

Feature selection of unbalanced breast cancer data using particle swarm optimization

Amal Elnawasany¹, Mohamed Abd Allah Makhoul¹, BenBella Tawfik¹, Hamed Nassar²

¹Department Information Systems, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

²Department of Computer Science, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

Article Info

Article history:

Received Jul 27, 2021

Revised May 28, 2022

Accepted Jun 14, 2022

Keywords:

Breast cancer

Feature selection

Genetic algorithm

Particle swarm optimization

Surveillance epidemiology

and end result

Synthetic minority

oversampling technique

Unbalance data

ABSTRACT

Breast cancer is one of the significant deaths causing diseases of women around the globe. Therefore, high accuracy in cancer prediction models is vital to improving patients' treatment quality and survivability rate. In this work, we presented a new method namely improved balancing particle swarm optimization (IBPSO) algorithm to predict the stage of breast cancer using unbalanced surveillance epidemiology and end result (USEER) data. The work contributes in two directions. First, design and implement an improved particle swarm optimization (IPSO) algorithm to avoid the local minima while reducing USEER data's dimensionality. The improvement comes primarily through employing the cross-over ability of the genetic algorithm as a fitness function while using the correlation-based function to guide the selection task to a minimal feature subset of USEER sufficiently to describe the universe. Second, develop an improved synthetic minority over-sampling technique (ISMOTE) that avoid over-fitting problem while efficiently balance USEER. ISMOTE generates the new objects based on the average of the two objects with the smallest and largest distance from the centroid object of the minority class. The experiments and analysis show that the proposed IBPSO is feasible and effective, outperforms other state-of-the-art methods; in minimizing the features with an accuracy of 98.45%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Amal Elnawasany

Department Information Systems, Faculty of Computers and Informatics, Suez Canal University

4.5 km the Ring Road, Ismailia, Egypt

Email: aml.elnawasany@ci.suez.edu.eg

1. INTRODUCTION

The surveillance epidemiology and end result (SEER) database [1] is an open cancer database that provides different cancers indicators for prognosis prediction. It contains information about the occurrence, frequency, survivability, and mortality of cancer. Cancer is typically labeled in stages from 1 to 4, with 4 being the most serious. The information collected in SEER mostly comes in high dimensionality [2]. Also, it is unbalanced, i.e., the objects in stages 2 and 3 are too larger than those in stages 1 and 4. Therefore, the database is referred to as an unbalance SEER (USEER) database. The two classes with the least number of objects are referred to as minority classes, while the other two are referred to as majority classes. The high dimensionality and unbalanced problems often hamper the breast cancer early prediction task and lead to delayed and inaccurate results, which degrade the patient's survival chance. Many research papers have been recently directed to address either the high dimensionality problem or the unbalance problem and the motivation behind this work is to propose an approach to address both.

Data reduction is a data preprocessing technique that aims at preparing the data for prediction. Instead of overwhelming the classifier with a huge amount of data, potentially causing many prediction errors, the classifier will have an easier job. Although data is shrinking, the fundamental and integrity of the original data should be retained. Data reduction decreases the processing time, storage space and computational complexity. Data reduction techniques include feature selection and instance selection. We will employ both techniques. SEER database contains instances for many types of cancers, while the cancer of interest is breast cancer. In this case, only the instances of breast cancer will be selected. This is called instance selection.

Feature selection (FS) tool is an approach that enables prediction algorithms to be applied to high dimensional data with less computations [3]. FS progresses by two steps: feature evaluation and feature set search. Feature evaluation evaluates each feature in the dataset separately in terms of its relevance to the class variable. On the other hand, feature set search tries various combinations of the evaluated features to arrive at a shortlist of features that sufficiently describes the objects [4]. Among the feature evaluation tools is correlation-based feature selection (CFS) [5]. The strength of the CFS tool comes in its ability to find a feature subset with features that are highly correlated with the class, yet uncorrelated with each other. CFS tool measures the goodness $f(\mathcal{M})$ of set \mathcal{M} of features ($\mathcal{M} \subset \mathcal{A}$) as (1):

$$f(\mathcal{M}) = \sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \frac{|\mathcal{M}| \rho_{a_i, a_j}}{\sqrt{|\mathcal{M}| + |\mathcal{M}|(|\mathcal{M}| - 1) \rho_{a_i, d}}}, \quad (1)$$

where ρ_{a_i, a_j} is the Pearson's correlation coefficient between features a_i and a_j and is given by (2):

$$\rho_{a_i, a_j} = \frac{\text{Cov}(a_i, a_j)}{\sigma_{a_i} \sigma_{a_j}}, \quad (2)$$

where $\text{Cov}(a_i, a_j)$ is the covariance which measures of the strength of the correlation between two features a_i and a_j and σ_{a_k} is the standard deviation of feature a_k .

The hurdle is that the size of the search space increases exponentially concerning the number of features, whereas CFS tool needs to assess 2^n feature subsets for USEER with n features. Therefore, CFS fails miserably when it confronts the USEER. A gap filled by swarm intelligent (SI) algorithms [6] is a low cost to search for a feature subset. Some examples of these algorithms are particle swarm optimization (PSO) algorithm [7], genetic algorithm (GA) [8], ant colony optimization (ACO) algorithm [9], artificial bee colony optimization (BCO) algorithm [10], bat search algorithm (BSA) [11], cuckoo optimization (CO) algorithm [12] and elephant herding optimization (EHO) algorithm [13]. The first two are the core of the approach proposed in the present article. Its simple operators characterize PSO algorithm and it is computationally inexpensive in terms of both memory and cost. PSO is an algorithm that solves FS problem by iteratively improving each particle position regarding a given measure of quality. The particle position is represented by a pivot vector pointing at a subset of features of the balanced SEER (BSEER). PSO algorithm assumes having a swarm of $P \geq 10$ particles moving in the search-space according to simple mathematical formula, known as a velocity function. The minimum number of particles is 10 because most of the swarms in nature have 10 particles on average. In each iteration $t \geq 1$, the particle that achieves the highest performance, being closer to the food, is referred to as the commander and the rest are slaves. The commander is chosen afresh in each iteration $t \geq 1$. The commander guides other particles to update their position to converge towards the food. Therefore, each particle $i = 1, 2, \dots, N$ in iteration $t + 1$ updates its position towards a better position according to the commander's position and the velocity function. At the end of the user defined number of iterations N , this exercise is expected to move the swarm toward their food's best solution. The good thing about PSO is that it does not make assumptions about the problem under study and can search high dimensional BSEER for minimal feature subset.

Initially, we consider that we have $P \geq 10$ particles. Each particle $i = 1, 2, \dots, P$ at iteration $t \geq 1$ has its own pivot vector $X_i = [x_1, x_2, \dots, x_n]$, where $x_j \in \{0, 1\}$. All the particles start at iteration $t = 1$ by pivoting randomly on a feature subset from the whole BSEER features. For each pivot vector X_i , we construct its corresponding feature subset $\mathbb{A} \subset \mathcal{A}$ by:

$$\mathbb{A} = \{a_j | x_j = 1\},$$

which represents the set of features whose corresponding values in X_i is 1. For example, consider $X_{12} = [1, 1, 0, 1, 0] \implies \mathbb{A} = \{a_1, a_2, a_4\}$.

PSO computes the goodness $f(\mathcal{A})$ of the original set of features \mathcal{A} in BSEER by (1). Then, at the end of each iteration $t \geq 1$, each particle is assessed through computing the goodness of their corresponding feature subset. The pivot vector with the corresponding highest goodness $f(\mathbb{A})$ is considered the commander and is assigned to \mathcal{X}_t . Each slave particle $i = 1, 2, \dots, P - 1$ updates its pivot vector $X_{i,t+1}$ in iteration $t + 1$ with respect to \mathcal{X}_t in two steps. First, it computes its velocity $V_{i,t+1}$ by which the particle updates its pivot vector at iteration $t + 1$ and is given by:

$$V_{i,t+1} = V_{i,t} + c_1 f(\mathcal{A}) \mathcal{X}_t - c_2 f(\mathbb{A}) X_{i,t}, \quad (3)$$

where c_1 and c_2 are two positive constants, in which $c_1 + c_2 = 4$. All the components of the velocity vector $V_{i,t}$ at $t = 1$ is of value 0. Second, the slave particles $i = 1, 2, \dots, P - 1$ find its updated pivot vector $X_{i,t+1}$ at iteration $t + 1$:

$$X_{i,t+1} = \mathcal{X}_t + V_{i,t+1}. \quad (4)$$

However, the PSO algorithm creates an undesirable feature subset $\mathbb{A} \subset \mathcal{A}$ of USEER which is insufficient to describe the universe [14]. This is because it employs the k -nearest neighbor (k NN) classifier as a fitness function which is frustrated by the unbalanced nature of USEER. This issue is commonly known as local minima feature subset; given a feature subset $\mathbb{A} \subset \mathcal{A}$, \mathbb{A} is said to be local minima feature subset if $f(\mathbb{A}) \leq f(\mathcal{A})$ for all values in specific interval but not the whole domain. Additionally, PSO algorithm is considered classifier-dependent algorithm as the resultant feature subset depends heavily on the accuracy of the k NN classifier and this fact, in turn, may result in poor accuracy with other classifiers. This calls for an improved PSO algorithm to deal with USEER utilized in the present work. GA is an evolutionary algorithm that mimics the biological behavior of genes. In contrast to the PSO algorithm, GA employs a cross-over technique that can update the feature subset it has been found so far and avoids being trapped in the local minima.

Definition 1 (Cross-over) Given two pivot vectors $Y = [1, 0, 0, 1, 1, 1, 0, 1]$ and $Z = [0, 1, 1, 0, 0, 1, 1, 0]$ where 1 in the i^{th} position means that feature a_i is selected and 0 otherwise. The cross-over technique randomly chooses a position and all bits beyond that position is swapped between the two vectors to generate two new vectors. For example, consider that position 4 is chosen then, the two new vectors are $Y' = [1, 0, 0, 1, 0, 1, 1, 0]$ and $Z' = [0, 1, 1, 0, 1, 1, 0, 1]$.

Rostami and Zadeh [15] state that the unbalance format of USEER negatively impacts the early prediction task of breast cancer. This is because the prediction algorithms have unpromising results on a minority classes than on majority ones [16]. Attempts to mitigate the unbalance problem are through converting USEER to a balanced one. This involves using object sampling (OS) technique that aims to have normally distributed objects among classes. OS is classified into two groups: under-sampling and over-sampling. The former progress by removing a set of objects from majority classes, while the latter progress by generating a set of objects in the minority classes. Its low cost characterizes Under-sampling on the contrary, over-sampling do not lose information, but it may result in an over-fitting problem, where the prediction algorithms fit to a specified set of objects and result in poor prediction accuracy with un-previously seen objects. Synthetic minority over-sampling technique (SMOTE) [17], an example of over-sampling, has put a great effort into balancing the USEER. It randomly chooses an object from the minority decision class and finds its k neighbor objects. Then it generates a new object by averaging the feature values of the k objects. The process is repeated till we have an equal number of objects in each class. This article remedies the FS and OS tasks' limitations of USEER. The rest of this article is organized as follows. Section 2 covers the related work. Section 3 describes the proposed approach. In section 4, the experimental work is carried out and discussions are given. Finally, the concluding remarks are presented in section 5.

2. RELATED WORK

Zhao *et al.* [18] introduce a predictive model for USEER using univariate and multivariate linear regression (LR). They aim to predict the patient's cancer stage using age, race, tumor size, primary site, pathological grade, histologic type, and molecular subtype features. However, [19] state that social features are more and more emphasized in breast cancer progression. Therefore, they introduce a predictive model for USEER to assess the impact of marital status on breast cancer. Furthermore, they used a chi-square method in [20] to analyze the associations between marital status and other features and a Kaplan Meier method to estimate survival curves. By and large, the models mentioned above result in low accuracy with the prediction algorithms. This is because they do not consider the unbalance classes, the main characteristic of USEER. OS technique

has been the topic of much research in recent years to alleviate the unbalance nature of USEER. Bertorello and Koh [21] use a density-based synthetic minority over-sampling (DSO) method to balance USEER. They use different weights for objects in the minority classes. Then they generate new objects regarding objects with the highest weight. On the contrary, Luo *et al.* [22] state that using the objects with the least weight is better in sampling to avoid misclassification. Tao *et al.* [23] propose a new over-sampling technique referred to as self-organizing map over-sampling (SOMO) to balance USEER. The SOMO technique generates new objects by producing a 2D representation of the input objects in the minority class, then averaging the closest objects. Wang [24] combine the strengths of PSO algorithm and CFS tool with two synthetic over-sampling methods; borderline-SMOTE and DSO with bayesian network (BN) algorithm and LR. Mirjalili *et al.* [25] examine 11 over-sampling techniques and 7 under-sampling techniques on 15 types of cancer. According to the study, USEER degrades the performance of classifiers. They state that balancing methods enhance the classification of USEER. Han *et al.* [26] introduce a distribution-sensitive over-sampling technique for balancing USEER. They divide the objects into noise, unstable, boundary, and stable objects according to their location in the minority class. They use a set of different methods to assess which objects are suitable to generate new objects. They use a set of different methods to assess which objects are suitable to generate new objects. Anupama and Jena [27] introduce increment over sampling for data streams (IOSDS) algorithm which uses a unique over-sampling technique to almost balance USEER. The IOSDS algorithm identifies noisy and mostly misclassified objects from the majority and minority classes by employing k -NN classifier. Then, it generates new objects in the minority classes using artificial, replication and hybrid objects. The trouble with the above-mentioned attempts is that they are sequential in nature, resulting in a delayed prediction. Tarkhaneh and Shen [28] introduce a Mantegna Lévy flight PSO and neighborhood search (LPSONS) algorithm to reduce the dimensionality of the USEER. They combine the strength of a velocity function, PSO algorithm with a Mantegna Lévy distribution function. This formulation leads to a more diverse feature subset. Additionally, to avoid being trapped in local minima, they combine the strengths of both a neighborhood search algorithm and a Mantegna Lévy distribution function. Pashaei *et al.* [29] introduced a binary version of PSO (BPSO) algorithm to avoid being trapped in local minima. Then, they combined the strengths of both a BPSO algorithms and a binary black hole optimization (BBHO) algorithm to improve the exploration and exploitation steps of BPSO algorithm. Afterward, they build a predictive model using a k -NN classifier to predict the patient's cancer stage early. The above attempts have one thing in common they create an undesirable feature subset of USEER because they do not consider the unbalanced nature of USEER, a gap that is filled by the present work. Fern'andez-Delgado *et al.* [30] evaluate 179 classifiers from 17 different families Bayesian, neural networks, random forests (RF), logistic regression. The RF classifier is at the forefront of the best classifiers. Ganggayah *et al.* [31] build prediction models using decision tree (DT), neural networks, support vector machine (SVM), RF and logistic regression algorithms to detect the significant indicators of breast cancer. The results detect the cancer stage as one of the most important indicators. The results were close; with the lowest accuracy obtained from DT and the highest obtained from RF. A study of [32] analyzed breast cancer at an early stage by comparing the performance of DT, RF and SVM. The results find that the RF performance is better than the other techniques for predicting cancer at an early stage. This article circumvents these problems by introducing an improved SMOTE to avoid the over-fitting problem and introducing an improved PSO algorithm to avoid get trapped in local minimal while dealing with the unbalance SEER.

3. RESEARCH METHOD

As shown in Figure 1, improved balancing particle swarm optimization (IBPSO) in this work conducts the improvement in two main directions: i) feature selection using an improved PSO algorithm (IPSO) and ii) balancing USEER using an improved SMOTE (ISMOTE). The first direction is FS which consists of two main steps: feature evaluation using CFS and feature set search by IPSO. First, CFS evaluates the relevance between each feature in the database and the class variable to find the highest associated features. Second, IPSO attempts different combinations of evaluated features to develop the best shortlist of features that adequately describe the objects.

Accordingly, IPSO algorithm is designed as shown in algorithm 1. We try different particles to pick the best number that fits the problem until the optimum result is saturated. IPSO algorithm tries a swarm of $10 \leq P \leq 50$ particles. IPSO uses the goodness function, given by (1), to assess the goodness of the selected feature subset in each iteration. When $t = 1$, PSO calculates the fitness value of the given BSEER and call it

B_goodness; best goodness. To this end, the algorithm aims to search for a minimal feature subset having the same B_goodness. A simple loop iterates P times to randomly initialize the P pivot vectors of the P particles, calculate the goodness of each corresponding feature subset \mathbb{A} . The pivot vector with the highest fitness value is assigned to \mathcal{X}_t . Afterward, IPSO employing the cross-over ability of GA to update the P particles pivot vectors and iterates till a user-defined number of iterations N is reached or a minimal feature subset with B_goodness is found.

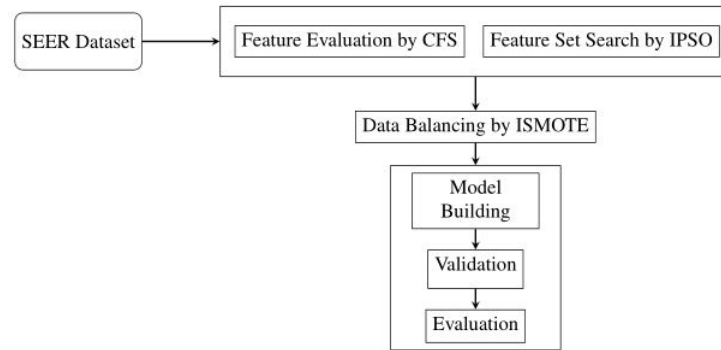


Figure 1. Process flow for the proposed IBPSO algorithm

The second direction is balancing data using ISMOTE which starts by finding the centroid for the minority decision class. Then, it computes the distance between the objects in the class and the centroid object using Euclidean distance. Finally, the newly generated object is the average of the two objects having the smallest and the largest distance from the centroid. This task is repeated until we have a BSEER.

Algorithm 1 Improved particle swarm optimization (IPSO) algorithm

Input: BSEER
 P , number of particles ($10 \leq P \leq 50$)
 N , number of iterations ($N \geq 2$)

Output: \mathbb{A} , a minimal BSEER feature subset ($\mathbb{A} \subset \mathcal{A}$)

$t := 1$
 B_goodness := $f(\mathcal{A})$ as per (1)

for $i=1$ to P **do**
 Construct randomly the pivot vector X_i
 $\mathbb{A} := \{a_j | x_j = 1\}$, set of features whose corresponding values in X_i is 1
 Calculate $f(\mathbb{A})$ as per (1)
end for

Assign the pivot vector with the corresponding highest $f(\mathbb{A})$ value to \mathcal{X}_t

for $k=2$ to N **do**
 $t := t + 1$
 Cross-over the two slave pivot vectors with the corresponding highest goodness as per Definition 1.
 for $i=1$ to P **do**
 Calculate the pivot vector X_i from X_{i-1} as per (3), (4)
 $\mathbb{A} := \{a_j | x_j = 1\}$
 Calculate $f(\mathbb{A})$ as per (1)
 end for
 Assign the pivot vector with the corresponding highest $f(\mathbb{A})$ value to \mathcal{X}_t .
 if $f(\mathbb{A}) = \text{B_goodness}$ **then**
 break.
 end if
end for

4. RESULT AND DISCUSSION

The experiments are conducted on SEER 1973-2016. SEER consists of 10,050,814 observations for all cancer types, only 1,631,572 cases diagnosed with breast cancer. From this population, we exclude 1,383,910 whose cause of death is not breast cancer. We further exclude 93,321 who have an unknown stage. Due to the impact of hurricane Katrina, 216 Louisiana cases diagnosed for those six-month period are excluded

from the research database. We excluded 280 cases that are not active follow-up, i.e., not keeping in touch with the patient for vital status, and exclude 2,770 cases that were not malignant cancers. The final cohort in our study on 151,075 with 160 variables. Table 1 shows the number of instances in different stages in BSEER before and after applying ISMOTE. The total number of balancing instances generated by ISMOTE is 177173. The final number of features selected by IPSO is 36 features. The selected features are survival months, first malignant primary indicator, total number of in situ/malignant tumors for a patient, radiation recode, chemotherapy recode, radiation sequence with surgery, laterality, histology, regional nodes positive, breast subtype, SEER cause-specific death classification, primary site, grade, tumors of adolescents and young adults site recode, breast-adjusted N (refers to the number of nearby lymph nodes that have cancer), breast-adjusted T (refers to the size and extent of the main tumor), scope of regional lymph node surgery (describes the performed procedure of removal, biopsy, or aspiration of regional lymph nodes), surgery of primary site (describes a surgical performed procedure that removes and/or destroys the tissue of the primary site), Appalachia, CS schema-AJCC 6th edition, Indian health service files to identify native Americans, Louisiana, month of diagnosis recode, hispanic identification algorithm (uses to classify cases as hispanic or not), record number (unique sequential number for each patient identifies the number of records submitted to SEER for that particular patient), SEER registry (used in conjunction with Patient ID to uniquely identify a patient), Site -mal+ins (mid detail) (which is should be used in conjunction with and only with the variables site specific (SS) sequence mal+ins (mid detail), SS sequence 1975+ mal+ins (mid detail), or SS sequence 1992+ mal+ins (mid detail) and they are already selected with our algorithm), SS sequence 1975+ -mal (most detail), SS sequence 1992+ mal (most detail), SS sequence 1992+ mal+ins (most detail), and Year of birth. Table 2 shows the performance measures of the IBPSO, the results without ISMOTE and the results without IPSO.

Table 1. Description of BSEER used in the experiments

Class	before ISMOTE (Percentage)	After ISMOTE (Percentage)
stage 1	9973 (6.60%)	19975(11.3%)
stage 2	1423 (34.04%)	51423 (29.0%)
stage 3	83581 (55.32%)	83581 (47.2%)
stage 4	6098 (4.04%)	22194 (12.5%)

The best number of particles that fit our problem was 20 particles. To validate the feature subset selected by IBPSO, we compare the results of IBPSO with five related SI algorithms namely, ACO, BCO, CO, BSA and EHO, using 10-fold cross-validation as shown in Table 2 and Table 3. We can see that the performance of the IBPSO is superior in selecting a smaller number of features while keeping its good classification performance.

Table 2. Performance measures for the IBPSO

Evaluation measure	IBPSO	Without ISMOTE	without IPSO
Accuracy	98.45%	64.79%	70.8%
Recall	0.985	0.648	0.708
Precision	0.985	0.652	0.714
F-Measure	0.984	0.621	0.694
MCC	0.974	0.343	0.489
ROC area	0.987	0.774	0.812
PRC area	0.986	0.706	0.724
MAE	0.0934	0.2255	0.2212
RMSE	0.1579	0.3351	0.325

To stress on the stability of the IPSO algorithm, Figure 2 shows the receiver operating characteristic (ROC) curve for the four cancer stages. ROC curve measures the classification algorithm's performance, the relation between classifier specificity and sensitivity at different thresholds. Classifier sensitivity represents the true positive rate, while specificity represents the truly negative rate. The farther the curve is from the diagonal line, the higher the overall accuracy of the model.

There are many problems with the SEER database:

- There are many blank(s) fields or unknown data. Unfortunately, excluding all blank(s) and unknown data leads to an empty matrix. This makes it impossible to remove all of them. So, we remove only blank and unknown fields from target features.

- Many features are gathered in only a specific period, which leads to inconsistency in the data. All these features are eliminated from our study. We select only data collected after 2010.

There were some degrees of missing in our data, but SEER encode the missing data; therefore, algorithms may be puzzled and deal with it as complete data. So, our further work would include other preprocessing techniques for missing data rather than deletion. Also, approach for hyperparameter optimization for the model parameter.

Table 3. Comparative experiments result of different SI algorithms

Evaluation measure	BSA	CO	ACO	BCO	EHO
# selected features	33	30	43	27	37
Accuracy	69.03%	70.74%	94.05%	64.36%	64.83%
Recall	0.69	0.707	0.941	0.644	0.648
Precision	0.694	0.713	0.941	0.663	0.654
F-Measure	0.672	0.694	0.94	0.604	0.625
MCC	0.455	0.488	0.899	0.362	0.371
ROC Area	0.775	0.812	0.991	0.775	0.765
PRC Area	0.669	0.728	0.986	0.689	0.684
MAE	0.2383	0.2208	0.0965	0.2454	0.228
RMSE	0.3364	0.3248	0.1797	0.3417	0.3376

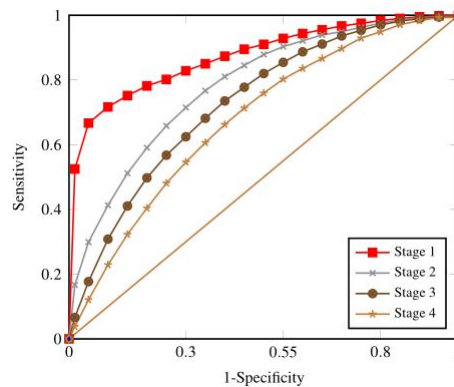


Figure 2. ROC curve for the four cancer stages

5. CONCLUSION

The proposed approach, IBPSO, is designed principally to process USEER data and predict the stage. IBPSO conducts the improvement in two directions. First, design and implement an IPSO algorithm to avoid being trapped in local minima while reducing USEER data's dimensionality. The improvement comes primarily through employing the cross-over ability of the genetic algorithm (GA) as a fitness function while using the correlation-based function to guide the selection task in IPSO algorithm. This idea leads to a minimal feature subset of USEER sufficiently to describe the universe. Second, develop an ISMOTE that avoid over-fitting problem while efficiently balance USEER. ISMOTE generates the new objects based on the average of the two objects with the smallest and largest distance from the centroid object of the minority class. The results show that IBPSO outperforms the related algorithms to find out a minimal feature subset with an accuracy of 98.45%. The classification accuracy of IBPSO is promising and superior to those achieved with different methods.

REFERENCES

- [1] SEER, "SEER*stat software." National Cancer Institute. <https://seer.cancer.gov/seerstat/> (accessed Jan. 2, 2020).
- [2] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *European Journal of Operational Research*, vol. 265, no. 3, pp. 993–1004, 2018, doi: 10.1016/j.ejor.2017.08.040.
- [3] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: a review and future trends," *Information Fusion*, vol. 52, pp. 1–12, Dec. 2019, doi: 10.1016/j.inffus.2018.11.008.




- [4] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease," *International Journal of Machine Learning and Computing*, vol. 5, no. 4, pp. 258–263, 2015, doi: 10.7763/ijmlc.2015.v5.517.
- [5] R.-J. Palma-Mendoza, L. De-Marcos, D. Rodriguez, and A. Alonso-Betanzos, "Distributed correlation-based feature selection in spark," *Information Sciences*, vol. 496, pp. 287–299, Sep. 2019, doi: 10.1016/j.ins.2018.10.052.
- [6] O. Ertenlice and C. B. Kalayci, "A survey of swarm intelligence for portfolio optimization: Algorithms and applications," *Swarm and Evolutionary Computation*, vol. 39, pp. 36–52, Apr. 2018, doi: 10.1016/j.swevo.2018.01.009.
- [7] S. Sengupta, S. Basak, and R. Peters, "Particle swarm optimization: a survey of historical and recent developments with hybridization perspectives," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 157–191, Oct. 2018, doi: 10.3390/make1010010.
- [8] S. Mirjalili, "Genetic algorithm," in *Evolutionary Algorithms and Neural Networks*, 2019, pp. 43–55, doi: 10.1007/978-3-319-93025-1_4.
- [9] M. Dorigo and T. Stützle, "Ant colony optimization: overview and recent advances," in *Handbook of Metaheuristics*, 2010, pp. 227–263, doi: 10.1007/978-1-4419-1665-5_8.
- [10] T. Dokeroglu, E. Sevinc, and A. Cosar, "Artificial bee colony optimization for the quadratic assignment problem," *Applied Soft Computing Journal*, vol. 76, pp. 595–606, 2019, doi: 10.1016/j.asoc.2019.01.001.
- [11] S. Fong, R. P. Biuk-Aghai, and R. C. Millham, "Swarm search methods in weka for data mining," in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, Feb. 2018, pp. 122–127, doi: 10.1145/3195106.3195167.
- [12] A. Rath, S. Samantaray, and P. C. Swain, "Optimization of the cropping pattern using cuckoo search technique," in *Studies in Fuzziness and Soft Computing*, vol. 374, 2019, pp. 19–35, doi: 10.1007/978-3-030-03131-2_2.
- [13] E. Tuba, D. Dolicanin-Djekic, R. Jovanovic, D. Simian, and M. Tuba, "Combined elephant herding optimization algorithm with k-means for data clustering," in *Smart Innovation, Systems and Technologies*, vol. 107, pp. 665–673, 2019, doi: 10.1007/978-981-13-1747-7_65.
- [14] L. Pitonakova, R. Crowder, and S. Bullock, "The information-cost-reward framework for understanding robot swarm foraging," *Swarm Intelligence*, vol. 12, no. 1, pp. 71–96, Mar. 2018, doi: 10.1007/s11721-017-0148-3.
- [15] S. M. Rostami and M. Ahmadzadeh, "Extracting predictor variables to construct breast cancer survivability model with class imbalance problem," *Journal of AI and Data Mining*, vol. 6, no. 2, pp. 263–276, 2018.
- [16] D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, and A. Bhardwaj, "Unbalanced breast cancer data classification using novel fitness functions in genetic programming," *Expert Systems with Applications*, vol. 140, Feb. 2020, doi: 10.1016/j.eswa.2019.112866.
- [17] K. Li and Y. Hu, "Research on unbalanced training samples based on SMOTE algorithm," *Journal of Physics: Conference Series*, vol. 1303, no. 1, Aug. 2019, doi: 10.1088/1742-6596/1303/1/012095.
- [18] Y.-X. Zhao, Y.-R. Liu, S. Xie, Y.-Z. Jiang, and Z.-M. Shao, "A nomogram predicting lymph node metastasis in T1 breast cancer based on the surveillance, epidemiology, and end results program," *Journal of Cancer*, vol. 10, no. 11, pp. 2443–2449, 2019, doi: 10.7150/jca.30386.
- [19] Y. Liu *et al.*, "Marital status is an independent prognostic factor in inflammatory breast cancer patients: an analysis of the surveillance, epidemiology, and end results database," *Breast Cancer Research and Treatment*, vol. 178, no. 2, pp. 379–388, 2019, doi: 10.1007/s10549-019-05385-8.
- [20] Y. Liu *et al.*, "The impact of marriage on the overall survival of prostate cancer patients: a surveillance, epidemiology, and end results (SEER) analysis," *Canadian Urological Association Journal*, vol. 13, no. 5, pp. 135–139, Oct. 2018, doi: 10.5489/cuaj.5413.
- [21] P. M. R. Bertorello and L. P. Koh, "SMate: synthetic minority adversarial technique," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3501279.
- [22] M. Luo, K. Wang, Z. Cai, A. Liu, Y. Li, and C. F. Cheang, "Using imbalanced triangle synthetic data for machine learning anomaly detection," *Computers, Materials and Continua*, vol. 58, no. 1, pp. 15–26, 2019, doi: 10.32604/cmc.2019.03708.
- [23] X. Tao *et al.*, "Real-value negative selection over-sampling for imbalanced data set learning," *Expert Systems with Applications*, vol. 129, pp. 118–134, Sep. 2019, doi: 10.1016/j.eswa.2019.04.011.
- [24] G. Wang, "A comparative study of cuckoo algorithm and ant colony algorithm in optimal path problems," *MATEC Web of Conferences*, vol. 232, Nov. 2018, doi: 10.1051/mateconf/201823203003.
- [25] S. Z. Mirjalili, S. Mirjalili, S. Saremi, H. Faris, and I. Aljarah, "Grasshopper optimization algorithm for multi-objective optimization problems," *Applied Intelligence*, vol. 48, no. 4, pp. 805–820, Apr. 2018, doi: 10.1007/s10489-017-1019-8.
- [26] W. Han, Z. Huang, S. Li, and Y. Jia, "Distribution-sensitive unbalanced data oversampling method for medical diagnosis," *Journal of Medical Systems*, vol. 43, no. 2, Feb. 2019, doi: 10.1007/s10916-018-1154-8.
- [27] N. Anupama and S. Jena, "A novel approach using incremental oversampling for data stream mining," *Evolving Systems*, vol. 10, no. 3, pp. 351–362, Sep. 2019, doi: 10.1007/s12530-018-9249-5.
- [28] O. Tarkhaneh and H. Shen, "Training of feedforward neural networks for data classification using hybrid particle swarm optimization, Mantegna Lévy flight and neighborhood search," *Heliyon*, vol. 5, no. 4, Apr. 2019, doi:

10.1016/j.heliyon.2019.e01275.




- [29] E. Pashaei, E. Pashaei, and N. Aydin, "Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization," *Genomics*, vol. 111, no. 4, pp. 669–686, Jul. 2019, doi: 10.1016/j.ygeno.2018.04.004.
- [30] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014
- [31] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-0801-4.
- [32] N. A. Farooqui, "A study on early prevention and detection of breast cancer using three-machine learning techniques," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 2, pp. 37–42, 2017.

BIOGRAPHIES OF AUTHORS






Amal Elnawasany    is a teaching assistant at the Faculty of Computers and Informatics. She obtained bachelor's degree in Information Systems from Suez Canal University in 2007. She obtained master's degree of information systems from the Suez Canal University (Egypt) in 2015. She can be contacted at email: aml.elnawasany@ci.suez.edu.eg.






Mohamed Abd Allah Makhlof    is currently an associate professor in Faculty of Computer Science and informatics Suez Canal University. He received his first degree in Computer Science and Operation Research, Faculty of Science, Master's degree in Expert Systems, Faculty of Science Cairo University. He received his Ph.D. degree in computer science, Faculty of Science, Zagazig University. He got the Post-Doctoral studies in Computer Science from Granada University, Spain in 2016. His research interests: machine learning, data mining, intelligent bioinformatics, metaheuristic optimization, decision support systems and predictive models. He can be contacted at email: m.abdallah@ci.suez.edu.eg.



BenBella Tawfik    born in September 1964, Cairo, Egypt, Graduated from Military Technical College in 1986. He got my master's in computer engineering in 1991 from the same school of graduation. He got his Ph.D. from Colorado State University, USA in 1998. In 2006, He visited USA-Colorado State University and earned Post Doctor Certificate in Computer Engineering. Besides his research work, he worked as a part time professor in many schools of Computer Engineering/Computers and Informatics in Egypt since 1998. He is working as an assistance professor in Information System Department in Faculty of Computer and Informatics – Suez Canal University - since October 2012. Three years ago, he is an associate professor working as a dean of the mentioned department. He can be contacted at email: benbellat@gmail.com.



Hamed Nassar    received the B.Sc. degree in electrical engineering from Ain Shams University, Egypt, in May 1979 and the M.Sc. degree in electrical engineering and the Ph.D. degree in computer engineering from the New Jersey Institute of Technology, USA, in May 1985 and May 1989, respectively. He was an Assistant Professor of computer engineering, in November 1989, an Associate Professor of computer science, in May 1998, Professor of computer science, in March 2004 and a Professor with the Department of Computers and Informatics, Faculty of Engineering, Beirut Arab University. Dr. Nassar is currently with the Department of Computer Science, Suez Canal University. His most recent publication is on a functional equation arising from a network model. His research interests include computing in mathematics, natural science, engineering and medicine, computer graphics and computer communications (networks). He can be contacted at email: nassar@ci.suez.edu.eg.