Journal of Green Engineering (JGE)

Volume-10, Issue-12, December 2020

Ensemble of Multi Objective Optimizer with Pareto Frontier Solutions for Feature Selection in Large- Scale Microarray Rule Datasets

[1]M. Sathya and [2]S. Manju Priya

[1]Research Scholar, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.
E-mail-sathya22joy@gmail.com
[2]Professor, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India,
E-mail-smanjupr@gmail.com

Abstract

A larger chuck of microarray data sets are generated with the fast forward by high-throughput technologies. An analysis of microarray will allow scientists to understand a disease more effectively at the molecular level. Nowadays, the analysis of microarray data has been a vital contribution to the detection and treatment of illness. Other processes of microarray data are investigated on microarray data classification. However, the microarray data classification is still challenging due to its high-dimensionality. Lots of techniques are involved in feature selection in place to resolve the high dimensionality for microarray data classification. Search space enhanced Modified Whale Optimization Algorithm (SMWOA) was an effective feature selection method to select the most discriminative features from microarray data. SMWOA is inspired from the hunting behavior of humpback whales and it achieved a better trade-off between local exploitation and global exploration by using a self-adaptive control parameter. However, it may get sure in a part of the Pareto-optimal problem since multiple objectives are used in SMWOA. To overcome this problem and give development an better final subset of features, an Ensemble of Multi-objective Search space enhanced Modified Whale Optimization Algorithmic method (EMSMWOA) is planned in this

Journal of Green Engineering, Vol. 10_12, 12800-12819.

paper. Initially, an evidential reasoning approach is introduced to choose optimal solution from the Pareto-optimal set by setting various decision solutions based on specificity, sensitivity, Area Under Curve (AUC) and relative distance. It returns a final subset of features for microarray data classification. Furthermore, an ensemble algorithm is proposed to generate a better final feature subset. In the ensemble algorithm, multiple SMWOA is initialized with various population sizes and different maximum iteration numbers. In each SMWOA, an evidential reasoning approach is processed and selects optimal features for data classification. The selected features from multiple SMWOA are combined based on feature-class and feature-feature mutual information. Thus, the EMSMWOA picks up a stable feature subset which improves the accuracy classification. The preferred features are processed in Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB) and Artificial Neural Network (ANN) for tumor detection.

Keywords: Microarray data, Search space enhanced Modified Whale Optimization Algorithm, evidential reasoning approach, ensemble algorithm, Ensemble of Multi-objective Search space enhanced Modified Whale Optimization Algorithm.

1 Introduction

Microarray gene expression is profiling based [1] has come up as a powerful tool to identify, predict and treat cancer. DNA microarray approach has a significant effect on recent years in the discovery of the explanatory genes which contributes to cancer. Main downside of the data processing in the microarray is the bane of dimensionality problem, which holds up valuable data and contributes to computational uncertainty. Hence, the choice of appropriate features leftover an imperative in the inquiry of microarray cancer data. Various studies have been carried out for feature selection [2-5] and classification based on machine learning techniques.

A feature selection technique using Particle Swarm Optimization (PSO) [6] was introduced for dimensionality reduction of the microarray data. Then, the selected features were given as input to the Naïve Bayes and Support Vector Machine (SVM) for microarray data classification. It has a slow convergence problem. So, a Modified Whale Optimization Algorithmic model (MWOA) [3] was suggested to select the most relevant features for microarray data classification. But, the MWOA has desired to get fast into local optima and has degraded accuracy in cancer detection.

Sharbaf *e*t a*l*., [7] it uses the filtering approach method Fischer score and a wrapper method to select genes from microarray data. Using Fischer criterion, features are ranked and the highest-ranking features are filtered. It selects the high ranking gene subset is and then applied to cellular learning automata to learn the gene relationships and optimized through ant colony

optimization which selects the final subset of genes. It finally has resulted in subset evaluation using tenfold cross validation on four different datasets on SVM, KNN and NB. After the feature selection process, the classification accuracy of SVM has improved from the range 58% to 95%, KNN from 70% to 95% and NB from 94% to 97%. Based on AUC, genes with higher AUC are selected and used for classification process, it gives better results in classification rate.

2 Literature Survey

A hybrid feature selection method [8] was proposed for microarray data study. The form used a Genetic Algorithm with Dynamic Parameter (GADP) sets location for generating a number of subsets of genetics and prioritized the genes based on their frequency sequence. After that, a 2-test for homogeneity was used in pick up proper number of the peak-rated genes for data inquiry. Conclusively, the selected genes were classified based on Support Vector Machine (SVM). However, this approach has a high computational complexity problem.

A fuzzy based feature selection method [9] was proposed from self-supporting integral subspace of machine learning classification on microarray data. This method selected the self-reliant components of DNA microarray data using Fuzzy Backward Feature Elimination (FBFE) which enhance the act of SVM and Naïve Bayes (NB) classifiers which made the computational expenses reasonable. However, the datasets were processed before by fixing thresholds which greatly influenced the accuracy of classifiers.

A novel feature extraction approach [10] was proposed on ensemble feature selection and altered discriminate self-reliant component analysis for classification of microarray data. This approach will be extended by applying a sampling approach which has avoided lack of data and non-balancing problem in microarray data which was a necessary in this stage for allocating problems.

A two-stage local dimension approach [11] was proposed for local dimension reduction and classification of microarray data. This approach had a stronger ability to deal with outliers and noises. However, regularization parameter influenced the accuracy of the two-stage local dimension approach. A multi-objective simplified swarm optimization [12] method was proposed for selection of gene in microarray data. However, this method is not more suitable for complex dataset. A computation gene selection model [13] was proposed for data classification for microarray through Adaptive Hyper graph Embedded Dictionary Learning (AHEDL). However, particular number of iterations the convergence speed of the computation gene selection model is slow.

A nested genetic algorithm [14] was proposed because feature selection

in a high dimensional cancer microarray dataset. It consisted of couple nested genetic algorithms such as interior GA and exterior GA. However, a nested genetic algorithm is computationally expensive. A Recursive Memetic Algorithm (RMA) [15] was proposed for selection of gene in a microarray dataset. It is a version of Memetic Algorithm (MA) and behaves in advance than MA along with Genetic Algorithm (GA). This algorithm would aims to engage RMA on miRNA and RNA chain of data which would be useful for experts as well as to the pharmacists in upcoming years. It uses Naive Bayes, SVM, and comparison algorithm logistic regression [16]. It achieves accuracy rate as 96%. It processes with Naive Bayes, Artificial Network and Multilayer Perpector [17]. It produces better in accuracy rate. It presents a hybrid feature selection method [18]. It results in depicted threshold values with appropriate features. It works with gene selection based on effective range [19]. It gives overlapping areas and including areas and gives effective results. It deals with Naive Bayes, Decision Tree, SVM, Random Forest, and Gradient Boosting [20]. It gives clarity in results. It derives with hybrid IWSS and Shuffled Frog Leaping Algorithm [21]. It gives the highest accuracy rate. It uses binary classifiers [22].It achieves low error rates with the help of advanced decision-support systems. It uses Naïve Bayes, J48 [23]. It produces greatest accuracy for each combination. A Feature Selection Algorithmic method is based on subjective related Information for tumour Microarray Data [24]. It gains better results. It uses the feature selection algorithm [25]. It gives a solution which deals with different sources of instability. Choosing of Feature and microarray data classification applying MapReduce based ANOVA and K-Nearest neighbor [26]. Its results are better in accuracy. A hybrid gene selection method for microarray realization is dealt in this approach [27]. It produces greatest accuracy in selecting genes.

3 Proposed Methodology

In this area, the Ensemble of Multi-objective Search Space Enhanced Modified Whale Optimization Algorithm (EMSMWOA) is described in detail for cancer detection. In SMWOA three objectives are used, there may not occur one result that is best or global optimum with esteem to all target. The existence of different targeted problem usually gives rise to a family of non-dominated solutions known as Pareto-optimal result, where each intention component of all results along the Pareto-front can be improved by lowering at least one of its objective of other components. Since none of the solutions in the non-dominated set is absolutely better than any other, any of them is an acceptable solution. It is tedious to select each particular solution for multi-objective optimization problem using a SMWOA. So, an evidential reasoning is introduced in SMWOA.

3.1 Evidential Reasoning Approach for Selection of Optimal Solution from Pareto-Optimal Set

The EMSMWOA is used for cancer detection. It usually works on the existence of multiple target problems and gives a result to a family of non-dominated solutions called Pareto-optimal solutions.

Consider, there are Obj_3 objectives and K solutions $N = \{N_1, N_2, \dots N_K\}$ contains Pareto solution set. On choosing the optimal result, various outcomes right to be set. These solutions include two types. One form is based on the objective function represented as O_1 and other form is suited on the prior and preference learning is represented as O_2. Hence, two objective functions such as specificity and sensitivity are taken as the first two solutions and O_{12} is given as follows:

$$O_{12}^k = \left\{o_{sen}^k, o_{spe}^k\right\}, k = 1,2, \dots K \tag{3.1}$$

Moreover, relative length is described to evaluate the result and it chooses a solution with balanced out specificity and sensitivity. The relative distance is given as follows:

$$o_{RD}^k = \left|o_{sen}^k - o_{spe}^k\right|, k = 1,2, \dots K \tag{3.2}$$

A significant evaluation criterion is the Area Under the Curve (AUC) which is worn to find whether the EMSMWOA is reliable and the AUC is adopted as the first solution in O_{34}.Thus solution O_{34} is,

$$O_{34}^k = \left\{o_{AUC}^k, o_{RD}^k\right\}, k = 1,2, \dots K \tag{3.3}$$

The final solution set R is obtained by combining O_1 and O_2 and the final set is given as follows,

$$R = \left\{o_1^k, o_2^k, o_3^k, o_4^k\right\}, k = 1,2, \dots K \tag{3.4}$$

In Eq. (3.4), the four solutions are given as follows:

$$o_1^k = o_{sen}^k \tag{3.5}$$

$$o_2^k = o_{spe}^k \tag{3.6}$$

$$o_3^k = o_{AUC}^k \tag{3.7}$$

$$o_4^k = o_{RD}^k \tag{3.8}$$

Consider T reference points for each result whereas R which are worn to fix all solution for a and the reference points are represented as$H = \{H_1, H_2, \dots, H_T\}$. For the first three solutions in R, the greater values show the better results. For the solution, four gives better results when the value is lower. Hence, the reference value $H_{i,j}$ for all solutions of i at reference point of j is mathematically represented,

$$H_{i,j}\begin{cases}\min\left(o_i^k\right) + (j-1) \times \dfrac{\max\left(o_i^k\right) - \min\left(o_i^k\right)}{T-1}, & i = 1,2,3 \\ \max\left(o_i^k\right) - (j-1) \times \dfrac{\max\left(o_i^k\right) - \min\left(o_i^k\right)}{T-1}, & i = 4\end{cases} \tag{3.9}$$

In Eq. (3.9), $j = 1,2, \ldots N,\ k = 1,2, \ldots K$. The optimal find out selection can be modeled used to meet expectations

$$S(N_k) = \{H_{i,j}, \beta_{i,j}(N_k), j = 1,2, \ldots T\}, i = 1,2, \ldots M, k = 1,2, \ldots K \quad (3.10)$$

In Eq. (3.10), $\beta_{i,j}(N_k) \geq 0$and$\sum_{j=1}^{N} \beta_{i,j}(N_k) = 1$. $\beta_{i,j}(N_k)$ denotes a degree of belief for outcomes N_k. Same to the reference value$H_{i,j}$, $\beta_{i,j}(N_k)$ is computed beneath the two situations such as solution 1-3 and solution 4.

$\beta_{i,j}(N_k)$ 1st decision solution is calculated for solutions 1-3 and it is given as follows,

$$\beta_{i,j}(N_k) = \frac{H_{i,j+1} - o_i^k}{H_{i,j+1} - H_{i,j}}, \beta_{i,j}(N_k) = 1 - \beta_{i,j}(N_k) \text{if } H_{i,j} \leq o_i^k \leq H_{i,j+1},$$
$$\beta_{i,j}(N_k) = 0, p = 1,2, \ldots T \text{ and } p \neq j, j+1,$$
$$i = 1,2,3 \quad (3.11)$$

$\beta_{i,j}(N_k)$ is 2nd decision solution computed using for the fourth solution which is given as follows,

$$\beta_{i,j}(N_k) = \frac{H_{i,j} - o_i^k}{H_{i,j} - H_{i,j+1}}, \beta_{i,j+1}(N_k) = 1 - \beta_{i,j}(N_k), \text{ if } H_{i,j+1} \leq o_i^k \leq H_{i,j},$$
$$\beta_{i,p}(N_k) = 0\ p = 1,2, \ldots T \text{ and } p \neq j, j+1, i = 4. \quad (3.12)$$

For all feasible solutions, the belief degrees for all solutions create a belief degree matrix's which is given as follows:

$$D(N_k) = \{(H_{i,j}, \beta_{k,j}), j = 1,2, \ldots T\}, k = 1,2, \ldots K \quad (3.13)$$

$$S_k = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,T} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,T} \\ \beta_{3,1} & \beta_{3,2} & \ldots & \beta_{3,T} \\ \beta_{4,1} & \beta_{4,2} & \ldots & \beta_{4,T} \end{bmatrix}, k1,2, \ldots K \quad (3.14)$$

Through the evidential reasoning approach, all the solutions are combined in the fourth step. Consider the weights for each solution are represented as $\omega_i = 1,2, \ldots 4$, which comforts the following constraints:

$$\leq \omega_i \leq 1, \sum_{i=1}^{M} \omega_i = 1 \quad (3.15)$$

The final assessment $D(N_k)$ for solution N_k is given as follows:

$$D(N_k) = \{(H_{k,j}, \beta_{k,j}), j = 1,2, \ldots T\}, k = 1,2, \ldots K \quad (3.16)$$

After that, the belief degree $\beta_{k,j}$ for the solution N_k at all reference point j in $D(N_k)$ is computed applying the evidential reasoning approach and $\beta_{k,j}$ is given as follows,

$$\beta_{k,j} = \frac{\mu \times \left[\prod_{i=1}^{M} \left(\omega_i \beta_{k,j}(N_k) + 1 - \omega_i \sum_{j=1}^{T} \beta_{i,j}(N_k) \right) - \prod_{i=1}^{M} \left(1 - \omega_i \sum_{j=1}^{T} \beta_{i,j}(N_k) \right) \right]}{1 - \mu \times [\prod_{i=1}^{M} (1 - \omega_i)]} \quad (3.17)$$

$$\mu = \left[\sum_{j=1}^{T}\prod_{i=1}^{M}\left(\omega_i\beta_{k,j}(N_k) + 1 - \omega_i\sum_{j=1}^{T}\beta_{i,j}(N_k)\right) - (T-1)\prod_{i=1}^{M}\left(1 - \omega_i\sum_{j=1}^{T}\beta_{i,j}(N_k)\right)\right]^{-1} \quad (3.18)$$

The utility for $N_k, k = 1,2,\dots K$ is calculated to choose the optimal solution. Because there are T evaluation grades are also required. Consider that the benifit of the grades $u(H_j)$ is equidistantly shared in the utility space, i.e., $u(H_j) = \frac{j-1}{T-1}, j = 1,2,\dots T$. After that the utility for N_k is computed as,

$$U(N_k) = \sum_{j=1}^{T} u_j\beta_{k,j}, k = 1,2,\dots K \quad (3.19)$$

The final solution U^* is selected by:

$$U^* = \max(U(N_k), k = 1,2,\dots K) \quad (3.20)$$

The final solution U^* returns subset of features from the Pareto-optimal set.

3.2 Ensemble Algorithm to Produce a Better Final Feature Subset

An ensemble algorithm is proposed to yield a improved final feature subset for the classification of microarray data. In the ensemble algorithm, multiple SMWOA is initialized with various community capacity and zenith number of repetition. In each SMWOA, an evidential reasoning approach is processed to select the optimal solution from the Pareto-optimal set. After that, the features from each SMWOA are combined using an ensemble algorithm. The ensemble algorithm uses a greedy search algorithm which merge the subset of selected features by different SMWOA. For ranking particularly a common feature is selected by all the SMWOA that feature is selected without using greedy search algorithms and put together with optimal subset. Or else, the approach calculates feature-class and feature-feature mutual information and chooses a feature which has peak feature-class mutual information but margin feature-feature mutual information.

The feature-class relevance defines the amount of feature-class mutual knowledge among a feature f and a class label$C$. The importance of feature-feature is defined as the degree of feature-feature mutual information among any two features f_i and$f_i$where$f_i, f_j \in F$. The feature-feature mutual materials is computed with each features which are prior selected as optimal features and if the feature-feature mutual information of that feature with all remaining selected features is lesser than a user specified origin δ then the features will be chosen. The feature-feature mutual information is used to calculate the relevant feature which is not the selected feature with features selected.

3.2.1 Pseudo Code of Ensemble of Multi-objective Search Space Enhanced Modified Whale Optimization Algorithm

Step 1: Process N number of SMWOA with different population and maximum number of iterations to select the optimal features from microarray

dataset.
Step 2: To select the optimal solution from Pareto-optimal set apply evidential reasoning approach
Step 3: Create solutions $O^k, k = 1,2, \dots K$ using equations (3.1-3.4)
Step 4: Compute reference values $H_{i,j}, i = 1,2, \dots M, j = 1,2, \dots T$ by equation (3.9)
Step 5: Compute belief degrees $\beta_{i,j}(N_k), i = 1,2, \dots M, j = 1,2,, \dots T, k = 1,2, \dots K$ using equation (3.17)
Step 6: Using equation (3.19) calculate utilities $U(N_k), k = 1,2, \dots K$
Step 7: Choose the final optimal solution U^* from each SMWOA using equation (3.20)
Step 8: If a feature is selected by all SMWOA
Step 9: Put that feature into the optimal subset
Step 10: else compute the feature-class and feature-feature mutual information
Step 11: If feature-class mutual information is maximum and feature-feature mutual information is minimum
Step 12: Put the feature into the optimal subset
Step 13: else remove the features
Step 14: Process the selected features in SVM, KNN, NB and ANN for cancer detection.
The overall flow of EMSMWOA is shown in Figure 1.

4 Result and Discussion

In this section, the efficiency of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers are tested in order of accuracy, precision, specificity, sensitivity and also F1-score for cancer detection. The proposed EMSMWOA uses MATLAB. It grants manipulations of matrix, function plotting, implementation of data by algorithms, creating of user interfaces, and interface with programs written in some other languages. This performance evaluates EMSMWOA on microarray dataset's three publicly available benchmark datasets. Three microarray datasets such as Leukemia, Lymphoma and prostate microarray datasets are used for the experimental purpose. This method is evaluated on the testing and training datasets. The datasets are split into testing and training set in the ratio of 60:40. The description of dataset is given in the Table 1. Table 2 provide Number of Features Selected by SMWOA and EMSMWOA

Table 1 shows the number of features selected SMWOA and EMSMWOA.

Dataset	Instances	Features	Classes
Leukemia	72	3572	2
Lymphoma	77	2647	2
Prostate	102	2135	2

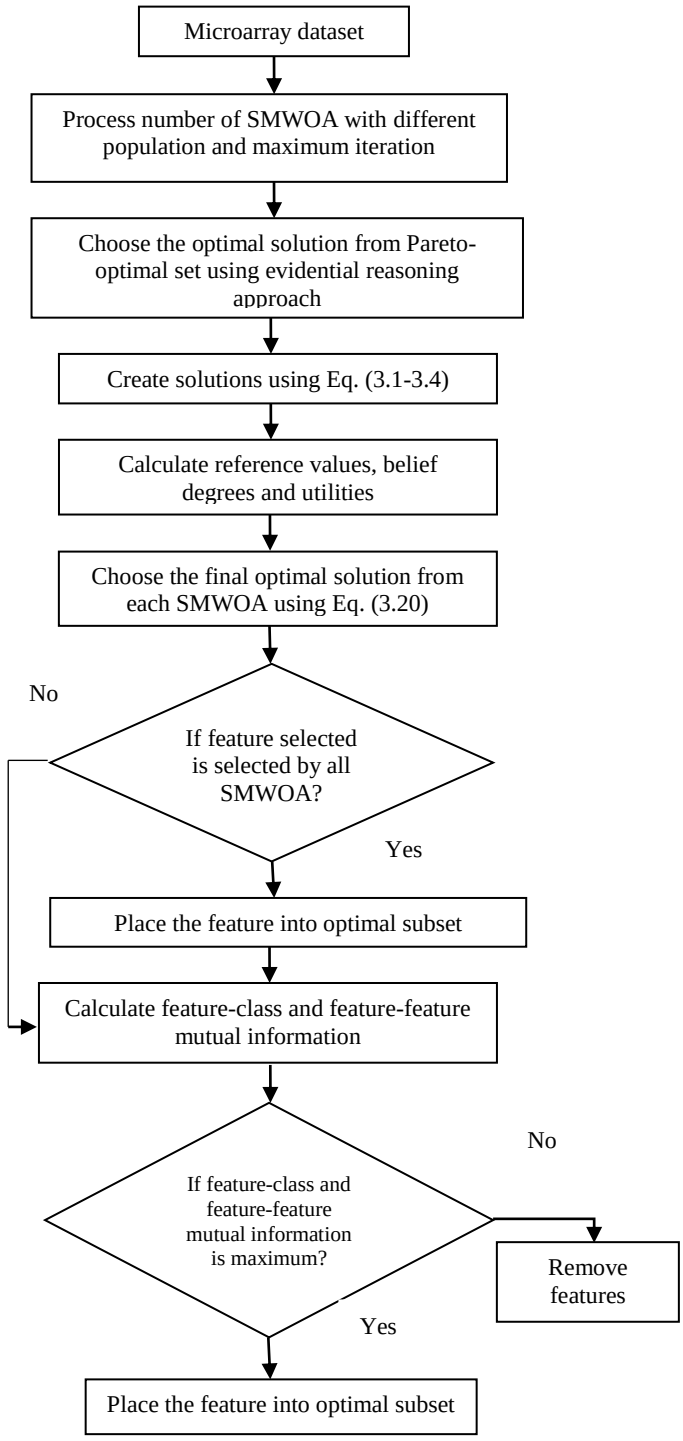

Figure 1 Overall flow of EMSMWOA

Table 2 Number of Features Selected by SMWOA and EMSMWOA

	No of features	SMWOA	EMSMWOA
Leukemia	3572	38	35
Prostate	2135	110	100
Lymphoma	2647	33	28

4.1 Accuracy

It deals as the proportion of instances that are classified correctly. The calculation is done by total number of correctly predicted sick people (true positive) and correctly predicted healthy people (true negative) over the total number of classifications. Mathematically, accuracy is defined as,

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

Where TP-sick people who are affected by cancer are classified correctly as sick. FP defines healthy people classified incorrectly as sick. TN-Healthy people are classified correctly as healthy. FN-Sick people are classified incorrectly as healthy.

Table 3 shows the accuracy of SMWOA and EMSMWOA with different classifiers on three different

Table 3 Comparison of Accuracy

	SMWOA-SVM	SMWOA-KNN	SMWOA- NB	SMWOA-ANN	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN
Leukemia	58.24	82.31	78.65	83.47	62.42	86.13	82.34	88.42
Prostate	87.26	80.42	75.87	89.02	92.13	85.36	79.24	91.87
Lymphma	92.56	81.83	83.43	95.62	94.15	85.67	86.98	96.14

The accuracy of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers is shown in the Figure 2. The classifiers are taken in X-axis and they are Y-axis shows the accuracy of feature selection methods. The accuracy of EMSMWOA-ANN is 51.82% greater than SMWOA-SVM, 7.42% greater than SMWOA-KNN, 12.42% greater than SMWOA-NB,

5.93% greater than SMWOA-ANN, 41.65% greater than EMSMWOA-SVM, 2.66% greater than EMSMWOA-KNN and 7.38% greater than EMSMWOA-NB.

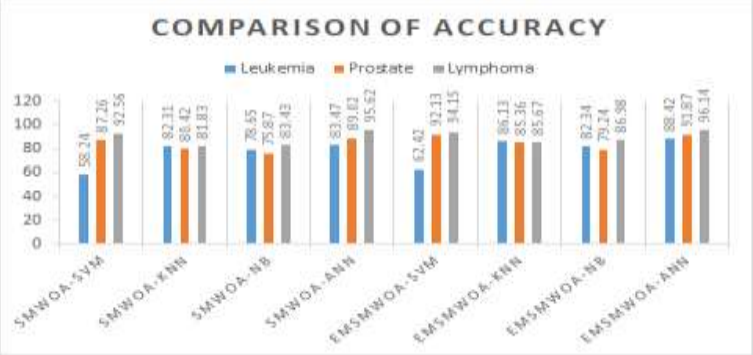

Figure 2 Accuracy Comparison on Leukemia, Prostate, Lymphoma dataset

From this analysis, it is proved that the proposed EMSMWOA-ANN has a higher accuracy than other methods on leukemia, prostate and lymphoma datasets for cancer detection.

4.2 Precision

Precision deals with the proportion of true positive instances are classified only as positive. Mathematically, precision is defined as,

$$Precision = \frac{TP}{TP + FP}$$

The Table 4 shows the precision of SMWOA and EMSMWOA with different classifiers on three different datasets.

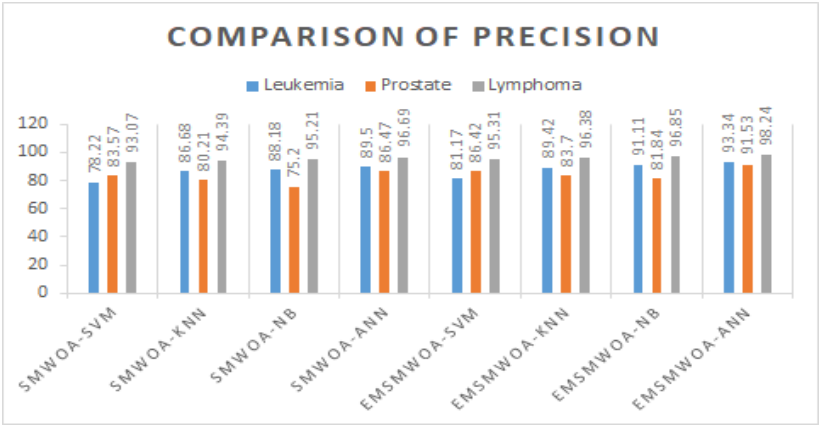

Figure 3 Precision comparison on Leukemia, Prostate, Lymphoma dataset

Table 4 Comparison of Precision

	SMWOA-SVM	SMWOA-KNN	SMWOA-NB	SMWOA-ANN	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN
Leukemia	78.22	86.68	88.18	89.50	81.17	89.42	91.11	93.34
Prostate	83.57	80.21	75.20	86.47	86.42	83.70	81.84	91.53
Lymphoma	93.07	94.39	95.21	96.69	95.31	96.38	96.85	98.24

The precision of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers is shown in the Figure 3. The classifiers are taken in X-axis and the precision of feature selection methods is shown in Y-axis. The precision of EMSMWOA-ANN is 19.33% greater than SMWOA-SVM, 7.68% is greater than SMWOA-KNN, 5.85% is greater than SMWOA-NB, 4.29% is greater than SMWOA-ANN, 14.99% greater than EMSMWOA-SVM, 4.38% is greater than EMSMWOA-KNN and 2.45% is greater than EMSMWOA-NB on leukemia dataset.

From this analysis, it is proved that the proposed EMSMWOA-ANN has high precision than the other methods on leukemia, prostate and lymphoma datasets for cancer detection.

4.3 Specificity or Recall

Specificity has the ability to measure the proportion of actual negatives that are correctly identified as negative (e.g., the percentage of healthy people who are correctly identified as having the condition). Mathematically, specificity is defined as,

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

The Table 5 gives the specificity of SMWOA and EMSMWOA with different classifiers on three different datasets.

The specificity of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers is shown in the Figure 4. The classifiers are taken in X-axis and the specificity of feature selection methods is shown in Y-axis. The specificity of EMSMWOA-ANN is 19.57% greater than SMWOA-SVM,

7.76% is greater than SMWOA-KNN, 5.92% greater than SMWOA-NB, 4.35% is greater than SMWOA-ANN, 15.16% is greater than EMSMWOA-SVM, 4.43% is greater than EMSMWOA-KNN and 2.47% is greater than EMSMWOA-NB.

Table 5 Comparison of Specificity

	SMWOA-SVM	SMWOA-KNN	SMWOA-NB	SMWOA-ANN	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN
Leukemia	77.25	85.72	87.21	88.52	80.21	88.45	90.14	92.37
Prostate	82.60	79.24	74.24	85.49	85.46	82.73	80.89	90.56
Lymphoma	92.12	93.42	94.23	95.73	94.34	95.41	95.89	97.28

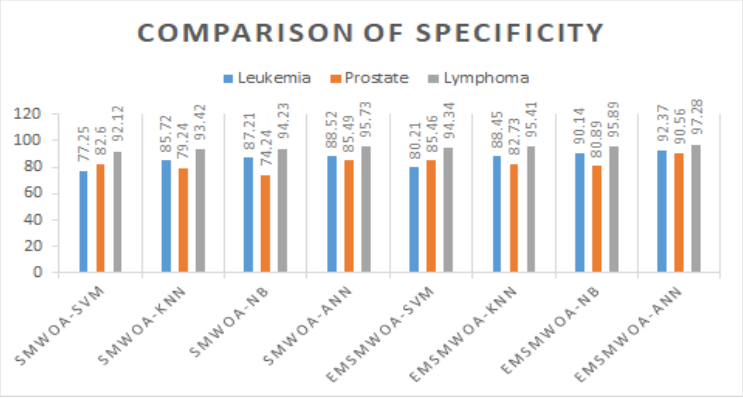

Figure 4 Specificity comparison on Leukemia, Prostate, Lymphoma dataset

From this analysis, it is proved that the proposed EMSMWOA-ANN has higher specificity than other methods on leukemia, prostate and lymphoma datasets for cancer detection.

4.3 Sensitivity

Sensitivity calculates the proportion of actual positives that are correctly identified as positive such (e.g., the percentage of sick people who are correctly identified as having the condition). Mathematically, to define sensitivity we use following,

$$Sensitivity = \frac{TP}{TP + FN}$$

The Table 6 shows the sensitivity of SMWOA and EMSMWOA with different classifiers on three different datasets.

Table 6 Comparison of Sensitivity

	SMWOA-SVM	SMWOA-KNN	SMWOA-NB	SMWOA-ANN	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN
Leukemia	64.25	79.39	82.61	84.24	66.78	82.45	85.67	89.39
Prostate	94.05	84.65	85.36	95.42	96.02	89.27	90.52	98.47
Lymphoma	73.46	76.35	80.35	84.65	78.16	80.61	85.19	90.04

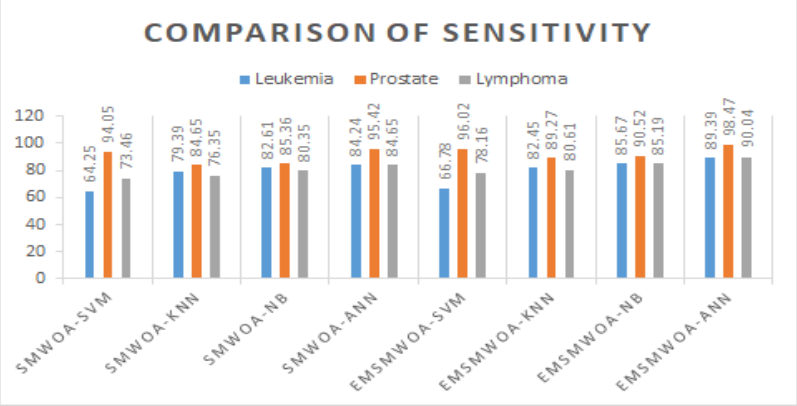

Figure 5 Sensitivity comparison on Leukemia, Prostate, Lymphoma dataset

The sensitivity of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers is shown in the Figure 5. The classifiers are taken in X-axis and the sensitivity of feature selection methods is shown in Y-axis. The sensitivity of EMSMWOA-ANN is 39.13% is greater than SMWOA-SVM, 12.6% is is greater than SMWOA-KNN, 8.21% is greater than SMWOA-NB, 6.11% is greater than SMWOA-ANN, 33.86% is greater than EMSMWOA-SVM, 8.42% is greater than EMSMWOA-KNN and 4.34% is greater than EMSMWOA-NB. From this analysis, it is proved that the proposed EMSMWOA-ANN has a higher sensitivity than other methods on leukemia, prostate and lymphoma datasets for cancer detection.

4.4 F1-score

F1-score produces the harmonic mean of precision, recall. Mathematically, F1-score is dealt by using formula below,

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

The Table 7 gives the F1-score comparison of SMWOA and EMSMWOA with different classifiers on three different datasets.

Table 7 Comparison of F1-score

	SMWOA-SVM	SMWOA-KNN	SMWOA- NB	SMWOA-ANN	EMSMWOA-SVM	EMSMWOA-KNN	EMSMWOA-NB	EMSMWOA-ANN
Leukemia	77.52	86.82	84.32	88.45	80.24	89.92	88.24	91.52
Prostate	90.25	85.65	79.24	91.54	93.18	89.24	84.68	95.02
Lymphoma	78.26	69.47	80.24	84.43	82.65	72.36	84.29	89.78

The F1-score of SMWOA and EMSMWOA with SVM, KNN, NB and ANN classifiers is shown in the Figure 6. The classifiers are taken in X-axis and the F1-score of feature selection methods is shown in Y-axis. The F1-score of

EMSMWOA-ANN is 18.06% is greater than SMWOA-SVM, 5.41% greater than SMWOA-KNN, 8.54% is greater than SMWOA-NB, 3.47% is greater than SMWOA-ANN, 14.06% is greater than EMSMWOA-SVM, 1.78% greater than EMSMWOA-KNN and 3.72% is greater than EMSMWOA-NB.

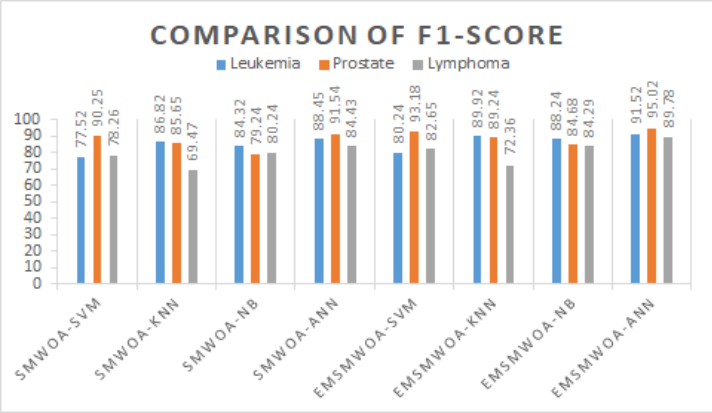

Figure 6 F1-score comparison on Leukemia, Prostate, Lymphoma dataset

From this analysis, it is proved that the proposed EMSMWOA-ANN has a higher F1-score than other methods on leukemia, prostate and lymphoma datasets for cancer detection.

4.5 Average Error

It is the average error of classifiers to classify the gene expression data with the selected features by SMWOA and EMSMWOA.

The Table 8 shows the average error of classifiers which processed the selected features by SMWOA and EMSMWOA.

Table 8 Comparison of Average Error

No. of iteration	SMWOA	EMSMWOA
100	0.079	0.075
200	0.06	0.059
300	0.055	0.04
400	0.05	0.03
500	0.05	0.025
600	0.0.5	0.025
700	0.05	0.025
800	0.05	0.025
900	0.05	0.025
1000	0.05	0.025

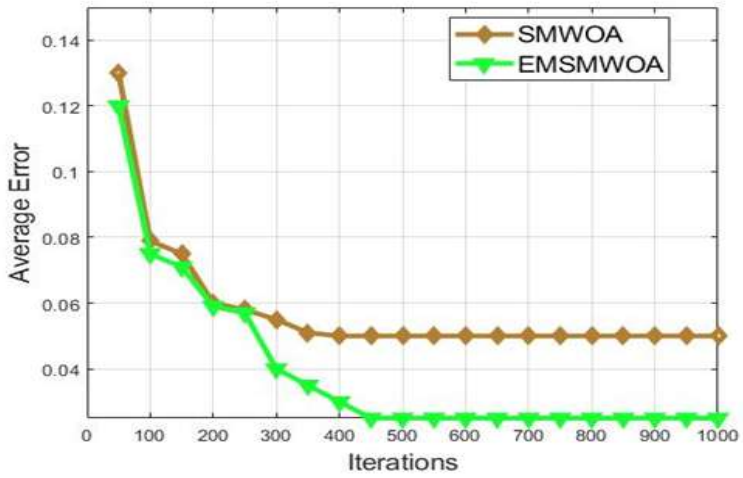

Figure 7 Comparison of Average Error

The average error of SMWOA and EMSMWOA with classifiers is shown in the 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 EMSMWOA is 1.67% less than SMWOA. From this analysis, it is proved that the EMSMWOA method has less average error than SMWOA for cancer detection.

5 Conclusion

In this paper, Ensemble of Multi-objective Search space enhanced Modified Whale Optimization Algorithm (EMSMWOA) is planned method to choose the most discriminative features in the dataset of microarray. Also, it solves the multi-objective optimization problem by selection optimal solution from the Pareto optimal set using evidential reasoning approach. In this approach, specificity, sensitivity, AUC and relative distance are used to evaluate the solutions and based on those, measures optimal solution for feature is selected. In addition to this, an ensemble algorithm is introduced to ensemble the results of different SMWOA based on feature-class and feature-feature mutual information to produce best final features subset for classification of microarray dataset. Finally, the selected features are given as input to SVM, KNN, NB and ANN for cancer detection. The experimental results prove that the proposed EMSMWOA with ANN classifier has better accuracy, precision, specificity, sensitivity, F1-score and average error than methods on three different datasets for cancer detection.

Reference

[1] Hsieh, S. Y., Chou, Y. C., "A faster cDNA microarray gene expression data classifier for diagnosing diseases", IEEE/ACM transactions on computational biology and bioinformatics, Vol. 13, no.1, pp.43-54, 2016.

[2] Yang, C. H., Chuang, L. Y., Yang, C. H., "IG-GA: a hybrid filter/wrapper method for feature selection of microarray data", Journal of Medical and Biological Engineering, Vol. 30, no.1, pp. 23-28, 2010.

[3] Jain, I., Jain, V. K., Jain, R., "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification", Applied Soft Computing, Vol. 62, pp.203-215, 2018.

[4] Cao, B., Zhao, J., Yang, P., Yang, P., Liu, X., Qi, J., & Muhammad, K., " Multiobjective feature selection for microarray data via distributed parallel algorithms", Future Generation Computer Systems, Vol. 100, pp. 952-981, 2019.

[5] Sathya, M., Manju Priya .S., "PSO search-based feature selection method for High Dimensional data, International Journal of Recent Technology and Engineering (IJRTE), Vol. 7, no. 5S3, pp. 485-488, 2019.

[6] Sathya, M., Manju Priya, S., "Modified Whale Optimization Algorithm for Feature Selection in Micro Array Cancer Dataset", International Journal of Scientific & Technology Research, Vol. 9, no. 9, pp. 549-556, 2020.

[7] Sathya, M., Manju Priya, S., "A search space enhanced modified whale optimization algorithm for feature selection in large-scale microarray datasets", Indian Journal of Science and Technology, Vol.13, no. 42, pp. 4396-4406, 2020.

[8] Lee, C. P., Leu, Y., "A novel hybrid feature selection method for microarray data analysis", Applied Soft Computing, Vol. 11, no. 1, pp. 208-213, 2011.

[9] Mollaee, M., Moattar, M. H., " A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification", Biocybernetics and Biomedical Engineering, Vol. 36, no.3, pp.521-529, 2016.

[10] Aziz, R., Verma, C., & Srivastava, N. "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data", Genomics Data, Vol. 8, pp.4-15, 2016.

[11] Guo, S., Guo, D., Chen, L., Jiang, Q., "A L1-regularized feature selection method for local dimension reduction on microarray data", Computational Biology and Chemistry, Vol. 67, pp. 92-101, 2017.

[12] Lai, C. M., "Multi-objective simplified swarm optimization with weighting scheme for gene selection", Applied Soft Computing, Vol. 65, pp. 58-68, 2018.

[13] Zheng, X., Zhu, W., Tang, C., Wang, M., "Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning", Gene, Vol. 706, pp. 188-200, 2019.

[14] Sayed, S., Nassef, M., Badr, A., Farag, I., "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets", Expert Systems with Applications, Vol. 121, pp. 233-243, 2019.

[15] Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., Maulik, U., "Recursive memetic algorithm for gene selection in microarray data", Expert Systems with Applications, Vol.116, pp. 172-185, 2019.

[16] Rajit Nair., Amit Bhagat., Feature selection method to improve the accuracy of classification algorithm", International Journal of Recent Technology and Engineering (IJRTE),Vol. 8, no. 6, pp. 124-127, 2019.

[17] Esar, Mahsereci, Karabulut., Selma, Ayse, Ozel., et al., " A Comparative study on effect of feature selection on classification accuracy", Procedia Technology , Vol.1, pp. 323-327, 2012.

[18] Divya, Jain., Vijendre, Singh., "An efficient hybrid feature selection model for dimensionality reduction", Procedia Computer Science, Vol. 132, pp. 333-341,2018.

[19] Jianzhong, Wang., Shunag, Zhou., et al., "An improved feature selection bases on effective range for classification", The scientific world journal, 2014.AvailableOnline: https://www.hindawi.com/journals/tswj/2014/972125/

[20] N. Bhalaji., K. B. Sundhara, Kumar., Chithra, Slevaraj., "Empirical study of feature selection methods over classification algorithms", International Journal of Intelligence Systems Technologies and Applications, Vol.17, no.1/2 , pp. 1-11, 2018.

[21] JamshidPigazi., Mohsen Alimoradi et al.., "An efficient hybrid filter-wrapper metaheuristicbased gene selection method for high dimensional datasets", Scientific Reports, Vol. 9, no. 1, 2019.

[22] Ghaddar,Bissan., Naoum-Sawaya. Joe., "High dimensional data classification and feature selection using support vector machine", European Journal of Operational Research , Vol. 265, no. 3, pp. 993-1004, 2018.

[23] Fifie, Francis., "Feature selection and classification accuracy of data mining algorithms", International research journal of engineering and technology, Vol. 5, no. 11, pp. 1280-1283, 2018.

[24] Nagpal, A., Singh, V. A., "Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data", Procedia Computer Science, Vol. 132, pp. 244-252, 2018.

[25] Utkarsh, Mahadeo, Khaire., R.Dhanalakshmi., "Stability of feature selection:A review", Journal of King Saud University - Computer and Information Science, 2019.

[26] Kumar, M., Rath, N. K., Swain, A., Rath, S. K., "Feature selection and classification of microarray data using MapReduce based ANOVA and K-Nearest neighbor", Procedia Computer Science, Vol. 54, pp. 301-310, 2015.

[27] Shukla, A. K., Singh, P., Vardhan, M., "A hybrid gene selection method for microarray recognition", Biocybernetics and Biomedical Engineering, Vol. 38, no. 4, pp. 975-991, 2018.

Biographies:

M. Sathya, Research Scholar, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.

S. Manju Priya, Professor, Dept. of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India,