ISSN 2228-9860 eISSN 1906-9642 CODEN: ITJEA8

International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

Simultaneous Feature and Sample Selection Using Ensemble Multi-objective Search Space Enhanced Modified Whale Optimization Algorithm

M. Sathya[1*], S. Manju Priya[1]

[1] Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, INDIA.
**Corresponding Author (Email: sathya22joy@gmail.com).*

Paper ID: 12A6C

Volume 12 Issue 6

Received 07 January 2021
Received in revised form 15 March 2021
Accepted 23 March 2021
Available online 26 March 2021

Keywords:
Microarray data analysis; F1-score; SMWOA; Feature Selection; Sample Selection; SFSS; Biological-inspired algorithm; Whale Optimization Algorithm (WOA); Levy Flight (LF); Pareto optimal problem; Cancer detection; Leukemia dataset; Support vector machine (SVM); Lymphoma, SFSS-EMSMWOA; Prostate; Lung cancer.

Abstract

An effective method called Ensemble of Multi-objective Search space enhanced Modified Whale Optimization Algorithm (EMSMWOA) was used to solve the high dimensionality reduction problem in microarray data classification. In EMSMWOA, multiple SMWOA, evidential reasoning approach, and ensemble algorithm were utilized to choose the most important features, choose the optimal solution for feature selection from the Pareto-optimal set and generate optimal features respectively. The selected features were processed in different classifiers for microarray cancer detection. This work focuses on the computational complexity of classifiers also cut down by selecting features and samples of microarray data simultaneously using SMWOA. It also enhances the accuracy classification. After choosing the most relevant features and samples in every iteration, the Pareto optimal problem because of using the multiple objectives in SMWOAs is solved by applying the evidential reasoning approach. Finally, the best feature and sample subset are selected by applying the ensemble algorithm among multiple SMWOA that has various sizes of population and maximum iteration count. The results of SMWOA are ensemble depends on the mutual information in-between feature, sample, and class. The preferred features and samples are given as input to different classifiers for microarray cancer detection. The experiment proves the proposed SFSS-EMSMWOA method is improved on the basis of accuracy, precision, specificity, sensitivity, F1-score and also with average error than EMSMWOA in four different datasets for microarray data classification.

Disciplinary: Computer Sciences & Knowledge Management Systems, Medicine & Health Technology, Biomedical & Bioinformatics.

Cite This Article:

Sathya. M., Priya, S. M. (2021). Simultaneous Feature and Sample Selection Using Ensemble Multi-objective Search Space Enhanced Modified Whale Optimization Algorithm. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies*, *12*(6), 12A6C, 1-14. http://TUENGR.COM/V12/12A6C.pdf DOI: 10.14456/ITJEMAST.2021.108

1 Introduction

Functional genomics requires extracting knowledge from vast collections from different biological studies (Dash, 2020). One form of the large-scale experiment includes simultaneously tracking the expression degree of thousands of genes in a single disease. The work focuses on the analysis of gene expression. The possibilities of microarray technologies as well as the volume of data or inputs produced are immense. The technology of microarray is one of several biologists' essential methods for tracking large-scale genome expression. The latest work has shown that the classification of microarray data methods was used for the diagnosis of diseases. These statistics are derived and obtained in the context of variations in gene expression from tissue samples.

The huge volume of research evidence poses several problems for the scientist to retrieve valuable knowledge by utilizing conventional methods in data mining. Very often, these data are asymmetrical. The volume of genes (or features) is millennium but the total sampling is generally lesser or considerably over than a hundred. Such analysis with these aspects reduces the performance of the classifiers and boosts the computational cost. Therefore, traditional classifiers on asymmetrical data are extremely difficult to use. So, dimension reduction is essential for microarray data analysis. Techniques of feature selection are put in use for reduction of dimensionality (Peng et al., 2010; Kumar et al., 2015) of microarray data.

Particle Swarm Optimization (PSO), Sathya et al. (2019) discussed a technique of feature selection which was used to decrease the dimensionality and the most relevant features are selected by microarray data. The features selected are given as input to the classifiers Naive Bayes (NB). Support Vector Machine (SVM) is used for the classification of microarray data. However, sometimes PSO has a slow convergence problem. A Modified Whale Optimization Algorithm (MWOA) (Sathya et al., 2020) was proposed one choose the most relevant features from microarray data. Even though, the MWOA improved in terms of dimensionality reduction in the microarray cancer dataset, the MWOA will get struck into local optima and degrades the cancer detection accuracy.

To resolve this issue and to equalize the exploration and also exploitation abilities of MWOA, Search space enhanced MWOA (SMWOA) (Dancey et al., 2012) was proposed where non-linear dynamic strategy and Levy-flight strategy were used for feature selection. But, SMWOA may trap in the Pareto-optimal problem because of using multiple objectives in it. This problem was resolved by proposing the Ensemble of Multi-objective SMWOA (EMSMWOA) (Kuo et al., 2004) method.

In this article, microarray data dimensionality is cut down and the efficiency of the classifier is enhanced further by selecting the features and samples simultaneously using EMSMWOA. The Pareto optimal problem during the selection of features and samples of microarray data is solved by applying the evidential reasoning approach in the selected features and samples. After the selection of optimal features and samples, the ensemble algorithm is processed to select the final feature and sample subset based on the mutual findings. Finally, features selected along with

samples are given as input to the different classifiers used for microarray detection of cancer. Hence, the simultaneous selection of features and samples from microarray data degrades the computational complexness of classifiers and also enhances the accuracy classification. This whole process is named Simultaneous Feature and Sample Selection using EMSMWOA (SFSS-EMSMWOA).

2 Literature Review

Bolón-Canedo et al. (2015) developed a feature selection method to classify the microarray data. Based on a vertical distribution, the microarray data were distributed by features. After the distribution of the data, a merge procedure is performed that updates the feature subset based on the enhancement in the classification accuracy. This method is appropriate only for huge datasets.

Bonilla-Huerta et al. (2015) did a hybrid framework to select appropriate features and classify the DNA microarray data. The hybrid framework used different classifiers for microarray data classification. And also computational complexity of this framework is high.

Mollaee et al. (2016) proposed an ensemble schema based on ensemble schema for the classification of microarray data. SVM was used to categorize the mapped features. However, the proper choice of the kernel function in SVM is more difficult.

Tang et al. (2016) proposed a feature selection approach based on microarray data according to the mutual information. This method is based on two strategies relevant to hiking and feature interaction improves microarray data classification. However, sometimes the computational complexity of this approach is high. Seijo-Pardo et al. (2016) presented an ensemble for feature selection that combines feature ranking. The individual rankings are integrated with various aggregation methods and a working subset of features chooses a data complexity means that is inverse of Fisher discriminant ratio. However, the performance is influenced by a threshold value.

Sun et al. (2016) presented a global dimensionality reduction method for microarray data. This method is processed according to the semi-definite programming model that minimizes the redundancy in the feature by using the quadratic programming model also maximizing relevant feature. In a semi-definite programming model, every feature had one constraint order that restricts the fitness or wellness function of the feature selection problem. The Lagrange multiplier is used for measurement proxy which finds the relevant features and these features are utilized in different classifiers, these features are utilized to classify or divide the microarray data. However, the strike of this method decides the threshold value selection process.

Wang et al. (2017) presented a feature selection method for the identification of cancer from microarray gene expression data. Despite it has confined by the demand to find on search space that is fit for better classification accuracy beyond past learning of datasets. Ke et al. (2018) used Score-based Criteria Fusion feature selection (SCF) for gene microarray data. SCF has combined ranking methods of two features through the evaluation of association among classes together with features. The selected features are processed in SVM and K-Nearest Neighbor (KNN) for cancer prediction. However, the assignment of the weighted parameter in SCF is a complex task.

Ebrahimpour et al. (2018) proposed a technique for high dimensional microarray data classification. The cooperating coevolution technique for large-scale feature selection (CCFS) algorithm split the dataset vertically in an unplanned way. With the principal theory of coevolution, the solution space has gained with a filter criterion in the objective function via a binary gravitational search algorithm. However, the computational complexity is high.

Brankovic et al. (2018) introduced a distributed feature selection algorithm according to the distance correlation for the classification of microarray data. It uses an approach called distance correlation (dCor) as a principle of dependency of estimated class in feature subset. However, this algorithm has a slow convergence problem. Kang et al. (2019) presented methods to choose the features and classify the tumor. Initially, the collected tumor dataset was normalized using a z-score. Relaxed Lasso and tumor multi-class support vector machines were used for feature selection in discriminative feature genes. However, it is not more suitable for the sparse dataset.

Wang et al. (2019) planned a method for feature selection and classification of microarray gene expression cancers. Along with the uniqueness of the planned method, multiple dimensionalities of the population were designed which can locate the feature selection issue and do not need the desired count of features. However, the computational time of BCO-MDP is high. (Yan et al. 2019) used a feature selection algorithm called Binary Coral Reef Optimization (BCRO) to select the most important features from the microarray dataset. Along with exploitation, and exploration BCRO would be a way enhanced in terms of fusing with strategies based on local search or Algorithm of swarm intelligence.

Cao et al. (2019) built a multi-objective feature selection model via distributed parallel algorithm. But, there may be chances for raising the conflicts between the multiple objectives. Zare et al. (2019) planned a supervised feature selection algorithm using Singular Value Decomposition (SVD) and matrix factorization for classification of microarray datasets. However, training for supervised learning requires a lot of computation time. Mazumder et al. (2019) proposed a method for the classification of microarray cancer data. It continued by taking the feature-feature redundancy.

3 Proposed Methodology

The area is planned with SFSS-EMSMWOA is described in detail for microarray cancer detection. The overflow of this proposed work is shown in Figure 3.

3.1 Simultaneous Feature and Sample Selection from the Microarray Dataset

Considering a crossover operator, the Quadratic Interpolation (QI) selects the best search agents for feature selection $X_f^* = (x_{1f}^*, x_{2f}^*, \dots x_{nf}^*)$ and other two parents $A_f = (a_{1f}, a_{2f}, \dots a_{nf})$,$B_f = (b_{1f}, b_{2f}, \dots b_{nf})$ and then generate a new solution $X_f = (x_{1f}, x_{2f}, \dots x_{nf})$ to choose the most relevant features from a microarray dataset. At the same time, assume a crossover operator QI selects the best search agents for sample selection $X_s^* = (x_{1s}^*, x_{2s}^*, \dots x_{ns}^*)$ and other two parents $C_S =$

$(c_{1s}, c_{2s}, \dots c_{ns})$,$D_s = (d_{1s}, d_{2s}, \dots d_{ns})$ and then generate a new solution $X_s = (x_{1s}, x_{2s}, \dots x_{ns})$ to choose the most relevant features from a microarray dataset. The new solution for feature selection and sample selection are given as

$$x_{if} = 0.5 \times \frac{\left(a_{if}^2 - b_{if}^2\right) \times f\left(X_f^*\right) + \left(b_{if}^2 - a_{if}^2\right) \times f(A_f) + \left(x_{if}^{*2} - a_{if}^2\right) \times f(B_f)}{(a_i - b_i) \times \left(b_i^2 - x_i^2\right) + \left(b_i^2 - a_i^2\right) \times f(A) + \left(x_i^{*2} - a_i^2\right) \times f(B)}, \forall i = 1,2, \dots n \quad (1),$$

$$x_{is} = 0.5 \times \frac{\left(c_{is}^2 - d_{is}^2\right) \times f(X_S^*) + \left(d_{is}^2 - c_{is}^2\right) \times f(C_S) + \left(x_{iS}^{*2} - c_{is}^2\right) \times f(D_S)}{(c_{is} - d_{is}) \times \left(d_{is}^2 - x_{is}^2\right) + \left(d_{is}^2 - c_{is}^2\right) \times f(C_S) + \left(x_{is}^{*2} - c_{is}^2\right) \times f(D_S)}, \forall i = 1,2, \dots n \quad (2).$$

In Equations (1) and (2), $f\left(X_f^*\right), f(A_f)$ and $f(B_f)$ are the fitness values for feature selection at X_f^*, A_f and B_f respectively, $f(X_S^*), f(A_s)$ and $f(B_s)$ are the fitness values for sample selection at X_S^*, A_s and B_s correspondingly and i denotes the i-th dimension. The fitness value of feature selection is

$$f\left(X_f\right) = IM\left(X_f\right) + cardinality\left(X_f\right)\ accuracy\left(X_f\right) \quad (3).$$

The fitness value of sample selection is given as

$$f(X_s) = IM(X_s) - cardinality(X_s) + \quad accuracy(X_s) \quad (4).$$

In Equations (3) and (4), $IM\left(X_f\right)$ is the information measure of each feature in the new solution X_f, $IM(X_s)$ is the information measure of each sample in the new solution X_s, $cardinality\left(X_f\right)$ is the proportion of the total features, there in the dataset to the features present in the subset, $cardinality\left(X_f\right)$ is the proportion of the total samples present in the dataset to the samples present in the subset, $accuracy\left(X_f\right)$ is the classifier accuracy on a classifier with the features selected and $accuracy\left(X_f\right)$ is the accuracy of the classifier is done by the select samples.

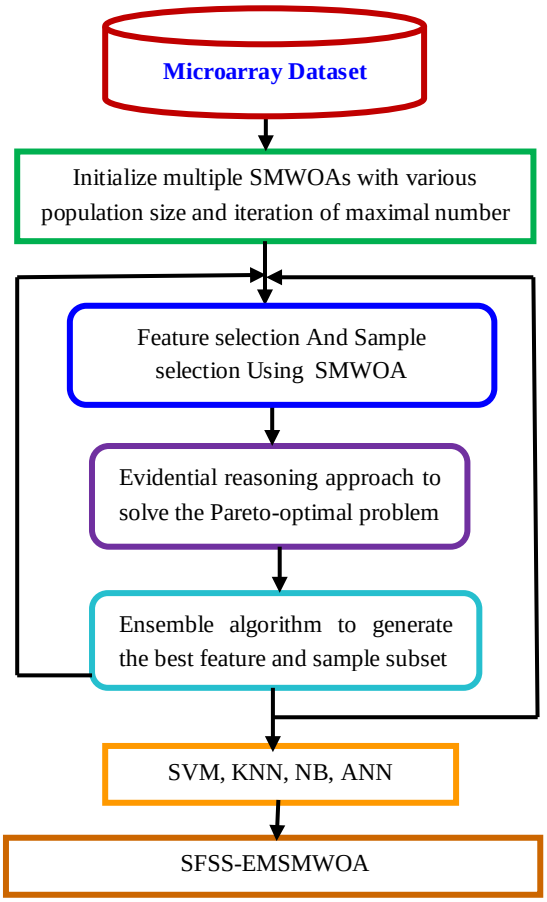

Figure 1: Overflow of SFSS-EMSMWOA.

The current best search agent $R_f^*(R_s^*)$ for feature selection (sample selection) in the quadratic crossover considers a leading role, to discover the global optimum solution for feature selection is done by the search agents and sample selection. The implementation of QI is in the exploitation phase to preserve the population diversity by enhancing the exploitation capability of SMWOA. Quadratic crossover and the spiral-shaped path are two components exploitation phase of SMWOA. To tradeoff between two components, a uniformly distributed parameter is used, the spiral-shaped path method is used when the probability is less than 0.6 for feature and sample selections are

$$X_f(t+1) = D_f' \times e^{bl} \times \cos(2\pi l) + X_f^*(t) \quad (5),$$

$$X_s(t+1) = D_s' \times e^{bl} \times \cos(2\pi l) + X_s^*(t) \quad (6).$$

In Equations (5) and (6),

$$D_f' = \left|X_f^*(t) - X_f(t)\right| \quad (7),$$

$$D_s' = \left|X_s^*(t) - X_s(t)\right| \quad (8).$$

In Equations (7) and (8), D_f' means the distance among the i-th whale along with the best solution (optimal features) got so far if there is a better solution, each iteration will be updated, D_s' shows the distance between i-th whale along with the best solution (optimal samples) attained, the update will be done in each iteration until a better solution is found, t -> current iteration, $|.|$ -> the absolute value operation, × gives an element-by-element multiplication, the constant is describing by $b$the shape of a logarithmic spiral and l is a random number which ranges from -1 to 1. The position of the whales is updated by quadratic crossover when the probability is greater than 0.6.

SMWOA uses Levy Flight (LF) that is used to avoid or escape from local optima problems by promoting the diversity of the population. A step size escaped the LF jumping of the design domain is changed. It is described as

$$Levy_f = random\left(size\left(D_f\right)\right) \oplus L(\beta) \sim \frac{0.01}{|v|^{\frac{1}{\beta}}\left(X_{if} - X_f^*\right)} \quad (9),$$

$$Levy_s = random(size(D_s)) \oplus L(\beta) \sim \frac{0.01}{|v|^{\frac{1}{\beta}}(X_{is} - X_s^*)} \quad (10).$$

In Equations (9) and (10), $random(.)$ denotes the random function, $size\left(D_f\right)$ denotes the scale of the feature selection problem, $size(D_s)$ denotes the scale of the sample selection problem, $L(\beta)$ is the Levy distribution and β -> index. Every whale position is renewed by

$$X_f(t+1) = X_f(t) + \frac{1}{sqrt(t)} \times sign(random - 0.5) \oplus Levy_f \quad (11),$$

$$X_s(t+1) = X_s(t) + \frac{1}{sqrt(t)} \times sign(random - 0.5) \oplus Levy_s \quad (12).$$

In Equations (11) and (12), $1/sqrt(t)$ denotes a parameter related to the current iteration number t, also $sqrt(t)$ represents the operation of square root. The exploration phase of SMWOA is expressed as

$$X_f(t+1) = \begin{cases} X_f(t) + \frac{1}{sqrt(t)} \times sign(random - 0.5) \oplus Levy_f, & if\ p < 0.5 \\ D'_f \times e^{bl} \times cos(2\pi l) + X^*_f(t), & if\ p \geq 0.5 \end{cases} \quad (13),$$

$$X_s(t+1) = \begin{cases} X_s(t) + \frac{1}{sqrt(t)} \times sign(random - 0.5) \oplus Levy_s, & if\ p < 0.5 \\ D'_s \times e^{bl} \times cos(2\pi l) + X^*_s(t), & if\ p \geq 0.5 \end{cases} \quad (14).$$

A non-linear control parameter is used to control a perfect harmony among exploration and also exploitation. The presence of multiple objectives generally gives rise to a family of non-dominated solutions called the Pareto-optimal solution. It can be solved by a final solution set that considers various objective functions such as specificity, sensitivity, Area Under Curve (AUC), and relative distance for effective feature selection and sample selection. To achieve high classification accuracy, an ensemble algorithm is applied in which the results of multiple SMWOA are ensembled using mutual information between feature, samples, and class. When common features and samples are chosen by all SMWOA that features and samples are chosen without using a greedy search algorithm and entered into the optimal feature and sample subset. The method measures feature-class, feature-feature, sample-class, sample-sample mutual information and chooses features which has maximal feature-class mutual information and minimal feature-feature mutual information. Samples are selected which have maximum sample-class mutual information and minimal sample-sample mutual information.

Pseudo code of SFSS-EMSMWOA

Step 1: The whale population is initialized by $X_i(i = 1,2,3,\dots n)$, max_{itr}.

Step 2: Each whale randomly selects the features and samples of microarray dataset.

Step 3: Compute the fitness of each search agent using Equations (3) and (4).

Step 4: Assign the best search agent to X^*_f and X^*_s

Step 5: while $(t < max_{itr})$

Step 6: if $(p_1 < 0.5)$

Step 7: if $(|T| < 1)$ // T is the vector coefficient

Step 8: The current search agent position is updated using Equations (11) and (12).

Step 9: else if $(|T| \geq 1)$

Step 10: Select random search agents X_{rand_f} and X_{rand_s} for feature selection and sample selection

Step 11: The current search agent position is updated by Equation (15) and Equation (16)

$X_f(t+1) = X_{rand\_f} - T \times D_f$ (15)

$X_s(t+1) = X_{rand\_s} - T \times D_s$ (16)

Step 12: end if

Step 13: else if $(p_2 \geq 0.6)$

Step 14: The current search agent position is updated by Equations (5) and (6).

Step 15: else if $(p_2 \geq 0.6)$

Step 16: The current search agent position is updated by Equations (1) and (3).

Step 17: end if

Step 18: end if

Step19: end for

Step 20: The fitness of each search agent is computed by using Equations (3) and (4).

Step 21: Update X_f^* and X_s^* if there is a better solution

Step 22: $t++$

Step 23: end while

Step 24: return X_f^* and X_s^*

Step 25: Process N number of SMWOA with different populations and a maximum number of iteration to choose the optimal features and samples from the microarray dataset.

Step 26: Process evidential reasoning approach to choose the optimal solution for feature selection and sample selection from the Pareto-optimal set.

Step 27: Choose the final optimal solution from each SMWOA based on the final solution set.

Step 28: Select the optimal features and samples based on the ensemble algorithm.

Step 29: Process the results of an ensemble algorithm in SVM, KNN, NB, and artificial neural network (ANN) for cancer detection.

Step 30: It has betterment results while using the SFSS-EMSMWOA ensemble algorithm for cancer detection.

4 Result and Discussion

The working of EMSMWOA and SFSS-EMSMWOA is tested in charge of accuracy, precision, specificity, sensitivity, F1-score, and average error. For experimental purposes, four datasets such as Leukemia, Lymphoma, prostate, and lung cancer microarray datasets are used. The leukemia dataset consists of 72 instances, 3572 features, and 2 classes, lymphoma dataset consists of 77 instances, 2647 features, and 2 classes. The prostate dataset has 102 cases, 2135 features, and 2 classes. The lung cancer dataset consists of 32 instances, 56 attributes, and 2 classes. From the collected data, 60% of data are used for testing and 40% are used for training datasets. EMSMWOA and SFSS-EMSMWOA MATLAB (2018a) are used for implementation and also runs on a Microsoft Windows 7 along with an Intel processor running at 2.70 GHz and 4GB memory. Table 1 and Table 4.2 reveals the count of features and samples chosen by EMSMWOA also SFSS-EMSMWOA methods for four datasets respectively.

Table 1: Total count of features selected by EMSMWOA and SFSS-EMSMWOA.

Feature	No of features	EMSMWOA	SFSS-EMSMWOA
Leukemia	3572	35	30
Prostate	2135	100	92
Lymphoma	2647	28	21
Lung cancer	12533	130	123

Table 2: Number of samples selected by SFSS-EMSMWOA

Feature	No of samples	EMSMWOA
Leukemia	72	65
Prostate	77	68
Lymphoma	102	90
Lung cancer	32	23

4.1 Accuracy

Accuracy is determined by the fraction of instances that are classified correctly. It is calculated by the total amount of correctly predicted sick people (true positive) and correctly predicted healthy people (true negative) over the total number of classifications. It is calculated as

$$Accuracy = \frac{True\ Positive\ (TP)+False\ Negative\ (FN)}{TP+True\ Negative\ (TN)+False\ Positive(FP)+FN} \quad (15),$$

where TP is the sick people that are affected by cancer are classified as sick correctly, FP is the healthy people classified as sick incorrectly, TN is the healthy people classified as healthy correctly and FN is the sick people are classified as healthy incorrectly. Table 3 shows the accuracy of EMSMWOA and SFSS-EMSMWOA methods for microarray data classification.

Table 3: Accuracy comparison of EMSMWOA and SFSS-EMSMWOA

	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN	SFSS-EMSMWOA-SVM	SFSS-EMSMWOA-KNN	SFSS-EMSMWOA-NB	SFSS-EMSMWOA-ANN
Leukemia	62.42	86.13	82.34	88.42	65.76	90.24	86.48	91.34
Lymphoma	92.13	85.36	79.24	91.87	95.06	89.24	82.24	96.45
Prostate	94.15	85.67	86.98	96.14	96.76	89.67	90.34	97.58
Lung cancer	92.08	82.75	84.21	93.45	95.35	87.69	89.37	96.40

From Figure 2, it is observed that the accuracy of SFSS-EMSMWOA-ANN method is 46.33%, 6.05%, 10.93%, 3.3%, 38.9%, 1.22%, and 5.62% greater than EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN and SFSS-EMSMWOA-NB respectively for leukemia dataset. Hence, it is proven that the planned SFSS-EMSMWOA-ANN method gives better results in charge of accuracy when relating with EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN and SFSS-EMSMWOA-NB methods.

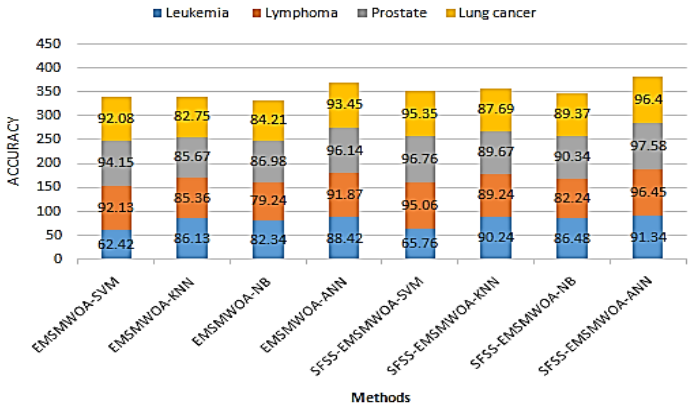

Figure 2: Accuracy comparison of EMSMWOA and SFSS-EMSMWOA.

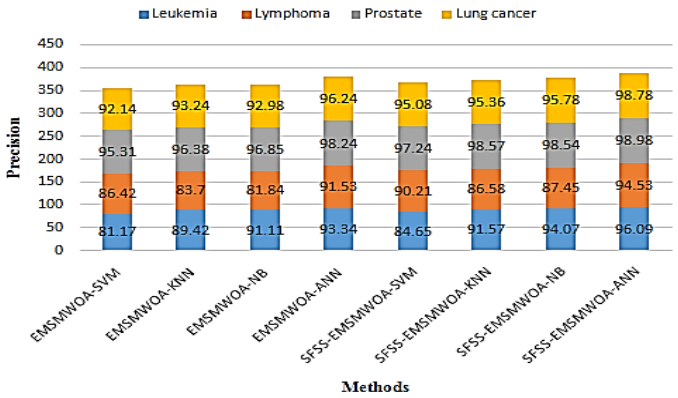

Figure 3: Precision Comparison of EMSMWOA and SFSS-EMSMWOA

4.2 Precision

The fraction of true positive instances which are classified as positiveis defined as precision. It is calculated from

$$Precision = \frac{TP}{TP+FP} \quad (16).$$

Table 4 shows the precision of EMSMWOA and SFSS-EMSMWOA method for microarray data classification.

Table 4: Precision Comparison of EMSMWOA and SFSS-EMSMWOA

	EMSM WOA-SVM	EMSM WOA-KNN	EMSM WOA-NB	EMSM WOA-ANN	SFSS-EMSM WOA-SVM	SFSS-EMSM WOA-KNN	SFSS-EMSM WOA-NB	SFSS-EMSMW OA-ANN
Leukemia	81.17	89.42	91.11	93.34	84.65	91.57	94.07	96.09
Lymphoma	86.42	83.7	81.84	91.53	90.21	86.58	87.45	94.53
Prostate	95.31	96.38	96.85	98.24	97.24	98.57	98.54	98.98
Lung cancer	92.14	93.24	92.98	96.24	95.08	95.36	95.78	98.78

The performance comparison between EMSMWOA and SFSS-EMSMWOA in terms of precision is shown in Figure 3. From the experimental results, the precision of SFSS-EMSMWOA-ANN is 18.38% greater than EMSMWOA-SVM, 7.46% greater than EMSMWOA-KNN, 5.47% greater than EMSMWOA-NB, 2.95% greater than EMSMWOA-ANN, 13.51% greater than SFSS-EMSMWOA-SVM, 4.94% greater than SFSS-EMSMWOA-KNN, and 4.94% greater than SFSS-EMSMWOA-NB for leukemia dataset. From this comparison, it results that the SFSS-EMSMWOA-ANN achieves high precision compared with other methods for four different datasets.

4.3 Specificity

The measurement of specificity is done by the proportion of actual negatives that identify people correctly without illness within all people which is termed as free from illness. It is calculated by using

$$Specificity = \frac{TN}{FP+TN} \quad (17).$$

Table 5 shows the specificity of EMSMWOA and SFSS-EMSMWOA methods for microarray data classification.

Table 5: Specificity Comparison of EMSMWOA and SFSS-EMSMWOA

	EMSM WOA-SVM	EMSM WOA-KNN	EMSM WOA-NB	EMSM WOA-ANN	SFSS-EMSM WOA-SVM	SFSS-EMSM WOA-KNN	SFSS-EMSM WOA-NB	SFSS-EMSMW OA-ANN
Leukemia	80.21	88.45	90.14	92.37	83.65	90.12	93.14	95.52
Lymphoma	85.46	82.73	80.89	90.56	89.18	85.46	84.15	93.79
Prostate	94.34	95.41	95.89	97.28	96.09	97.24	97.68	98.88
Lung cancer	93.14	94.63	94.74	96.09	95.06	96.91	97.02	98.98

In Figure 4, the effect of the method SFSS-EMSMWOA is analyzed by specificity comparison with EMSMWOA with different classifiers. The specificity of SFSS-EMSMWOA-ANN is 19.09%, 7.99%, 5.97%, 3.41%, 14.19%, 5.99%, and 2.56% greater than EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN and SFSS-EMSMWOA-NB respectively for leukemia dataset. This study has proved that the proposed SFSS-EMSMWOA-ANN method attains improved results in charges of specificity when compared to other methods.

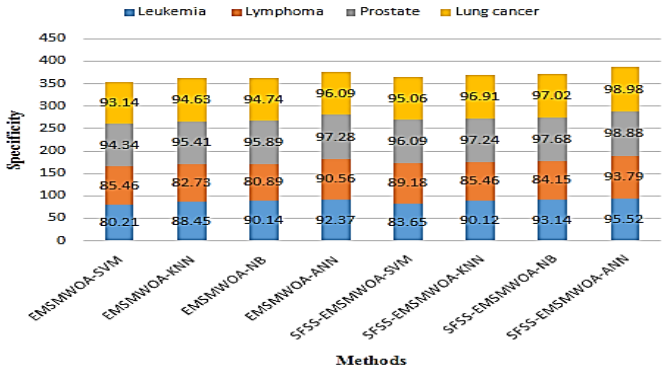

Figure 4: Specificity comparison of EMSMWOA and SFSS-EMSMWOA

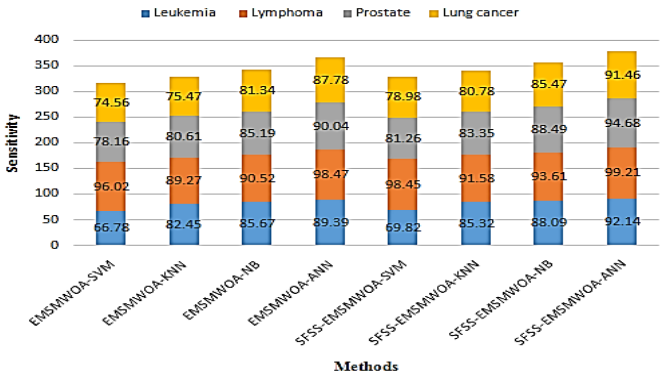

Figure 5: Sensitivity comparison of EMSMWOA and SFSS-EMSMWOA

4.4 Sensitivity

Sensitivity is measured by the fraction of actual positives that correctly identifies the people with illnesses. The following formula is used to calculate sensitivity,

$$Sensitivity = \frac{TP}{TP+FN} \quad (18).$$

Table 6 shows the sensitivity of EMSMWOA and SFSS-EMSMWOA method for microarray data classification.

Table 6: Sensitivity Comparison of EMSMWOA and SFSS-EMSMWOA

	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN	SFSS-EMSMWOA-SVM	SFSS-EMSMWOA-KNN	SFSS-EMSMWOA-NB	SFSS-EMSMWOA-ANN
Leukemia	66.78	82.45	85.67	89.39	69.82	85.32	88.09	92.14
Lymphoma	96.02	89.27	90.52	98.47	98.45	91.58	93.61	99.21
Prostate	78.16	80.61	85.19	90.04	81.26	83.35	88.49	94.68
Lung cancer	74.56	75.47	81.34	87.78	78.98	80.78	85.47	91.46

In Figure 5, the effect of this method SFSS-EMSMWOA is analyzed in terms of sensitivity by comparing it with EMSMWOA with different classifiers. The sensitivity of SFSS-EMSMWOA-ANN is 37.98%, 11.75%, 7.55%, 3.08%, 31.97%, 7.99% and 4.6% greater than EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN and SFSS-EMSMWOA-NB respectively for leukemia dataset. From this study, it is proved that the suggested method SFSS-EMSMWOA-ANN method gives better result its sensitivity obtained when it is compared with other methods.

4.5 F1-Score

F1-score is determined by harmonic average of precision and recall. It is calculated by

$$\text{F1-score} = \frac{2\times precision\times recall}{precision+recall} \quad (19).$$

Table 7 shows the F1-score of EMSMWOA and SFSS-EMSMWOA method for microarray data classification.

Table 7: F1-Score comparison of EMSMWOA and SFSS-EMSMWOA

	EMSM WOA-SVM	EMSM WOA-KNN	EMSM WOA-NB	EMSM WOA-ANN	SFSS-EMSM WOA-SVM	SFSS-EMSM WOA-KNN	SFSS-EMSM WOA-NB	SFSS-EMSMW OA-ANN
Leukemia	80.24	89.92	88.24	91.52	83.16	89.92	88.24	94.75
Lymphoma	93.18	89.24	84.68	95.02	96.45	89.24	84.68	95.02
Prostate	82.65	72.36	84.29	89.78	85.37	72.36	84.29	89.78
Lung cancer	79.09	69.78	81.47	85.47	83.39	69.34	81.47	87.67

From Figure 6, it is observed that the F1-score of the SFSS-EMSMWOA-ANN method is 18.08%, 5.37%, 7.38%, 3.53%, 13.94%, 5.37%, and 7.38%greater than EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN, and SFSS-EMSMWOA-NB respectively for leukemia dataset. Hence, it is proved that the planned method SFSS-EMSMWOA-ANN gives better results in terms of F1-score when compared to EMSMWOA-SVM, EMSMWOA-KNN, EMSMWOA-NB, EMSMWOA-ANN, SFSS-EMSMWOA-SVM, SFSS-EMSMWOA-KNN, and SFSS-EMSMWOA-NB methods.

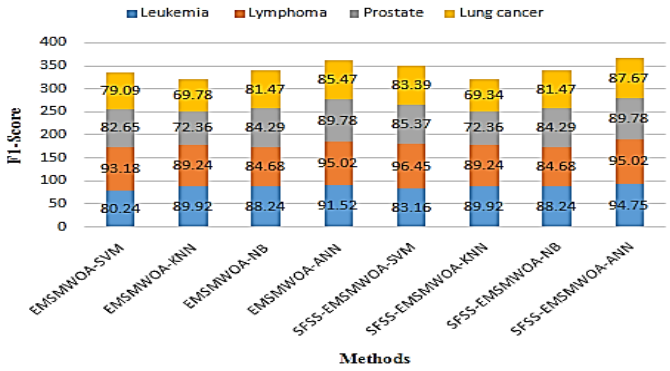

Figure 6: F1-score comparison of EMSMWOA and SFSS-EMSMWOA

4.6 Average Error

It is the average error of classifiers to classify the gene expression data with the selected features by EMSMWOA and SFSS-EMSMWOA. Table 8 shows the average error of classifiers that processed the selected features by SMWOA and EMSMWOA.

Table 8: Comparison of Average Error

No. of iteration	EMSMWOA	SFSS-EMSMWOA
100	0.075	0.064
200	0.059	0.051
300	0.04	0.024
400	0.03	0.018
500	0.025	0.012
600	0.025	0.012
700	0.025	0.012
800	0.025	0.012
900	0.025	0.012
1000	0.025	0.012

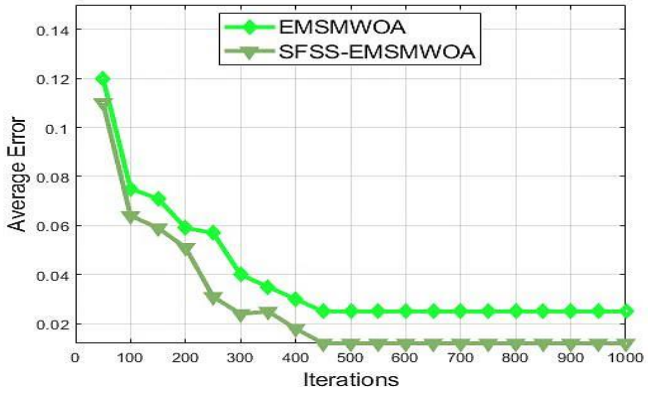

Figure 7: Comparison of Average Error.

In average error of EMSMWOA and SFSS-EMSMWOA with classifiers is shown in Figure 7. The number of iteration is taken in X-axis and the average error of feature selection methods is

shown in Y-axis. When the number of iteration is 200, the average error of SFSS-EMSMWOA is 13.56% less than SEMSMWOA. From this analysis, it is proved that the SFSS-EMSMWOA method has less average error than EMSMWOA for cancer detection.

5 Conclusion

In this paper, SFSS-EMSMWOA is proposed to further enhance the accuracyclassification and reduce the complexity in the computation of classifiers for microarray data classification. Initially, microarray data are collected and the most relevant features and samples are selected simultaneously using SMWOA. Because of using multiple objectives in the fitness function, the Pareto optimal problem is solved by applying the evidential approach. The multiple SMWOA are initialized with various size of population and maximal count of iteration and the results of multiple SMWOA are ensembled using mutual information of features, samples and lass. Finally the selected features and samples are processed in SVM, KNN, NB and ANN classifiers for microarray data classification. The experiment proves the proposed SFSS-EMSMWOA method is improved on the basis of accuracy, precision, specificity, sensitivity, F1-score and also with average error than EMSMWOA in four different datasets for microarray data classification.

6 Availability of Data and Material

Data can be made available by contacting the corresponding author.

7 References

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Distributed feature selection: An application to microarray data classification. *Applied soft computing*, 30, 136-150.

Bonilla-Huerta, E., Hernandez-Montiel, A., Morales-Caporal, R., & Arjona-Lopez, M. (2015). Hybrid framework using multiple filters and an embedded approach for an efficient selection and classification of microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(1), 12-26.

Brankovic, A., Hosseini, M., & Piroddi, L. (2018). A distributed feature selection algorithm based on distance correlation with an application to microarrays. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(6), 1802-1815.

Cao, B., Zhao, J., Yang, P., Yang, P., Liu, X., Qi, J., & Muhammad, K. (2019). Multiobjective feature selection for microarray data via distributed parallel algorithms. *Future Generation Computer Systems*, *100*, 952-981.

Dancey, J.E, Bedard, P.L, Onetto, N& Hudson, T.J (2012). The genetic basis for cancer treatment decisions. *Cell*, *148*(3), 409-420.

Dash, R. (2020). A two-stage grading approach for feature selection and classification of microarray data using Pareto-based feature ranking techniques: *A case study. Journal of King Saud University-Computer and Information Sciences*, *32*(2), 232-247.

Ebrahimpour, M. K., Nezamabadi-Pour, H., & Eftekhari, M. (2018). CCFS: A cooperating coevolution technique for large-scale feature selection on microarray datasets. *Computational biology and chemistry*, *73*, 171-178.

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of theoretical biology*, 463, 77-91.

Ke, W., Wu, C., Wu, Y., & Xiong, N. N. (2018). A new filter feature selection based on criteria fusion for gene microarray data. *IEEE Access*, *6*, 61065-61076.

Kumar, M., Rath, N. K., Swain, A., &Rath, S. K. (2015). Feature selection and classification of microarray data using MapReduce based ANOVA and K-Nearest neighbor. *Procedia Computer Science*, *54*, 301-310.

Kuo, W. P, Kim, E. Y, Trimarchi, J, Jenssen, T. K, Vinterbo, S. A& Ohno-Machado, L (2004). A primer on gene expression and microarrays for machine learning researchers. *Journal of Biomedical Informatics, 37*(4), 293-303.

Mazumder, D. H., & Veilumuthu, R. (2019). An enhanced feature selection filter for classification of microarray cancer data. *ETRI Journal, 41*(3), 358-370.

Mollaee, M., & Moattar, M. H. (2016). A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. *Biocybernetics and Biomedical Engineering, 36*(3), 521-529.

Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics, 43*(1), 15-23.

Sathya, M. & Manju Priya, S. (2019). PSO search-based feature selection method for high dimensional data. *International Journal of Recent Technology & Engineering, 7*(583), 485-488.

Sathya, M. & Manju Priya, S. (2020). Modified whale optimization algorithm for feature selection algorithm in microarray cancer datasets. *International Journal of Scientific & Technology Research*, 9(3), 549-556.

Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2016, April). Using a feature selection ensemble on DNA microarray datasets. In *ESANN*.

Sun, S., Peng, Q., & Zhang, X. (2016). Global feature selection from microarray data using Lagrange multipliers. *Knowledge-Based Systems, 110*, 267-274.

Tang, J., & Zhou, S. (2016). A new approach for feature selection from microarray data based on mutual information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*(6), 1004-1015.

Wang, H., Jing, X., & Niu, B. (2017). A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. *Knowledge-Based Systems, 126*, 8-19.

Wang, H., Tan, L., & Niu, B. (2019). Feature selection for classification of microarray gene expression cancers using bacterial colony optimization with multi-dimensional population. *Swarm and Evolutionary Computation, 48*, 172-181.

Yan, C., Ma, J., Luo, H., & Patel, A. (2019). Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. *Chemometrics and Intelligent Laboratory Systems, 184*, 102-111.

Zare, M., Eftekhari, M., & Aghamollaei, G. (2019). Supervised feature selection via matrix factorization based on singular value decomposition. *Chemometrics and Intelligent Laboratory Systems, 185*, 105-113.

M.Sathya is a Research Scholar at Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India. Her research interest includes Data mining, Networking.

Dr. S. Manju Priya is Professor at the Department of Computer Science, Karpagam Academy of Higher Education. She holds a Ph.D Degree in Computer Science from Karpagam University. Her research interest includes Data mining, IoT, Sensor Networks.