International Journal of Swarm Intelligence Research
Volume 14 • Issue 1

Breast Cancer Classification With Microarray Gene Expression Data Based on Improved Whale Optimization Algorithm

S. Sathiya Devi, University College of Engineering, Bharathidasan Institute of Technology, Trichy, India*
Prithiviraj K., University College of Engineering, Bharathidasan Institute of Technology, Trichy, India

ABSTRACT

Breast cancer is one of the most common and dangerous cancer types in women worldwide. Since it is generally a genetic disease, microarray technology-based cancer prediction is technically significant among lot of diagnosis methods. The microarray gene expression data contains fewer samples with many redundant and noisy genes. It leads to inaccurate diagnose and low prediction accuracy. To overcome these difficulties, this paper proposes an Improved Whale Optimization Algorithm (IWOA) for wrapper based feature selection in gene expression data. The proposed IWOA incorporates modified cross over and mutation operations to enhance the exploration and exploitation of classical WOA. The proposed IWOA adapts multiobjective fitness function, which simultaneously balance between minimization of error rate and feature selection. The experimental analysis demonstrated that, the proposed IWOA with Gradient Boost Classifier (GBC) achieves high classification accuracy of 97.7% with minimum subset of features and also converges quickly for the breast cancer dataset.

KEYWORDS

Accuracy, Crossover and Mutation, Feature Selection, Gradient Boosting Classifier, Multi Objective Optimization, Support Vector Machine, Whale Optimization Algorithm

1. INTRODUCTION

Cancer is a second dangerous disease that causes 9.6 million deaths worldwide. There are approximately 21. 7 Million people in the world is suffering from cancer by 2030 and predicted 30 million deaths (Aldryan et al., 2018). There are different types of cancer among which breast cancer is the common (prevalent) among females. Nearly one fourth of female population is affected by this cancer irrespective of age factor in India and is common in rural India. The majority of the cancer types can be caused due to either genetic (hereditary) or epigenetic changes and generally 90% of the breast cancer is due to genetic abnormalities. The variations in high penetrance genes such as BRCA1, BRCA2, p53, PTEN, ATM, NBS1, LKB1, etc. can produce genetic abnormalities (Dumitrescu

DOI: 10.4018/IJSIR.317091 *Corresponding Author

& Cotarla, 2005). The common symptoms of breast cancer are (i). a lump in the breast, (ii) blood discharge from the nipple and (iii). shape or texture changes in the nipple or breast. Since the breast cancer cannot be accurately diagnosed by a single clinical test, it requires many tests along with the complete history of the patient and their physical examination. Based on these results, the physician could (i). Identify and confirm the disease, (ii). Consistently monitoring the disease progress and (iii). Schedule for and assess the viability of the treatment. The above mentioned classical diagnosis methods result in uncertain diagnosis and prone to human error. It also requires skilled labors and is time consuming, which causes stress throughout the diagnostic process. Hence the early detection of cancer to reduce the risk of death requires an accurate and reliable diagnosis processes as well as the use of robust tools and techniques.

DNA Microarray technology based cancer prediction is technically significant among lot of diagnosis methods and is used by researchers and clinicians for the past two decades. The recent technologies made the availability of thousands of benchmark gene expression assays through online for microarray data analysis to predict different types of cancerous tumors. The microarray dataset consists of huge number of genes corresponding to small sample size and the genes are highly correlated. The high dimensionality of the genes and small sample size is a challenge for the effective analysis and diagnosis of microarray data resulting in poor diagnosis and prediction accuracy. Hence this paper addresses this issue by proposing an Improved Whale Optimization Algorithm (IWOA) for feature selection and ensemble based classification for breast cancer prediction.

The remaining portion of this paper is organized as follows: section 2 reviews the relevant and significant previous work. The classical Whale Optimization Algorithm (WOA) is described in section 3. The proposed IWOA based feature selection with Support Vector Machine (SVM) and GBC is described in section 4. Section 54 discusses the experimental and result analysis with data set and performance measure. Conclusion is presented in section 6.

2. RELATED WORK

In the microarray dataset, the high ratio between the huge dimension of the genes (features) and the few number of samples resulted in inaccurate and imbalanced cancer prediction. In common, most of the genes in the microarray data are uninformative and redundant. These types of the genes are to be identified with the machine learning technique called as feature subset selection. Though the feature selection techniques not only identify the significant genes, it also improves the classification accuracy. There are three approaches for feature selection: (i) Filter method, (ii) Wrapper method and (iii) Hybrid method. In the filter method, the feature importance is measured with properties of the dataset and order the features based on the relevance score (feature importance score). This method is simple and fast and not considering the correlation among the features. The wrapper method generally incorporates any predefined classification algorithm to search for and select the relevance features. This method considers the feature dependencies and computationally intensive and slower. The hybrid method is the combination of filter and wrapper methods. This method, first apply the filter technique to reduce the feature space then use the wrapper method for feature subset selection. Since the wrapper method is expensive, it is proved to be beneficial in finding feature subsets that suit a predetermined classifier (Alshamlan et al., 2015).

A. K. Shukla et. al. (2019) have introduced a hybrid wrapper approach called TLBOSA, which is the combination of Teaching Learning based Optimization (TLBO) and Simulated Annealing (SA) with SVM for gene expression data. It overcomes the exploitation issue and produces better classification accuracy with small subset of genes. To identify the more discriminative subset of genes and to reduce the dimensionality, the Gravitational Search Algorithm (GSA) is combined with TLBO called TLBOGSA has been described in (Shukla et al., 2020). This method achieves higher classification accuracy with less computational cost when compared with six datasets. P. Gunasekhar et. al. (2020) have used six different filter based approaches for biomarker feature selection. From the

selected features, two high ranked features are extracted with Modified Social Ski Driver Optimization (MSSO) algorithm. Then the cancerous tissues are predicted based on Sunflower optimization based Deep Neural Network (DSFNN) approach.

An Intelligent Decision Support System (IDSS) for early detection of cancer based on microarray data has been introduced in (Abdelnabi et al., 2020). The IDSS incorporates Information Gain (IG) as a initial measure to select the subset of genes and Gray Wolf Optimization (GWO) algorithm is further applied to select an optimal subset for prediction and also overcomes the overfiting problem. Finally SVM is used to classify the cancerous genes. The reliability of this system is tested by incorporating other benchmark datasets with binary and multi class dataset.

V. Nandagopal et. al. (2019) have described the feature (genes) selection based on fuzzy logistic regression with LASSO Logistic Regression (LLR) for prediction. This model eliminates unnecessary covariates and produces a classification accuracy of 94.05%. This method imputes the missing data using Expectation Maximization (EM) algorithm.

The two stage feature selection approach utilizing Spearman's Correlation (SC) and distributed filter FS methods which can select the highly discriminative genes for distinguishing samples from high dimensional datasets have been introduced in (Shukla & Tripathi, 2019). The distributed filter method quantify the relation between gene-gene and the gene-class and simultaneously detect subsets of essential genes. The method is verified with four classifiers among which, SVM produces high classification accuracy.

Md. Maniruzzaman et. al. (2019) have utilized the statistical tests and Machine Learning (ML) strategy to identify the high risk differential genes and prediction of cancer genes respectively. The ML strategy used ten classifiers namely Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naïve Bayes (NB), Gaussian Process Classification (GPC), Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression (LR), Decision Tree (DT), Adaboost (AB), and Random Forest (RF). This strategy produced the 99.81% accuracy with Wilcoxon sign rank sum (WCSRS) test and RF based classifier.

The t-test and a nested Genetic Algorithm (GA) based feature subset selection for microarray gene data have been discussed in (2019). The nested GA consists of outer and inner GA that run on two different kinds of datasets. The Outer Genetic Algorithm (OGA-SVM) works on Microarray gene expression datasets, whereas the Inner Genetic algorithm (IGA-NNW) runs on DNA Methylation datasets. After applying nested GA, the Incremental Feature Selection (IFS) method is applied to obtain the optimal subset of genes. The authors claimed that this method achieves better accuracy when compared with existing methods.

To obtain minimum subset of genes and maximum classification accuracy, a new frame work called C-HMOSHSSA has been introduced in (2019) by utilizing the meta heuristic algorithms. The combination of Multi Objective Spotted Hyena Optimizer (MOSHO) and Salp Swarm Algorithms (SSA) are used to identify the optimal subset of genes. This hybrid algorithm improves the exploration and exploitation capability and yields better classification accuracy. Similarly, the Whale Optimization Algorithm (WOA) and Mixed Kernel Function (MKF) based SVM have been introduced in (Zhao et al., 2019). This approach not only obtains the optimal subset of features and also solves the data imbalance issue. Recently Altruistic Whale Optimization Algorithm (AltWOA) has been proposed in (Kundu et al., 2022) to obtain the global optima in the feature subset selection process and also to increase the classification accuracy.

Ant Colony Optimization (ACO) and Modified Back Propagation Conjugate Gradient Polak Ribiere based feature selection and microarray gene classification is presented in (Aldryan et al., 2018). The hybrid feature selection algorithm that combines the Mutual Information Maximization (MIM) and the Adaptive Genetic Algorithm (AGA) have been described in (Lu et al., 2017). The MIMAGA feature Selection method significantly reduces the dimension and redundancy of the gene expression data and increases the classification accuracy.

An hybrid Enhanced ANFIS (EANFIS) with Manta ray Foraging Optimization (MaFO) algorithm has been described in (Mishra & Bhoi, 2021) to increase the classification accuracy and to minimize the convergence time of the learning process. S. Belciug (2020) has developed a method to solve the curse of dimensionality and the curse of sparsity issue by merging the two ML approaches such as Logistic regression and single hidden layer feed forward neural network. This combination obtained the classification accuracies between 64.70% and 98.66% depending on the dataset. The author claimed that this approach is problem dependent and yields better result with less computational cost and high speed. Similarly the combination of Information Gain (IG) and Grey Wolf Optimization (GWO) algorithm are used to select the subset of genes in (El Nabi et al., 2020). In this work initially, IG is applied to the Microarray data to select the genes and further the sub set is reduced by applying GWO. A. Tahmouresi (2022) et. al. have described the Pyramid Gravitational Search Algorithm (PGSA) to solve high-dimensional gene selection problems. It is a hybrid approach that cyclically reduces the number of genes and selects the genes producing high classification accuracy.

From the literature review, it is noticed that searching an optimal subset of features (informative genes) from the high dimensional, redundant and large scale micro array data is a crucial task and it also affects the performance of the classification accuracy. The cancer classification based on microarray data can be viewed as a non-deterministic polynomial time (NP – hard) problem because of the high dimensionality and correlation in the dataset. Therefore meta heuristic based feature selection with predefined classifier is an appropriate approach for cancer prediction using microarray data. Hence this paper proposes an Improved WOA (IWOA) incorporating cross over and mutation operator for feature (gene) subset selection and Gradient boost ensemble classifier for cancer prediction. The classical WOA method is described in the next section.

3. WHALE OPTIMIZATION ALGORITHM (WOA)

The WOA algorithm is one type of population based optimization algorithm proposed recently in (Mirjalili & Lewis, 2016) for obtaining global optimum. This algorithm has three operations such as (i). Search for prey, (ii) Encircling prey and (iii) Bubble net foraging behavior of humpback whales. These three operations are grouped into two phases such as:

1. **Exploitation (Intensification) Phase:** In this, the prey is encircled and spiral bubble net attacking is performed.
2. **Exploration (Diversification) Phase:** In which, searching a prey randomly is performed.

The mathematical formulation and the two phases of WOA are elaborated as follows: The WOA model the movement and to update the positions of a whale around a prey mathematically using the equations (1) and (2):

$$D = \left| \vec{C} \cdot \overrightarrow{X_{best}}\left(t\right) - \vec{X}\left(t\right) \right| \tag{1}$$

$$\vec{X}\left(t+1\right) = \vec{X}_{best}\left(t\right) - \vec{A} \cdot \vec{D} \tag{2}$$

where t is the current iteration number, X_{best} represents the position of the best solution obtained so far and X is the current solution position. The coefficient vectors A and C can be calculated from the equations (3) and (4) respectively:

$$\vec{A} = 2\,\vec{a}\ .\ \vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2\ .\ \vec{r} \tag{4}$$

where *r* is a random vector belongs to the interval [0, 1] and *a* decreases linearly through the iterations from 2 to 0. The values of *A* and *C* vectors are adjusted to control the areas where the solution is located in the neighborhood of the best solution.

3.1 Exploitation (Intensification) Phase

It consists of two steps. The first one is the shrinking encircling a prey which can be obtained by reducing the values of *a* using Equation (3) according to equation (6):

$$\vec{a} = 2 - \vec{t}\,\frac{2}{\left(Max\_Iteration\right)} \tag{5}$$

$$\vec{a} = 2\,(1 - \frac{\vec{t}}{Max\_Iteration}) \tag{6}$$

where *t* is the iteration number and *Max_Iteration* is the total number of iterations.

The second step is the spiral bubble net attacking method to update the position. To simulate the spiral path and to calculate the distance between whale (or) solution (*X*) and the prey (or) best solution (X_{best}), the equation (7) is used. This equation produces helix shaped curve to mimic the behavior of the spiral path:

$$\vec{X}\,(t+1) = D^{l} \,.\, e^{b} \,.cos\,(2\pi l) + \overrightarrow{X}_{\text{best}}\left(t\right) \tag{7}$$

where $D^{l} = \left| \overrightarrow{X_{best}}\left(t\right) - \vec{X}\left(t\right) \right|$ is the distance between prey (best solution obtained so far) and the whale, *b* is a constant for defining the shape of the spiral and *l* is a random number in range $[-1,1]$.

A probability of 50% is considered to model shrinking encircling and the spiral shaped path during optimization and is represented mathematically as:

$$\vec{X}(t+1) = \begin{matrix} \vec{X}_{best}\left(t\right) - \vec{A}\cdot\vec{D} & if(p < 0.5) \\ D^{l}\cdot e^{b}\cdot\cos\left(2\pi l\right) + \vec{X}_{best}(t) & if(p \geq 0.5) \end{matrix} \tag{8}$$

where *p* is the random number in the uniform distribution between [0,1].

3.2 Exploration (Diversification) Phase

In order to perform global search and to compute the best solution, the value of the vector A is assumed between 1 < A < -1 to force the whale to move far away from the reference whale (best known solution). The updated position is mathematically modeled as:

$$D = \left| \vec{C}\cdot\overrightarrow{X}_{rand} - \vec{X} \right. \tag{9}$$

$$\vec{X}\left(t+1\right) = \vec{X}_{best} - \vec{A}\cdot\vec{D} \tag{10}$$

where $\vec{X}_{rand}$ is the random whale taken from current iteration.

The two phases explained above are iteratively used until the termination condition is reached. To find the solution, the WOA perform the following steps:

- **Initialization:** The whales are initialized randomly to constitute the initial population. The fitness values of each whale are computed based on the objective function and choose the best whale position from this population.
- **Whale Position Updation:** In this step, the position of the whale is updated with bubble net hunting method. The current whale position depends on the position of the best whale identified in the so far iterations or the random whale position is chosen if the specific condition prevails. Based on the above conditions, the best position is updated with the new whale that have better fitness value.
- **Check for termination Condition:** If termination condition which is the total number of generations or iterations is reached, the algorithm returns the best position of the whale as a optimum solution. Otherwise go back to Whale Position Updation step.

The objective function and the termination criteria are problem dependent. Though the WOA is efficient in identifying global optimum solution, the convergence speed is slow and exhibits high computational cost. Generally the low convergence of WOA is due to randomization and inability to obtain diverse optimal solution from the population. In order to improve the exploration capability and higher convergence speed, this paper proposes an Improved Whale Optimization algorithm (IWOA) incorporating crossover and mutation operator of GA into WOA. The IWOA balances between exploration and exploitation and to increase the convergence speed. The next subsection describes the proposed IWOA and IWOA based feature selection.

4. THE PROPOSED IWOA BASED FEATURE SELECTION

4.1 Proposed IWOA

The IWOA algorithm is the hybridization of WOA and the modified GA operator such as crossover and mutation. The IWOA uses modified crossover and mutation operator when the whale updates its position towards the best solution. Due to this, the diverse optimum solution is computed and the converge speed also get increased. In WOA, each whale (solution) updates its position linearly based on the equation (2) and (10). In these equations, the values of distance control parameter a guarantees the faster convergence and better exploration. However, the search process of WOA is nonlinear and the linearly decreasing value of a does not sufficiently fit to this process. Therefore, the proposed IWOA incorporates the modified GA operators such as crossover and mutation for quick convergence. But the crossover operation that is used in the proposed method is not the usual crossover. A modified crossover is defined as follows.

The crossover used in the proposed IWOA is different from the conventional crossover operator in GA. In the traditional method, the cross over point is chosen by random for processing. But in the proposed method, the cross over is performed between two solutions using the combination of AND and XOR logic. The solution (whale) is divided into even number of partitions, even though in some cases the MSB (Most Significant Bit) posses unequal number of bits. The crossover operation is performed between two solutions from LSB to MSB with the combination of XOR and AND operation alternatively. Similarly, the mutation is carried out in a middle partition of the solution by altering only the '1' into '0'. The reason behind this crossover and mutation operation is to reduce the number of features and randomness in the operations. The diagrammatic representation of modified cross over and mutation operators are given in figure 1.

The overall modified cross over operation is given as follows:

1. From the population, the parent set is generated by joining the solutions.
2. Divide the parent solution into even number of segments, though the segment with the MSB bits is greater or lesser than the remaining segments bit.

3. Apply AND – XOR operation on the segments from LSB to MSB in an alternative way to generate child (or) offspring.

Based on this concept the proposed IWOA is outlined and is shown in figure 2.

In this paper, wrapper based feature selection with SVM and GBC based classification is adapted. So the main objective of this work is to obtain high classification accuracy (minimum error rate) with less number of features.

- **Solution representation:** In the proposed work, the feature selection is considered as a binary optimization problem. Hence, each solution (whale) is represented as one dimensional vector of binary values, where the length of the vector is equal to the total number of features (attributes) of the dataset. The binary value 1 represents the presence of the feature and 0 represents the feature is not selected.
- **Fitness function:** The fitness function designed in the proposed work is a multi-objective where it has to satisfy the two contradictory objectives such as minimum subset of features and high classification accuracy. The feature subset selection is a minimization approach and obtaining the increased classification accuracy is a maximization approach. So the fitness function should be to balance between these two objectives. The fitness function used in this work is the minimization function and is represented in equation (11) to evaluate the solution as:

$$fitness = \alpha * err + \beta * \frac{LS}{L} \tag{11}$$

where *err* is the error rate (or) misclassification accuracy of the classifier, *err* = 1 – *classification accuracy*. *LS* and *L* represents the length of the selected feature subset and total number of features

Figure 1. (a) Modified Cross over; and (b) Modified Mutation

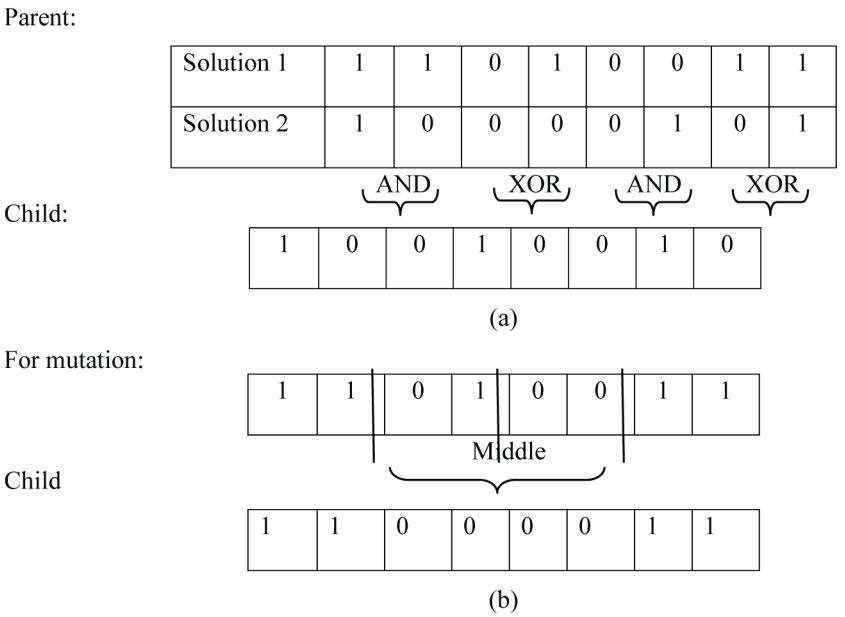


Figure 2. Improved Whale Optimization Algorithm

Input:
max_Iteration, *n*, FitnessFunction // n = population size
Begin:
Initialize $t = 1$, T = max_Iteration;
Generate initial population randomly, $X(t) = \{ X^1(t), X^2(t), \ldots, X^i(t), \ldots, X^n(t)\}$;
Compute the fitness value of each solution;
$Z_{best}(t)$ = BestSolution($X(t)$);
best_Index = BestSolutionIndex($Z_{best}(t)$);
$X_{best}(t) = X_{best\_index}(t)$;
While $(t < T)$
for each solution, $i = 1,2,3, \ldots, n$
Choose the random value of p and update the value of A with Eq. (3)
if $(p < 0.5)$
if $(|A| >= 1)$
// **Explore** ($X(t)$)
$X_R(t)$ = RankSelection($X(t)$)
$X_M(t)$ = Mutation($X_R(t)$)
$X_C(t)$ = Crossover($X_R(t)$, $X_i(t)$)
$X_i(t)$ = Best($X_M(t)$, $X_C(t)$)
else if $(|A| < 1)$
// **Shrink** ($X_{best}(t)$)
$X_i(t)$ = Crossover(Mutation($X_{best}(t)$), $X_{best}(t)$)
End if
else if $(p >= 0.5)$
// **Sprial** ($X_{best}(t)$, $X_i(t)$)
D = distance($X_{best}(t)$, $X_i(t)$)
$X_i(t) = D.e^b .cos(2\pi l) + X_{best}(t)$
end if
end for
Calculate the fitness value of each solution, $X(t) = \{ X^1(t), X^2(t), \ldots, X^i(t), \ldots, X^n(t)\}$;
$Z_{best}(t)$ = BestSolution($X(t)$);
best_Index = BestSolutionIndex($Z_{best}(t)$);
$X_{best}(t) = X_{best\_index}(t)$;
t = t +1
end while
return X_{best}

Table 1. Parameter Settings of the proposed IWOA and WOA

S.no.	Parameters	Values
1	Population Size	10,30,50
2	Number of Generations	50,70,100
3	Number of runs	5
4	Performance Measure	Accuracy, Precision, Recall, Error rate, Specificity, ROC Score, Convergence

in the dataset respectively. α and β are the control (or) weight parameter for classifier error rate and feature subset respectively where $\alpha \varepsilon (0,1)$ and $\beta = (1 - \alpha)$. The value of these parameters is set as 0.99 and 0.01 respectively. The values are chosen in such a way the, the proposed approach select minimum features for ensuring the less error rate.

In the proposed IWOA approach, the initial solutions (whales) are randomly generated. The fitness value of each solution is computed with classifier and the positions of each whale updates accordingly. The WOA solutions are updated by considering one solution with some mathematical operations, whereas in GA the solutions are selected based on the fitness value and combined to produce better solution using cross over operator. Similarly mutation is performed in one solution by randomly modifying the bit position. This concept improves the exploration ability and the WOA's of adaptive mechanism accelerate exploitation proportional to the number of iterations. Hence the hybridization between global and local search approach produce better solution for feature selection.

5. EXPERIMENTS AND RESULTS

5.1 Experimental Setup and Datasets

The proposed IWOA approach is implemented using Python in a Spyder Integrated Development Environment (IDE). The packages such as Collections, MatplotLib, NumPy, Pandas and Scikit are used to conduct the experiments in Intel Core i5 machine, 2.2 GHz CPU and 4GB of RAM. The dataset considered for this experiment is taken from Gene Expression Omnibus (GEO) which can be accessed in the NCBI cancer portal (http://www.ncbi.nlm.nih.gov/geo/).

The size of the dataset *Breast.txt* is 40.41 MB of data. It contains 47294 features and 128 samples. The features are the genes that are collected from the microarray. The 47294 features are the gene spots which contain probes that are presented in the microarray and 128 samples are the gene nucleotides which are hybridized with the sample mRNA genes and probes. There are two class labels available which are Luminal and Non Luminal. The Luminal class label mentions that genes lead to cancer and non luminal tells that genes are non cancerous.

The experiments are conducted initially with the proposed IWOA using SVM classifier with linear and non linear kernels. Since the proposed IWOA and classical WOA are naturally inspired algorithms and runs with a set of suitable parameter setting for comparing the efficiency between them. The parameters are shown in table 1.

The experiment divides the dataset into training and validation set with the ratio of (70, 30) respectively. The experiments consider K – fold cross validation with M times to obtain the accuracy for the training dataset. The performance of the proposed method is evaluated with the metrics such Accuracy, Precision, Recall, Specificity, ROC Score and Error rate. The model accuracy is computed by validating the validation dataset. Throughout the experiment, the proposed approach considers 5 – fold cross validation with $M = 5$. The proposed IWOA adapts wrapper approach for feature selection and fitness evolution, requires classifier. The classifier used in this paper is SVM and Gradient Boost Classifier (GBC). The experiment considers both linear and non linear SVM

Table 2. Parameter set for SVM and Gradient Boost classifiers for tuning

Name of the classifier	Parameter set used for tuning
Support Vector Machine (SVM)	Kernel: Linear Kernel, Polynomial kernel with order = 3, Radial Basis Function (RBF) Gamma: 1 (for RBF kernel) Regularization (C): 0.1
Gradient Boost Classifier (GBC)	Learning rate: 0.01 Max_Depth: 3 Subsample: 0.6 n_estimators: 100

with the classic kernel functions. The hyper parameter set used for tuning of classifiers such as SVM and GBC are given in table 2.

5.2 Performance Measure

The performance of the proposed IWOA is evaluated by various measures such as Accuracy, Precision, Recall, Error rate, Specificity and ROC Score. These measures are calculated from the confusion matrix and the elements of this matrix are:

1. **Positive (P):** The sample belongs to positive class.
2. **Negative (N):** The sample belongs to negative class.
3. **True Positive (T_P):** The classifier model correctly predicts the instance belongs to positive class.
4. **True Negative (T_N):** The classifier model correctly predicts the instance belongs to negative class.
5. **False Positive (F_P):** This is also called as type 1 error and the classifier model incorrectly predicts the instance as positive, but it is actually negative.
6. **False Negative (F_N):** This is also called as type 2 error and the classifier model incorrectly predicts the instance as negative, but it is actually positive.

These measures are defined based on confusion matrix from Equation (12) to Equation (18):

- **Accuracy:** Accuracy is defined as the ratio of total number of instances predicted correctly to total number of instances in the dataset, which is defined in the Equation (12) as:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_N + F_P} \tag{12}$$

- **Error Rate:** Error rate is the misclassification of the samples and is defines as:

$$Errorrate = 1 - Accuracy \tag{13}$$

- **Precision:** Precision is the ratio of total number of relevant instances to the retrieved instances and is given as:

$$Precision = \frac{T_P}{T_P + F_P} \tag{14}$$

- **Recall:** Recall which is also known as sensitivity is the ratio of correct positive instances retrieved to the total number of positive instances and is represented as:

$$Recall = \frac{T_P}{T_P + F_N} \tag{15}$$

- **Specificity:** Specificity is defined as the proportion of actual negative instances, which is also predicted as the negative instances and is calculated as:

$$Specifity = \frac{T_N}{T_N + F_P} \tag{16}$$

- **ROC Score:** The ROC score is the plot between True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds of the classification model. The computation of TPR and FPR are given as follows:

$$TruePositiveRate\left(TPR\right) = \frac{T_P}{T_P + F_N} \tag{17}$$

$$FalsePositiveRate\left(FPR\right) = \frac{F_P}{T_N + F_P} \tag{18}$$

5.3 Experimental Results Discussion and Analysis

During experimentation, the initial population (whales) for the proposed IWOA is set randomly with size as 10 and the number of iteration is 50. Initially the SVM with linear kernel is used for classification along with the validation setting parameter as discussed in the previous subsection 5.1. The obtained result with various measures is presented in table 3. The classification accuracy of the proposed IWOA is 93% with the feature size as 383 and the optimum solution is converged in 22th iteration, which is half of the total number of iterations considered. The size of the initial population and the number of iterations influences the accuracy of the classification. Hence, to improve the accuracy, convergence and error rate, the size of the population and the number of iterations are set as (30,70), and (50,100) respectively.

The experiments are carried out with new set of search parameters and the maximum iterations and the obtained results are appended into the same table 3. The classification accuracy is improved from 93% to 95% for the population size and number of iterations as 30,70 respectively. Similarly the Error rate, Precision, Recall, Specificity and AUC score are also significantly improved. The feature size and the convergence are also reduced significantly and the population converges in 12th iteration and the size of the feature subset is 326. The experiments are conducted with the same set of parameters with polynomial kernel of order 3 and RBF kernel of SVM along with the parameters given in table 2. The results are also appended into the same table 3. From the table, it is inferred that, the linear kernel performs well when compared to polynomial and RBF kernel. The parameters such as size of the population and total number of iterations are influenced the classification accuracy. The performance of the proposed IWOA is compared with classical WOA using same parameter setup. The WOA has been experimented and the results are presented in table 4.

The classic WOA algorithm also produces high accuracy 92.5% for the linear kernel when compared with other kernels. But the accuracy is 3% lesser than the proposed IWOA. The remaining measures are

Table 3. Performance of IWOA with various SVM kernel based on different parameters

Number of whales	10	30	50
Number of iteration	50	70	100
No. of trails	10	10	10
SVM classifier with Linear Kernel			
Feature size	383	326	370
Accuracy (%)	93	95	94
Error rate	0.0692	0.05	0.06
Precision (%)	93	95	92
Recall (%)	92	100	94
Specificity (%)	86	97	96
AUC score	0.876	0.91	0.89
convergence	22	12	8
SVM classifier using Polynomial kernel with order=3			
Feature size	395	648	1489
Accuracy (%)	94	90	86.5
Error rate	0.06	0.097	0.135
Precision (%)	94	90	87
Recall (%)	96	95	93
Specificity (%)	92	95	80
AUC score	0.91	0.88	0.85
convergence	24	4	4
SVM classifier using RBF kernel			
Feature size	601	484	351
Accuracy (%)	83	88.2	86
Error rate	0.17	0.12	0.14
Precision (%)	82	86	98.5
Recall (%)	97	98	100
Specificity (%)	97	89	70.5
AUC score	0.77	0.83	0.77
convergence	41	21	30

also yield better results for this kernel. It is noticed from the table 4 that, the classic WOA gives better performance with population size (no. of whales) 10 and with maximum iteration 50. The solution converged in 49th iteration itself. Though the highest accuracy of the proposed IWOA algorithm is obtained with the population size as 30 and maximum number of iteration as 70, the feature size is much reduced in this method than the classic WOA. From the set of experimental results, it is inferred that feature dimension and size of the population play a vital role in increasing the accuracy of the classifier. The proposed IWOA and WOA both converges within 50 iterations itself. Hence the total number of iterations should be decided based on the empirical result. The proposed IWOA with modified crossover and mutation operation significantly reduces the dimension of the features and provides better accuracy.

Table 4. Performance of WOA with various SVM kernel based on different parameters

Number of whales	10	30	50
Number of iteration	50	70	100
No. of trails	10	10	10
SVM classifier with Linear Kernel			
Feature size	430	631	602
Accuracy (%)	92.5	89	92
Error rate	0.0745	0.11	0.08
Precision (%)	93	93	89
Recall (%)	96.3	96.9	97
Specificity (%)	83	81	90
AUC score	0.82	0.84	0.83
convergence	49	39	44
SVM classifier using Polynomial kernel with order=3			
Feature size	787	631	523
Accuracy (%)	91	90	91.5
Error rate	0.09	0.97	0.083
Precision (%)	96	93.6	94
Recall (%)	96.43	96.1	96
Specificity (%)	95	93	85
AUC score	0.88	0.86	0.86
convergence	40	48	48
SVM classifier using RBF kernel			
Feature size	601	637	618
Accuracy (%)	83	87	89.9
Error rate	0.17	0.14	0.101
Precision (%)	94	86	91
Recall (%)	93	98	89
Specificity (%)	87	89	93
AUC score	0.77	83	0.84
convergence	21	49	86

The proposed IOWA and WOA along with Gradient Boosting Classifier (GBC) as a wrapper is used for feature selection and also measuring the accuracy while comparing with SVM. The experiments are conducted with same set of parameters and the results are tabulated in table 5 and 6. The proposed IWOA with GBC achieves best classification accuracy as 97.7% with minimal subset of features, which is 318. The accuracy is improved 3% when compared with the proposed IWOA with SVM linear kernel and the solution is converged in 29th iteration in GBC. The WOA also performs better with GBC than SVM classifier. The proposed IWOA with GBC achieves the best accuracy, error rate and convergence with minimum number of population size and number of iterations when compared with SVM linear kernel.

The figures 3 and 4 represent and error rate and the feature subset size of the proposed IWOA and WOA with various SVM kernels and GBC. The error rate is drastically reduced in IWOA with GBC than IWOA with SVM linear kernel. It shows that the GBC classifier is performing well than SVM. From the figure 4, it is noticed that the feature subset of both classifiers are consistent and also inferred that the feature sub set of the proposed IWOA is minimum than WOA. Hence, it is evident that the proposed IWOA with modified cross over and mutation operation yields minimum subset of features.

Generally the convergence of WOA algorithm is uncertain and is influenced by the combination of encircling prey and spiral updating position. The proposed IWOA converge to the global optima quickly with few iterations when compared with WOA. From the table 3, the proposed IWOA converges at 8^{th} and 4^{th} iteration itself for linear and polynomial kernel of order 3 respectively. Though it converges quickly, it produces a classification accuracy of 94% and 90% with the number of whales as 50 and 30 for the linear and polynomial kernel respectively. But, the linear kernel with 30 whales produces highest accuracy of 95% and converges in the 12^{th} iteration and is shown in the figure 5.

Table 5. Performance of IWOA with Gradient boosting classifier

Number of whales	10	30	50
Number of iteration	50	70	100
No. of trails	10	10	10
Gradient Boosting Classifier (GBC)			
Feature size	318	731	690
Accuracy (%)	97.7	95	89
Error rate	0.0247	0.05	0.11
Precision (%)	94	92	91
Recall (%)	99	96	89
Specificity (%)	97	96.09	91
AUC score	0.92	0.91	0.88
convergence	29	19	18

Table 6. Performance of WOA with Gradient boosting classifier

Number of whales	10	30	50
Number of iteration	50	70	100
No. of trails	10	10	10
Gradient Boosting Classifier (GBC)			
Feature size	639	380	456
Accuracy (%)	89.3	93	92.3
Error rate	0.11	0.07	0.0746
Precision (%)	92	91	91
Recall (%)	92	93	91
Specificity (%)	96	97	97.3
AUC score	0.85	0.89	0.89
convergence	30	26	5

Figure 3. Average Error rate of the proposed IWOA and WOA with SVM and GBC classifier

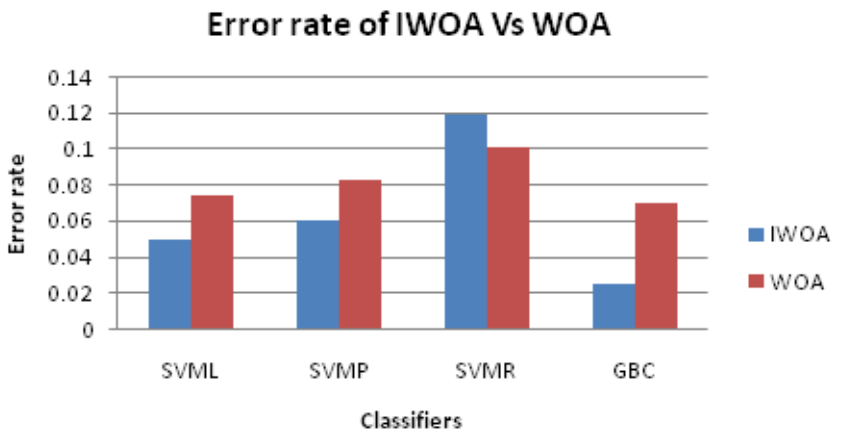


Figure 4. Average feature subset size of the proposed IWOA and WOA with SVM and GBC classifier

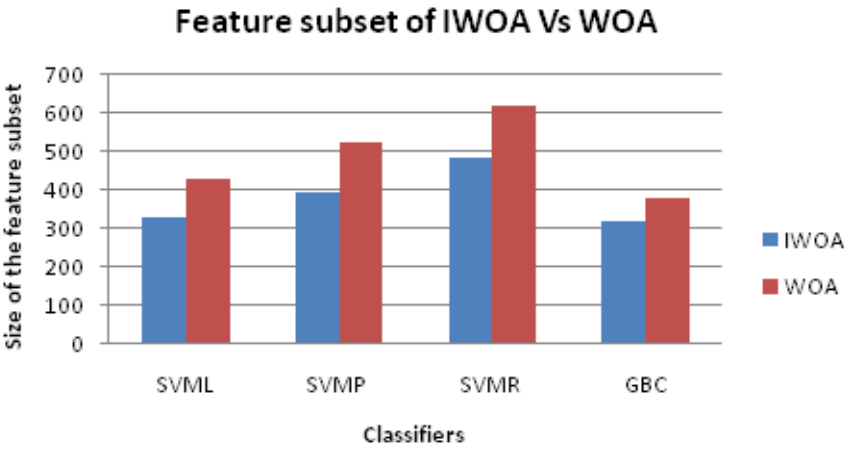
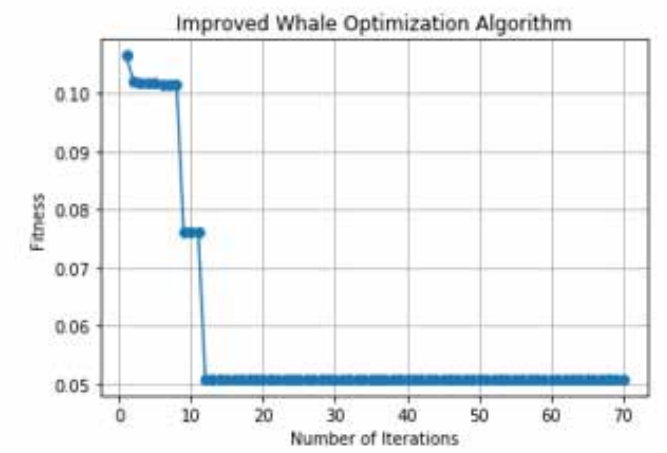


Figure 5. Convergence of the proposed IWOA with linear SVM kernel and population size as 30



The WOA converges at 49th iteration to produce a classification accuracy of 92.5% with the number of whales as 10 and is shown in figure 6. The proposed IWOA converges 6.125 times quickly when compared with WOA. Though the proposed IWOA with gradient boosting classifier produces high accuracy as 97.7%, it converges at 29th iteration, which is 27.59% slower than the proposed IWOA with SVM linear kernel. The convergence of proposed IWOA with GBC is shown in figure 7. The WOA with GBC classifier combination converges unexpectedly in the 5th iteration with number of whales as 50 and maximum number of iteration as 100, which is shown in figure 8. But the accuracy is 92.3%.

The comparative analysis of the convergence criteria alone of the proposed IWOA and WOA with SVM and GBC classifier without considering the accuracy, feature size, number of whales and number of iteration are shown in the figure 9.

The proposed IWOA is compared with Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Firefly Algorithm (FM) with GBC along with parameters as number of instances, number

Figure 6. Convergence of the WOA with the population size as 10 for linear SVM kernel

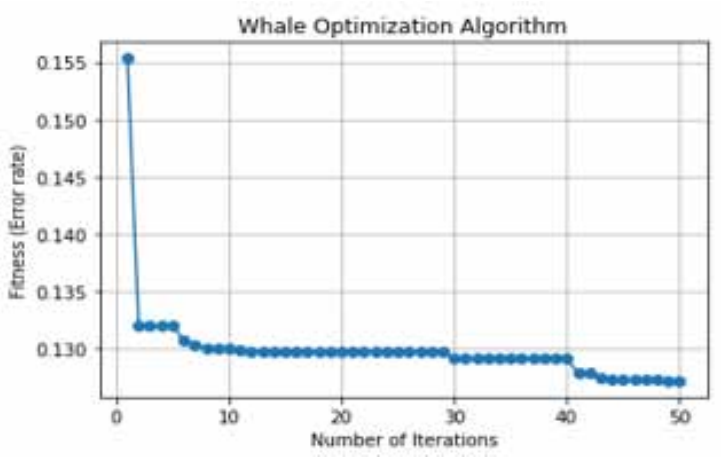


Figure 7. Convergence of the proposed IWOA with GBC

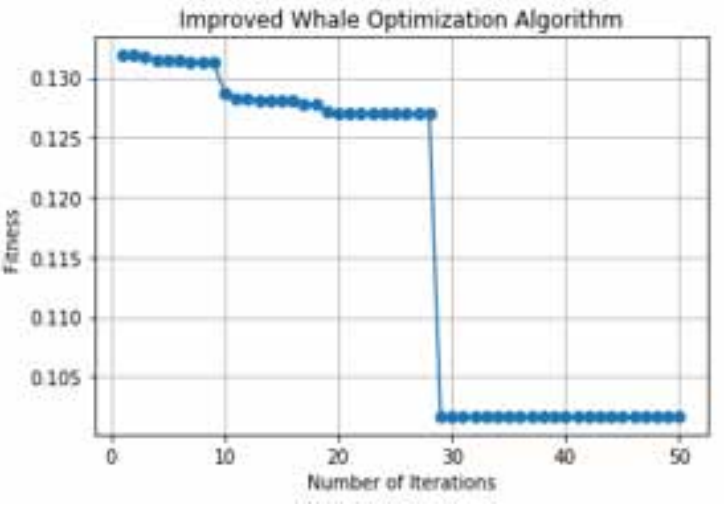


Figure 8. Convergence of WOA with GBC

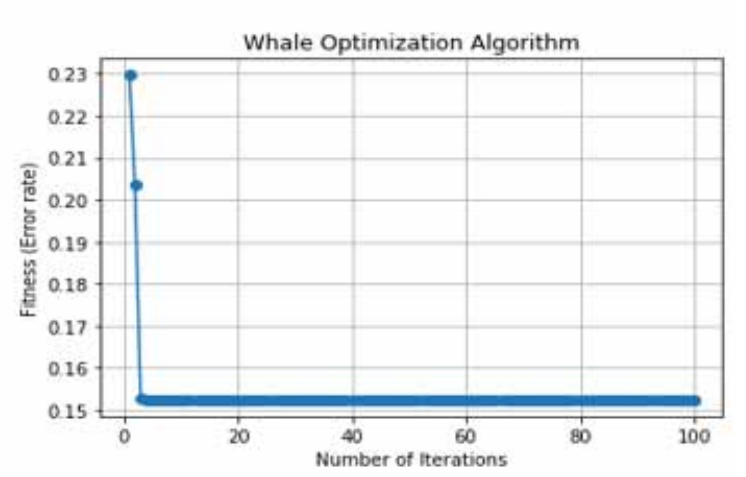
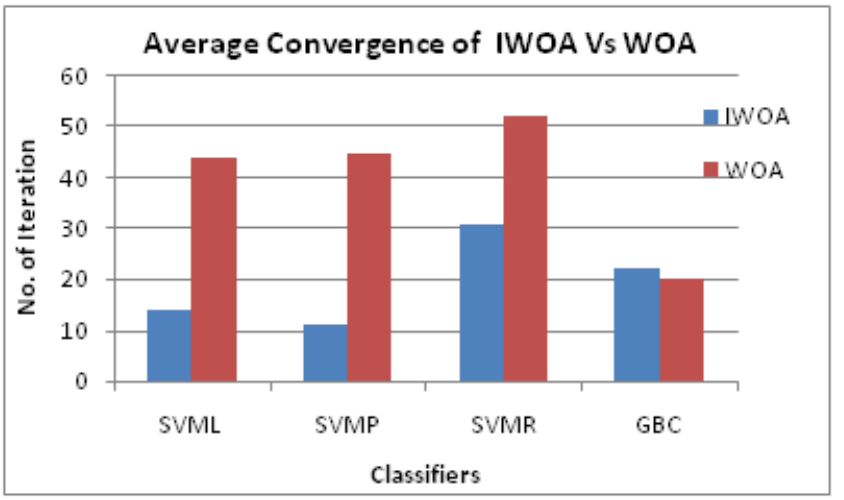


Figure 9. Average convergence of Proposed IWOA and WOA with SVM and GBC classifier



of iterations and number of trails as 10,50,10 respectively. The obtained results are shown in table 7. It is evident from the table that, the proposed IWOA significantly outperforms with other methods and produces better classification accuracy. The convergence and feature size is also minimum when compared with other methods.

The experiments are also conducted to evaluate model accuracy with validation data. The validation dataset excluding the class label is given to the model (IWOA + SVM) and the obtained performance measures are presented in table 8. The validation accuracy of the proposed IWOA with SVM linear kernel is 94% with error rate as 0.06. The validation accuracy of the proposed IWOA with GBC is given in table 9. From the table 9, it in inferred that, the validation accuracy of IWOA with GBC is 96% which is nearly equal to the training accuracy. It is realized from the table that, the obtained model is better to balance between the over fitting and under fitting. This clearly depicts that the proposed IWOA with GBC performs good to attain the minimal subset of

Table 7. Performance comparison of the proposed IWOA with other methods using GBC

	Accuracy (%)	Precision (%)	Recall (%)	Convergence	Feature Size
PSO	92	96	92	47	3565
GA	90.4	88	92	42	1034
FA	93.2	94	96	49	1125
Proposed IWOA	97.7	94	99	29	318

features for the breast cancer classification. The experimental result shows that, the model exhibits better generalization ability and produces the accuracy as 96% for the proposed IWOA with GBC which is as equal as the training data accuracy of the same approach. It is noticed from the table 8 and 9 that the AUC score is low when compared with training data for proposed IWOA with SVM and GBC respectively. It is inferred that, the minimum values of AUC score such as 0.77 and 0.81 for the proposed IWOA with SVM and GBC respectively is due to imbalanced validation data set and the increase the validation accuracy to training accuracy by applying appropriate method to balance the dataset in future.

6. CONCLUSION

In this paper, an Improved Whale Optimization Algorithm (IWOA) is proposed for feature selection with micro array gene expression data for breast cancer prediction. The proposed IWOA algorithm utilizes the modified crossover and mutation operators for better exploration and exploitation of the optimum search. The proposed IWOA incorporates the wrapper based approach, in which SVM and GBC classifiers are used to evaluate the fitness of the solutions. The experimental result shows that, the proposed IWOA with GBC produces 97.7% classification accuracy for the training

Table 8. Validation accuracy of the proposed IWOA algorithm with SVM Linear kernel

Metrics	Performance
Accuracy (%)	94
Error rate	0.06
Precision (%)	96
Recall (%)	100
AUC score	0.77

Table 9. Validation accuracy of the proposed IWOA with GBC

Metrics	Performance
Accuracy (%)	96
Error rate	0.04
Precision (%)	98
Recall (%)	100
AUC score	0.81

dataset and 96% for validation dataset with minimum subset of features and high specificity when compared with classical WOA and SVM classifier. It is identified from the result that, the classic WOA and the proposed IWOA both perform well for the SVM linear kernel than polynomial of order 3 and RBF kernels. The proposed IWOA obtains the optimal feature subset between 20 to 30 iterations. The AUC score of the proposed IWOA with SVM and GBC are relatively low value for the validation dataset which is 0.77 and 0.81 respectively. This is due to the improper distribution of class data and this imbalance issue is to be rectified in future. The proposed IWOA is compared with PSO,GA and FA and the proposed IWOA produces higher classification accuracy than the existing method.

REFERENCES

Abdelnabi, M. L. R., Jasim, M. W., Bakry, H. M. E. L., Taha, M. H. N., & Khalifa, N. E. M. (2020). Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. *Symmetry*, *408*(12), 1–15. doi:10.3390/sym12030408

Aldryan, D., Adiwijaya, P., & Annisa, A. (2018). Cancer Detection Based on Microarray Data Classification with Ant Colony Optimization and Modified Back Propagation Conjugate Gradient Polak-Ribiére. *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA),* (*vol. 1*, pp. 13 – 16). Semantic Scholar.

Alshamlan, H., Badr, G., & Alohali, Y. (2015). mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling. *BioMed Research International*, *2015*, 1–15. doi:10.1155/2015/604910 PMID:25961028

Belciug, S. (2020). Logistic Regression Paradigm for Training a Single - Hidden Layer Feed Forward Neural Network. Application to Gene Expression Datasets for Cancer Research. *Journal of Biomedical Informatics*, *102*(103373), 1–8. PMID:31901506

Dumitrescu, R. G., & Cotarla, I. (2005). Understanding Breast Cancer Risk - Where Do We Stand in 2005? *Journal of Molecular and Cellular Medicine*, *9*(1), 208–221. doi:10.1111/j.1582-4934.2005.tb00350.x PMID:15784178

El Nabi, A. M. L. R., Jasim, W., El Bakry, H. M., Taha, H. N., & Khalifa, N. E. M. (2020). Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. Symmetry, *12*(3), 408.

GEO. (n.d.) Gene Expression Onmibus. NCBI. https://www.ncbi.nlm.nih.gov/geo/

Gunasekhar, P., & Vijayalakshmi, S. (2020). Optimal Biomarker Selection using Adaptive Social Ski-Driver optimization for Liver Cancer Detection. *Biocybernetics and Biomedical Engineering*, *40*(4), 1611–1625. doi:10.1016/j.bbe.2020.10.005

Kundu, R., Chattopadhyay, S., Cuevas, E., & Sarkar, R. (2022, May). AltWOA: Altruistic Whale Optimization Algorithm for Feature Selection on Microarray Datasets. *Computers in Biology and Medicine*, *144*(105349), 105349. doi:10.1016/j.compbiomed.2022.105349 PMID:35303580

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., & Gao, Z. (2017). A Hybrid Feature Selection Algorithm for Gene Expression Data Classification. *Neurocomputing*, *256*, 56–62. doi:10.1016/j.neucom.2016.07.080

Maniruzzaman, M., Jahanur Rahman, M., Ahammed, B., Abedin, M. M., Suri, H. S., Biswas, M., El-Baz, A., Bangeas, P., Tsoulfas, G., & Suri, J. S. (2019). Md. J. Rahman, B. Ahammed, Md. M. Abedin, H. S. Suri, M. Biswas, A.E.Baz, P.Bangeas, G. Tsoulfas and J. S. Suri, " Statistical Characterization and Classification of Colon Microarray Gene Expression Data using Multiple Machine Learning Paradigms. *Computer Methods and Programs in Biomedicine*, *176*, 173–193. doi:10.1016/j.cmpb.2019.04.008 PMID:31200905

Mirjalili, S., & Lewis, A. (2016, May). The Whale Optimization Algorithm. *Advances in Engineering Software*, *95*, 51–67. doi:10.1016/j.advengsoft.2016.01.008

Mishra, P., & Bhoi, N. (2021). Cancer Gene Recognition from Microarray data with Manta Ray based enhanced ANFIS Technique. Biocybernetics and Biomedical Engineering, 41, 916 – 932.

Nandagopal, V., Geeitha, S., Vinoth Kumar, K., & Anbarasi, J. (2019). Feasible Analysis of Gene Expression – a Computational based Classification for Breast Cancer. *Measurement*, *140*, 120–125. doi:10.1016/j.measurement.2019.03.015

Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A Nested Genetic Algorithm for Feature Selection in High-dimensional Cancer Microarray Datasets. *Expert Systems with Applications*, *121*, 233–243. doi:10.1016/j.eswa.2018.12.022

Sharma, A., & Rani, R. (2019). C-HMOSHSSA: Gene Selection for Cancer Classification using Multi-objective Meta heuristic and Machine Learning Methods. *Computer Methods and Programs in Biomedicine*, *178*, 219–235. doi:10.1016/j.cmpb.2019.06.029 PMID:31416551

Shukla, A. K., Singh, P., & Vardhan, M. (2019). A New Hybrid Wrapper TLBO and SA With SVM Approach for Gene Expression Data. *Information Sciences*, *503*, 238–254. doi:10.1016/j.ins.2019.06.063

Shukla, A. K., Singh, P., & Vardhan, M. (2020). Gene selection for cancer types classification using novel hybrid metaheuristics approach. *Swarm and Evolutionary Computation*, *54*(100661), 1–16. doi:10.1016/j.swevo.2020.100661

Shukla, A. K., & Tripathi, D. (2019). Identification of Potential Biomarkers on Microarray Data using Distributed Gene Selection Approach. *Mathematical Biosciences*, *315*(108230), 1–15. doi:10.1016/j.mbs.2019.108230 PMID:31326384

Tahmouresil, , ARashedi, , EYaghoobi, , MRezaei, , M. (2022). Gene Selection using Pyramid Gravitational Search Algorithm. *PLoS One*, *17*(3), 1–15.

Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). Whale Optimized Mixed Kernel Function of Support Vector Machine for Colorectal Cancer Diagnosis. *Journal of Biomedical Informatics*, *92*(103124), 1–11. doi:10.1016/j.jbi.2019.103124 PMID:30796977