INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A search space enhanced modified whale optimization algorithm for feature selection in large-scale microarray datasets

M Sathya[1]*, **S Manju Priya**[2]

1 Research Scholar, Dept. of CS, Karpagam Academy of Higher Education, Coimbatore, 641 021, India
2 Professor, Dept. of CS, Karpagam Academy of Higher Education, Coimbatore, 641 021, India

OPEN ACCESS

Received: 19.09.2020
Accepted: 28.11.2020
Published: 04.12.2020

Citation: Sathya M, Manju Priya S (2020) A search space enhanced modified whale optimization algorithm for feature selection in large-scale microarray datasets. Indian Journal of Science and Technology 13(42): 4396-4406. https://doi.org/10.17485/IJST/v13i42.767

* **Corresponding author**.

sathya22joy@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To enhance the microarray data classification accuracy, to accelerate the convergence speed of classifier, and Modified Whale Optimization Algorithm (MWOA), refine the best balance among local exploitation and global exploration, a Search space enhanced Modified Whale Optimization Algorithm (SMWOA) is the proposed task. **Methods:** The SMWOA selects the optimal features stands on the Levy flight method and quadratic interpolation method. Levy flight which employs for acceleration convergence speed of SMWOA and also holds the result from local optima builds up by the population assortment. A quadratic interpolation takes up the exploitation stage for deeper searching within the search area. **Finding:** In addition to this, a self-adaptive control parameter is introduced to make a clear variation to the solution quality. It refines the best equity among the local exploitation method by global exploration method. After selection of features, those are processed in Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN) classifiers for cancer detection. **Novelty:** The classification accuracy is improved by processing the most discriminative features in the classifiers. The overall accuracy, specificity, sensitivity, F1-score and average error of SMWOA-ANN are 6.7%, 5.6%, 7.3% and 5.6% greater than MWOA-ANN respectively for cancer detection.

Keywords: Gene expression data; dimensionality reduction; feature selection; modified whale optimization algorithm (MWOA); search space enhanced modified whale optimization algorithm (WOA)

1 Introduction

A principle of microarray technology is established and verified for antibody microarrays in a registered series of patents[(1)]. In a single experiment, Microarray technology allows biologists to determine expression rates of thousands of genes. The microarray data consists of a small sample and huge dimensional data. A disadvantage of microarray data on gene expression is that the number of genes greatly exceeds the

sample size generally referred to as the curse of dimensionality. As a result of efforts to improve the drug discovery process, microarrays have been developed. In Order to avoid the complication of cursing dimensionality, dimension reduction shows a pivotal role in DNA microarray investigation.

Microarray studies supply the scientist with tremendous data, but important information and knowledge contained in this database cannot be found without proper instruments and methodologies. Analytical and computational problems emerge due to a large amount of raw gene expression data. The challenging work of the analyst is the texture or nature of the microarray data. The best analytical models greatly depend on the sum of all possible gene combo. Hence, the hit of microarray technologies depends biologically on massive data mining and analytical methods. The field of data mining performs a chief role which plans to solve the problem to bane of dimensionality [2–4].

The selection of feature [5–8] is a data mining technique that is used to figure out the curse of dimensionality. The selection of feature is a method used to find the difference between relevant and irrelevant features and eliminates irrelevant features. A feature selection strategy using Particle Swarm Optimization (PSO) [9] has been introduced to cut the dimensions of the microarray data. It reduces the features in microarray data which are processed in Naïve Bayes classifier and Support Vector Machine classifier (SVM) for data classification. Sometimes, PSO has a slow convergence problem. The advantages of WOA [10–15] such as the low number of parameters and low possibility of getting struck into the local optimal problem are intended to use the feature selection in cancer detection. In order to effectively select feature subsets, a Modified Whale Optimization Algorithmic method (MWOA) [16] acts as a planned method for selecting the most relevant features in the microarray cancer dataset. With the simplicity and less dependency on the parameters of WOA, the local minima can be extended to fit the best solution at random through which the exploration can be tuned to fit the best position of the agents. The MWOA focus to improve the exploration of the Whale Optimization Algorithm which employs on fitness function and aims to find the location of the agents at minimum length. However, sometimes the MWOA will get struck into local optima problem (i.e., the solution is optimal within a neighboring set of candidate solution) which degrades cancer detection accuracy.

In this paper, Search space enhanced MWOA (SMWOA) is proposed to enhance the cancer detection accuracy and to improve the exploration and exploitation strategy of MWOA. A non-linear dynamic strategy is introduced in SMWOA to update the control parameters of MWOA and it also balances the exploration and exploitation abilities of MWOA. In order to escape from local optima problem for cancer detection, a Levy-flight strategy is introduced in SMWOA. Moreover, the leader of the population is done by a quadratic interpolation method which boosts the local exploitation ability which enhances cancer detection accuracy.

2 Literature Survey

A global feature selection method [17] was done by semi-definite programming models using Lagrange multipliers. However, a threshold value used in Lagrange multipliers greatly influenced the classification accuracy. A framework [18] was proposed to choose top-ranked features for microarray data. However, if the datasets were imbalanced, efforts can be given to resolve the imbalance issue in the dataset. A bi-objective genetic algorithm [19] was proposed for ensemble-based feature selection. The classification accuracy will be improved by using multiple objectives in genetic algorithm.

A two-step attribute selection method [20] was introduced for cancer diagnosis using kernel-based learning. Other objective functions such as the correlation between genes or class separability distances will be considered for cancer diagnosis. A recursive Particle Swarm Optimization (PSO) [21] was proposed for gene selection in the microarray dataset. One of the key future direction of recursive PSO based gene selection has included exploring of other soft computing approaches and extending it for further minimizing the number of genes. Improved-binary particle swarm optimization method results in correlation based feature selection [22] was proposed for gene selection and cancer classification However, this method further improved in terms of accuracy.

A multi-objective simplified swarm optimization [23] method was proposed for picks up gene in microarray data. However, this method was not more suitable for the complex datasets. A Partial Maximum Correlation Information (PMCI) method was proposed [24] for microarray data classification. However, it has a poor F-score. A Binary Coral Reef Optimization (BCRO) algorithm [25] was proposed to select the most significant features from the microarray datasets. The exploration and exploitation of BCRO would be improved by combining with other local search strategies or swarm intelligence algorithms. A multi-objective feature selection model [26] was constructed for microarray data through a distributed parallel algorithm. However, there may be chances for rosining conflicts between the multiple objectives.

A hybrid metaheuristic using binary black hole algorithmic method and Particle Swarm Optimization method(PSO) [27] was proposed for gene selection. However, it was not suitable for complex datasets. A feature selection method [28] was proposed for microarray data classification according to the Hidden Markov Model (HMM). This method will be extended by using more

feature selection methods with HMM-based method and KNN, ANNs [29] gave better results and classification rate was higher. A metaheuristic method [30] was proposed for gene selection based on binary shuffled for the leap algorithm. This method will be extended for the high dimensional multi-class classification problem.

A feature selection strategy [31] was presented to improve the classification performance over high dimensional datasets. However, analyzing each cluster with Multi-Layer Perceptron (MLP) in sequential order was a highly time-consuming process which was the major drawback of this strategy. A method of filter wrapper hybrid feature selection approaches on the Genetic Algorithm (GA) using the penalty scheme and weighted occurrence frequency [32] was proposed for dimensionality reduction in biomedical datasets. However, this approach required a large number of records for reliable sampling. A wrapper-based feature selection [33] method was proposed to increase the ability of the intrusion detection system. This method will be extended by considering more groups so that the algorithm can categorize diverse types of intrusions within the datasets of the intrusion detection rule. An enhanced Artificial Bee Colony algorithm according to the Whale Optimization Algorithm (ABCWOA) [34] was proposed for data clustering. The performance of ABCWOA will be enhanced by using new operators concerning the optimization problem.

3 Proposed Methodology

The paper focuses to bring down the dimensionality of microarray data and raise convergence speed of classifiers, improving the detection accuracy. There are three modifications such as tracking the large scale global optimization problem using Quadratic Interpolation (QI), securing the solution from local optima using Levy Flight (LF), and maintaining perfect harmony between exploration and exploitation by using non-linear control parameter strategy is carried out in SMWOA of selectively picks up with the most discriminative features in the microarray dataset.

- QI is carried out in the exploitation time of SMWOA to keep and maintain the population diversity.
- LF is introduced in SMWOA to get away local optima by boosting up the diversity of population.
- The non-linear control parameter strategy has introduced a parameter that makes perfect harmony between exploration and exploitation.

By making these modifications, the SMWOA selects the most discriminative features from the dataset and the selected features are used in NB, SVM, KNN, and ANN for data classification.

3.1 Tracking the large scale global optimization problem using quadratic interpolation

Even though the MWOA with the skirting mechanism and spiral way is exploring well in the search space, it still requires enhancement to track the large scale global optimization problem. The Quadratic Interpolation (QI) method inhibits in SMWOA to enchase with exploitation capability and also enhances the data classification accuracy. Mathematically, the minimum point of quadratic curve QI is driven bypassing three selected solutions in n-dimensional space. A crossover operator, Quadratic Interpolation (QI) chooses the excellent search agents $X^* = (x_1^*, x_2^*, \ldots x_n^*)$ and other two parents $A = (a_1, a_2, \ldots a_n)$, $B = (b_1, b_2, \ldots b_n)$ as three parents next it creates a newest solution $X = (x_1, x_2, \ldots x_n)$ to select the optimal features from microarray dataset. The newest solution is expressed as follow

$$x_i = 0.5 \times \frac{\left(a_i^2 - b_i^2\right) \times f(x^*) + \left(b_i^2 - a_i^2\right) \times f(A) + \left(x_i^{*2} - a_i^2\right) \times f(B)}{\left(a_i - b_i\right) \times \left(b_i^2 - x_i^2\right) + \left(b_i^2 - a_i^2\right) \times f(A) + \left(x_i^{*2} - a_i^2\right) \times f(B)}, \ \forall i = 1, 2, \ldots n \tag{3.1}$$

In Equation (3.1), $f(X^*)$, $f(A)$ and $f(B) \rightarrow$ fitness values at X^*, A also B correspondingly, i-th dimension is stands by i. Fitness value is calculated by formula

$$f(X) = IM(X) - cardinality(X) + \ accuracy(X) \tag{3.2}$$

In Equation (3.2), $IM(X)$ is the information measure of each feature in the new solution X, $cardinality(X)$ is the proportion of the total features present in the dataset to the features present in the subset and $accuracy(X)$ is the classifier accuracy on a classifier.

In the quadratic crossover, the present good search agent C^* takes up the leading role, which allows search agents to better discover the global optimum solution. The population diversity is preserved and needed to enhance the ability of exploitation using SMWOA, Exploitation phase uses Quadratic Interpolation method. In exploitation phase of SMWOA thus in two

elements such as the spiral-shaped path, and quadratic crossover. The parameter is evenly shared to control the two elements. The spiral-shaped path method is done if the probability is less than 0.6, is calculated by

$$X(t+1) = D^{'} \times e^{bl} \times cos(2\pi l) + X^{*}(t) \tag{3.3}$$

In Equation (3.3),

$$D^{'} = |X^{*}(t) - X(t)| \tag{3.4}$$

The Equation (3.4), $D^{'}$ denotes space between the ith whale (actor) has a good solution (optimal features) got so far, all iteration will restore if there is the best result, t -> present iteration, $|.|$ denotes absolute value operation, $\times$ produces element-by-element multiplication, b gives the constant for describing the structure of a logarithmic spiral along with l denotes the random number which ranges from -1 to 1. If the probability is greater than 0.6, to execute the quadratic crossover which updates the position of whales.

3.2 Securing the solution from local optima using Levy Flight (LF)

Levy Flight (LF) process is used to secure the solution of local optima which has driven the meeting velocity of powerful global search. Hence, LF works in SMWOA to get away from local optima by upgrading the diversity of population. LF results in non-Gaussian random process in addition to step length following a Levy distribution. The Levy distribution of simple power-law vision is given as:

$$L(s) \sim |s|^{-l-\beta}, 0 < \beta \leq 2 \tag{3.5}$$

In Equation (3.5) , β denotes an index, s denotes the step length of the LF. The step length s is calculated by using Mantegna's algorithm

$$s = \frac{\mu}{|v|^{1/\beta}} \tag{3.6}$$

In Equation (3.6), μ and v obey normal distribution which are mathematically expressed as,

$$\mu \sim N\left(0, \sigma_{\mu}^{2}\right) \tag{3.7}$$

$$v \sim N\left(0, \sigma_{v}^{2}\right) \tag{3.8}$$

$$\sigma_{\mu} = \left[\frac{\Gamma(1+\beta) \times \sin(\pi \times \beta/2)}{\Gamma((1+\beta)/2) \times \beta \times 2^{(\beta-1)/2}}\right]^{1/\beta} \tag{3.9}$$

$$\sigma_{v} = 1 \tag{3.10}$$

The LF jumping out is of the design domain is accepted to avoid step size. The calculation is given below

$$Levy = random(size(D)) \oplus L(\beta) \sim \frac{0.01}{|v|^{\frac{1}{\beta}}(X_i - X^{*})} \tag{3.11}$$

In Equation (3.11), $size(D)$ -> range of the problem considered, $\oplus$ shows entry-wise multiplications and X_i gives the i^{th} solution vector.

LF also carries out long-distance movement to facilitate the opportunity to explore the limitless variety of the Lévy distribution, and short-distance movement, is used to maximize exploration potential. This advantage can ensure that the meta-heuristic algorithmic method jumps out of local optima. Exploring the search space more effectively, need a shrinking encircling mechanism to be replaced by LF. Based on the following rule, the new position is updated as

$$X(t+1) = X(t) + 1/sqrt(t) \times sign(random - 0.5) \oplus Levy \tag{3.12}$$

In Equation (3.12), $1/sqrt(t)$gives the framework related to the present iteration number t also $sqrt(t)$ represents the operation of square root. To the point, search movements larger range that can be implemented in the early stage during little ones are involved for future period. $sign\,(random-0.5)$ Sign function has 3 values -1, 0 and 1, which results a random search. During phase of exploration SMWOA is as follows,

$$X(t+1) = \begin{cases} X(t) + \dfrac{1}{sqrt(t)} \times sign\,(random-0.5) \oplus Levy, & if\ p < 0.5 \\ D' \times e^{bl} \times \cos(2\pi l) + X^*(t), & if\ p \geq 0.5 \end{cases} \tag{3.13}$$

3.3 Maintaining perfect harmony between exploration and exploitation by using non-linear control parameter strategy

In order to achieve a good performance, a perfect harmony has to be maintained between exploration and exploitation. In MWOA, A coefficient is an important factor in balancing exploration and exploitation. Already mentioned above, whales make charge towards the prey (exploitation) by value $|A|<1$ and whales, to examine a search area by value$|A|>1$. A is the coefficient, directly affected by the linearly decreasing parameter$'$ a $'$. However, the linearly decreasing parameter a cannot represent or respond correctly to the difficult and also to the non-linear search process. Bringing this into view, SMWOA uses a control parameter nonlinear to specifically influence the consistency of the solution. SMWOA has employed a function of cosine to update $'a'$ each iterations which is given as follows:

$$a = 2 \times cos\left(\frac{t}{max_{itr}}\right) \tag{3.14}$$

In Equation (3.14) , max_{itr} denotes the maximum iteration.

Search space enhanced modified whale optimization algorithmic method for feature selection (SMWOA)

1. Initialize the whale population $X_i(i=1,2,3,\ldots n)$, max_{itr}
2. Each whale randomly chooses the features of microarray dataset
3. Each search agent fitness is measured (classification accuracy)
4. The best search agent *assigned to* X^*
5. while $(t < max_{itr})$
6. for every search agent
7. Equation (3.14) is used to update a
8. Update A, c, l, p_1 and p_2
9. if $(p_1 < 0.5)$
10. if $(|A| < 1)$
11. the current search agent position updated using Equation (3.12)
12. else if $((A| \geq 1)$
13. hoose search agent randomly by (X_{rand})
14. The current search agent position is updated by$X(t+1) = X_{rand} - A \times D$ by using (3.15)
15. end if
16. else if $(p_1 \geq 0.5)$
17. if $(p_2 < 0.6)$
18. the current search agent position is updated by using Equation (3.3) ,
19. else if $(p_2 \geq 0.6)$
20. The position of the current search agent is updated by using Equation (3.1),
21. end if
22. end if
23. end for
24. verifies & repairs to ensure duplicate gene that every search agent is valid or not
25. Every search agent fitness of is computed by using Equation (3.2)
26. If there is a better solution update X^*

27. $t++$
28. end while
29. return X^*

SMWOA is used to selectively give the most compelling features in microarray dataset. The selected features are processed in NB, SVM, KNN and ANN for cancer detection. The overall flow of SMWOA is shown in [Figure 1].

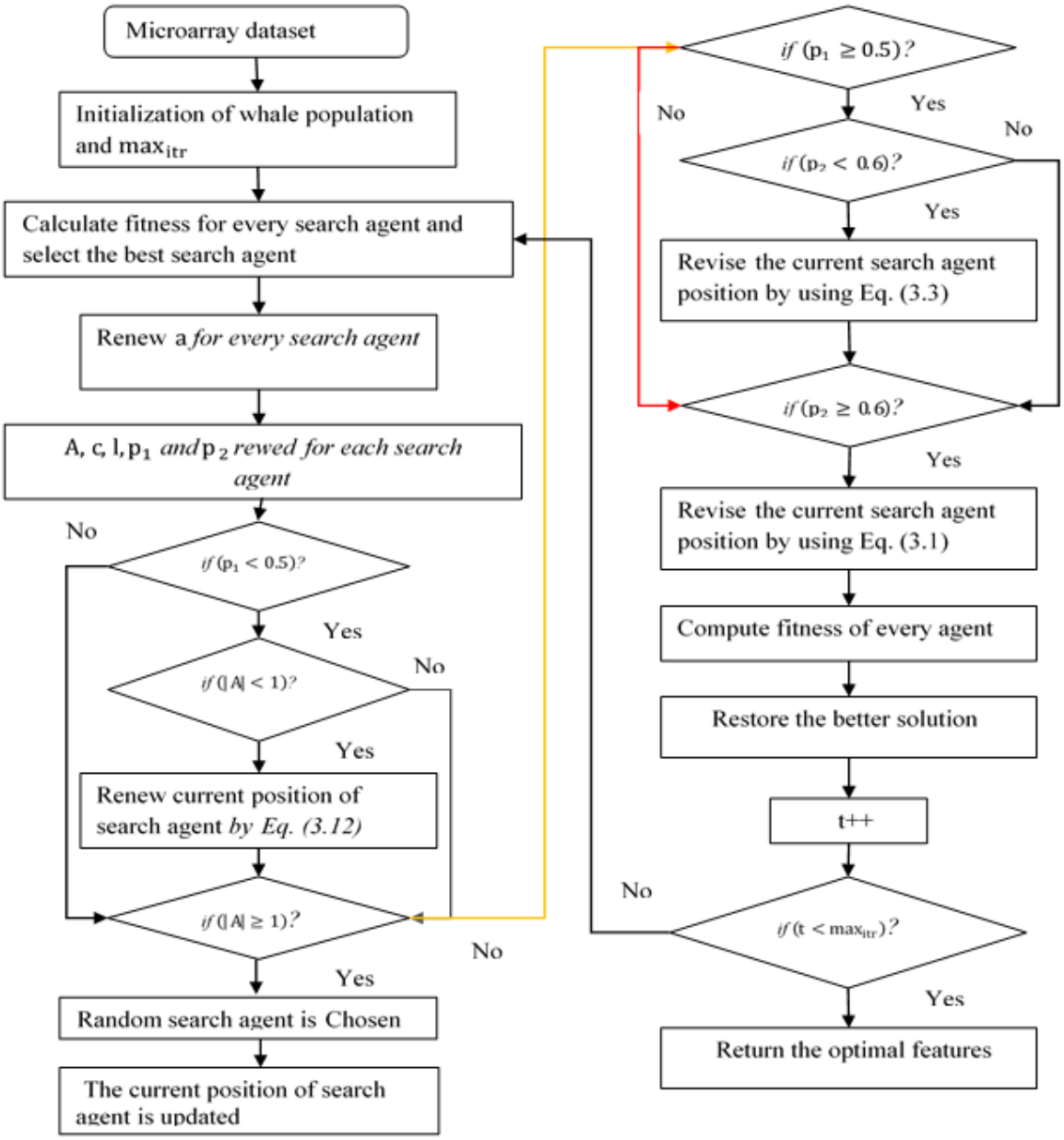

Fig 1. Overall flow of SMWOA

4 Results and Discussion

The work of this part shows the performance of MWOA and SMWOA with different classifiers for cancer detection which is analyzed in terms of accuracy, specificity, sensitivity, F1-score and average error. For the experimental purpose, three microarray

datasets such as Leukemia, Lymphoma and prostate microarray datasets are used. These datasets are publicly available on the internet. Dataset of Leukemia has 72 instances, features count as 3572 and classes count as 2, lymphoma dataset consists of 77 instances, 2647 features, and 2 classes. Prostate dataset has 102 cases, 2135 features and 2 classes. The execution of this pattern is evaluated on the testing and training datasets. The datasets are divided into testing and training set in the ratio of 60:40. Table 1 shows the number of features selected by MWOA and SMWOA methods.

Table 1. Features selected by MWOA and SMWOA

	No of features	MWOA	SMWOA
Leukemia	3572	43	38
Prostate	2135	135	110
Lymphoma	2647	39	33

4.1 Accuracy

Accuracy gives the various records in the microarray data which are correctly classified among all number of records in the dataset. It is calculated as

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

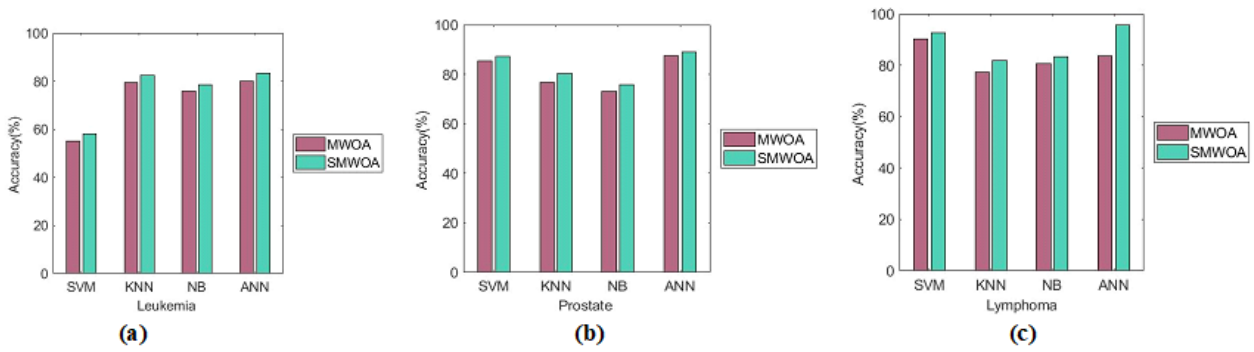

Fig 2. Comparison of Accuracy

[Figure 2] gives the accuracy between MWOA and SMWOA with different classifiers for cancer detection. X-axis stands for the classifiers and Y-axis stands for the accuracy based on feature selection methods. For the leukemia dataset, the accuracy of SMWOA-ANN is 51.30% greater than MWOA-SVM, 5.25% greater than MWOA-KNN, 10.03% greater than MWOA-NB, 4.18% greater than MWOA-ANN, 43.32% greater than SMWOA-SVM, 1.41% greater than SMWOA-KNN and 6.13% greater than SMWOA-NB. Similarly, the accuracy of SMWOA-ANN is 4.29% greater than MWOA-SVM, 16.21% greater than MWOA-KNN, 21.66% greater than MWOA-NB, 1.69% greater than MWOA-ANN, 2.02% greater than SMWOA-SVM, 10.69% greater than SMWOA-KNN and 17.33% greater than SMWOA-NB for prostate dataset. For the lymphoma dataset, the accuracy of SMWOA-ANN is 5.86% greater than MWOA-SVM, 23.52% greater than MWOA-KNN, 18.58% greater than MWOA-NB, 14.34% greater than MWOA-ANN, 3.31% greater than SMWOA-SVM, 16.85% greater than SMWOA-KNN and 14.61% greater than SMWOA-NB. From this analysis it is concluded that the planned SMWOA-ANN method produces high accuracy than the methods for cancer detection.

4.2 Specificity

The specificity in the clinical test can identify correct people without illness, within all people free from illness. The following formula is used to calculate,

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

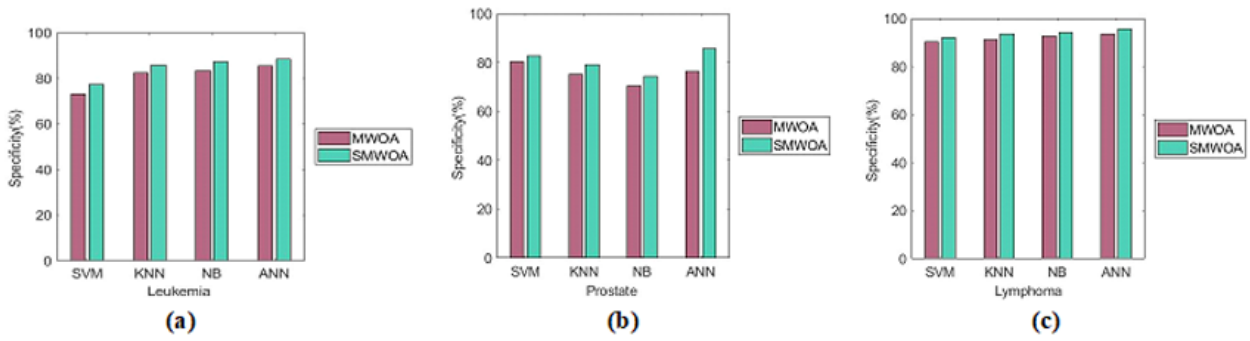

Fig 3. Comparison of Specificity

Figure 3 results the specificity inbetween MWOA and SMWOA with different classifiers for cancer detection. X-axis expresses the classifiers and Y-axis expresses the specificity about feature selection methods. For the leukemia dataset, the specificity of SMWOA-ANN is 20.86% greater than MWOA-SVM, 7.49% greater than MWOA-KNN, 6.30% greater than MWOA-NB, 3.58% greater than MWOA-ANN, 14.59% greater than SMWOA-SVM, 3.27% greater than SMWOA-KNN and 1.50% greater than SMWOA-NB. Similarly, the specificity of SMWOA-ANN is 6.37% greater than MWOA-SVM, 13.65% greater than MWOA-KNN, 21.31% greater than MWOA-NB, 11.72% greater than MWOA-ANN, 3.50% greater than SMWOA-SVM, 7.89% greater than SMWOA-KNN and 15.15% greater than SMWOA-NB for prostate dataset. For the lymphoma dataset, the specificity of SMWOA-ANN is 5.99% greater than MWOA-SVM, 5.14% greater than MWOA-KNN, 3.36% greater than MWOA-NB, 2.44% greater than MWOA-ANN, 3.92% greater than SMWOA-SVM, 2.47% greater than SMWOA-KNN and 1.59% greater than SMWOA-NB. From this analysis it is proven that the executed SMWOA-ANN method has high specificity than alternative methods used for cancer detection.

4.3 Sensitivity

Sensitivity in the clinical test can find out exact people with the illness. It produces the proportion of people with the disease who are positive, expressed in percentages. It is calculated as,

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

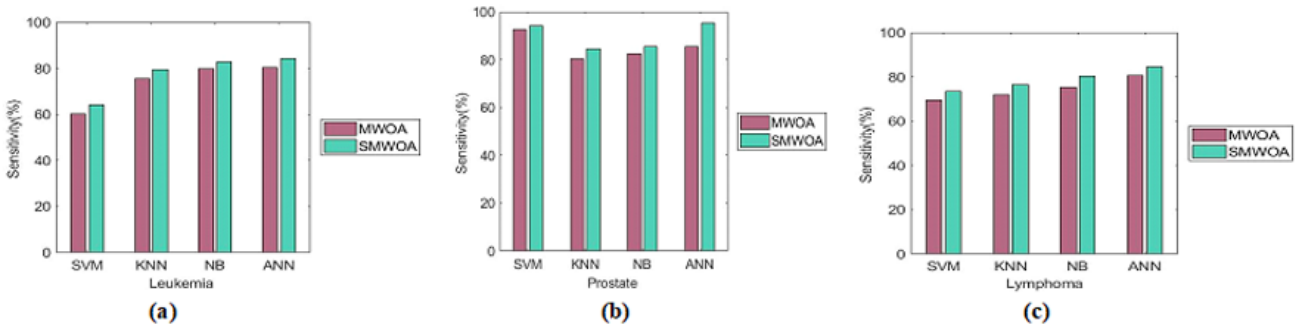

Fig 4. Comparison of Sensitivity

[Figure 4] displays the sensitivity between MWOA and SMWOA with different classifiers for cancer detection. X-axis shows the classifiers and Y-axis shows the sensitivity based on feature selection methods. For the leukemia dataset, the sensitivity of SMWOA-ANN is 39.77% greater than MWOA-SVM, 11.37% greater than MWOA-KNN, 5.84% greater than MWOA-NB, 4.99% greater than MWOA-ANN, 31.11% greater than SMWOA-SVM, 6.11% greater than SMWOA-KNN and 1.97% greater than SMWOA-NB. Similarly, the sensitivity of SMWOA-ANN is 2.81% greater than MWOA-SVM, 18.69% greater

than MWOA-KNN, 15.34% greater than MWOA-NB, 11.44% greater than MWOA-ANN, 1.27% greater than SMWOA-SVM, 12.51% greater than SMWOA-KNN and 11.57% greater than SMWOA-NB for prostate dataset. For the lymphoma dataset, the sensitivity of SMWOA-ANN is 21.94% greater than MWOA-SVM, 17.50% greater than MWOA-KNN, 12.33% greater than MWOA-NB, 5.08% greater than MWOA-ANN, 15.23% greater than SMWOA-SVM, 10.87% greater than SMWOA-KNN and 5.35% greater than SMWOA-NB. From this analysis it is proved that the planned SMWOA-ANN method has high sensitivity than another methods for cancer detection.

4.4 F1-score

F1-score is used to calculate harmonic mean of precision and recall. It is done by using,

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

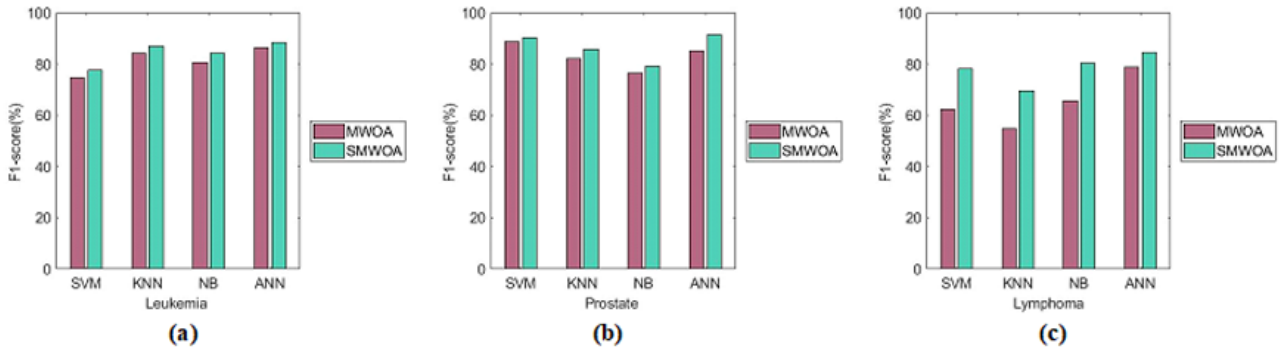

Fig 5. Comparison of F1-score

[Figure 5] produces the F1-score among MWOA and SMWOA with different classifiers for cancer detection. X-axis shows the classifiers and Y-axis shows the F1-score based on feature selection methods. For the leukemia dataset, the F1-score of SMWOA-ANN is 18.61% greater than MWOA-SVM, 5.02% greater than MWOA-KNN, 9.58% greater than MWOA-NB, 2.57% greater than MWOA-ANN, 14.10% greater than SMWOA-SVM, 1.88% greater than SMWOA-KNN and 4.90% greater than SMWOA-NB. Similarly, the F1-score of SMWOA-ANN is 3.14% greater than MWOA-SVM, 11.43% greater than MWOA-KNN, 19.72% greater than MWOA-NB, 7.38% greater than MWOA-ANN, 1.43% greater than SMWOA-SVM, 6.88% greater than SMWOA-KNN and 15.52% greater than SMWOA-NB for prostate dataset. For the lymphoma dataset, the F1-score of SMWOA-ANN is 35.20% greater than MWOA-SVM, 53.87% greater than MWOA-KNN, 28.94% greater than MWOA-NB, 7.12% greater than MWOA-ANN, 7.88% greater than SMWOA-SVM, 21.53% greater than SMWOA-KNN and 5.22% greater than SMWOA-NB. From this analysis it is resulted that the executed SMWOA-ANN method has high F1-score than related to methods used 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 MWOA and SMWOA.

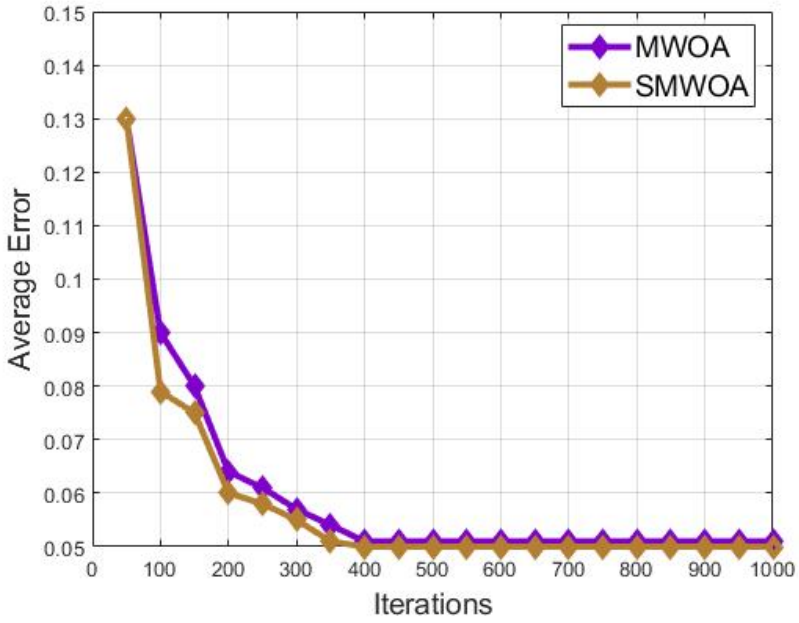

Fig 6. Average error comparison

[Figure 6] shows the average error of classifiers for cancer detection with the selected features by MWOA and SMWOA. X-axis stands for the number of iteration and Y-axis shows the average error of classifier. When the number of iteration is 100, the average error of SMWOA is 12.22% less than MWOA. From this analysis it is proved that the proposed SMWOA method has less average error than MWOA methods for cancer detection.

5 Conclusion

The proposed SMWOA accelerates the convergence speed of classifiers, enhances cancer detection accuracy and effectively improves the exploration and exploitation strategy of MWOA. LF and quadratic implementation are introduced in SMWOA to enhance the classification accuracy. With the help of Lévy flight MWOA jump or breaks out local optima to the repeated short-distance search step, rarely the longer-distance search step, quadratic interpolation solution accuracy is improve by enhancing the exploitation ability. In addition to, a self-adaptive control framework has gained improvement which balances the intervals among local exploitation and global exploration. This experiment concludes that the proposed SMWOA-ANN gives improvement in terms of accuracy, specificity, sensitivity, F1-score and average error than other methods for microarray data classification. However, SMWOA may get stuck in a part of the Pareto-optimal problem since multiple objectives are used in SMWOA. In the future, an efficient technique will be introduced to solve the Pareto-optimal problem and get a better final subset of features for microarray data classification.

References

1) Aziz R, Verma CK, Srivastava N. Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*. 2017;4(1):179–197. Available from: https://doi.org/10.3934/bioeng.2017.2.179.
2) Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*. 2010;43(1):15–23. Available from: https://doi.org/10.1016/j.jbi.2009.07.008.
3) Kumar M, Rath NK, Swain A, Rath SK. Feature selection and classification of microarray data using MapReduce based ANOVA and K-Nearest neighbor. *Procedia Computer Science*. 2015;54:301–310. Available from: https://doi.org/10.1016/j.procs.2015.06.035.
4) Elsebakhi E, Asparouhov O, Al-Ali R. Novel incremental ranking framework for biomedical data analytics and dimensionality reduction: Big data challenges and opportunities. *Journal of Computer Science & Systems Biology*. 2015;8(4):203–214. Available from: https://doi.org/10.4172/jcsb.1000190.

5) Yang CH, Chuang LY, Yang CH. IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. *Journal of Medical and Biological Engineering*. 2010;30(1):23–28.
6) Huerta EB, Duval B, Hao JK. A hybrid GA/SVM approach for gene selection and classification of microarray data. In: Workshops on Applications of Evolutionary Computation Springer. 2006;p. 34–44. Available from: https://doi.org/10.1007/11732242_4.
7) Dashtban M, Balafar M, Suravajhala P. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics*. 2018;110(1):10–17. Available from: https://doi.org/10.1016/j.ygeno.2017.07.010.
8) Gao W, Hu L, Zhang P, Wang F. Feature selection by integrating two groups of feature evaluation criteria. *Expert Systems with Applications*. 2018;110:11–19. Available from: https://doi.org/10.1016/j.eswa.2018.05.029.
9) Sathya M, Priya M, S. PSO search-based feature selection method for high dimensional data. *International Journal of Recent Technology and Engineering (IJRTE)*. 2019;7(583):485–488.
10) Zhai QH, Ye T, Huang MX, Feng SL, Li H. Whale optimization algorithm for multiconstraint second-order stochastic dominance portfolio optimization. 2019. Available from: https://doi.org/10.1155/2020/8834162.
11) Mohammadzadeh H, Gharehchopogh FS. A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection. 2020. Available from: https://doi.org/10.1111/coin.12397.
12) Feng Y, Chen H, Li T, Luo C. A novel community detection method based on whale optimization algorithm with evolutionary population. *Applied Intelligence*;2020:1–20. Available from: https://doi.org/10.1007/s10489-020-01659-7.
13) Sahu PR, Hota PK, Panda S. Modified whale optimization algorithm for coordinated design of fuzzy lead-lag structure-based SSSC controller and power system stabilizer. *International Transactions on Electrical Energy Systems*. 2019;29(4):e2797–e2797. Available from: https://dx.doi.org/10.1002/etep.2797.
14) Nasiri J, Khiyabani FM. A whale optimization algorithm (WOA) approach for clustering. *Cogent Mathematics & Statistics*. 2018;5(1). Available from: https://doi.org/10.1080/25742558.2018.1483565.
15) Gharehchopogh FS, Gholizadeh H. A comprehensive survey: Whale Optimization Algorithm and its applications. 2019. Available from: https://doi.org/10.1016/j.swevo.2019.03.004.
16) Sathya M, Priya M, S. Modified whale optimization algorithm for feature selection algorithm in microarray cancer datasets. *International Journal of Scientific & Technology Research*. 2020;1(1).
17) Sun S, Peng Q, Zhang X. Global feature selection from microarray data using Lagrange multipliers. 2016. Available from: https://doi.org/10.1016/j.knosys.2016.07.035.
18) Sahu B, Dehuri S, Jagadev AK. Feature selection model based on clustering and ranking in pipeline for microarray data. *Informatics in Medicine Unlocked*. 2017;9:107–122. Available from: https://dx.doi.org/10.1016/j.imu.2017.07.004.
19) Das AK, Das S, Ghosh A. Ensemble feature selection using bi-objective genetic algorithm. 2017. Available from: https://doi.org/10.1016/j.knosys.2017.02.013.
20) Medjahed SA, Saadi TA, Benyettou A, Ouali M. Kernel-based learning and feature selection analysis for cancer diagnosis. *Applied Soft Computing*. 2017;51:39–48. Available from: https://dx.doi.org/10.1016/j.asoc.2016.12.010.
21) Prasad Y, Biswas KK, Hanmandlu M. A recursive PSO scheme for gene selection in microarray data. *Applied Soft Computing*. 2018;71:213–225. Available from: https://doi.org/10.1016/j.asoc.2018.06.019.
22) Jain I, Jain VK, Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing*. 2018;62:203–215. Available from: https://doi.org/10.1016/j.asoc.2017.09.038.
23) Lai CM. Multi-objective simplified swarm optimization with weighting scheme for gene selection. *Applied Soft Computing*. 2018;65:58–68. Available from: https://doi.org/10.1016/j.asoc.2017.12.049.
24) Yuan M, Yang Z, Ji G. Partial maximum correlation information: A new feature selection method for microarray data classification. *Neurocomputing*. 2019;323:231–243. Available from: https://dx.doi.org/10.1016/j.neucom.2018.09.084.
25) Yan C, Ma J, Luo H, Patel A. Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. 2019. Available from: https://doi.org/10.1016/j.chemolab.2018.11.010.
26) Ca B, Zhao J, Yang P, Yang P, Li X, Qi J, et al. Multiobjective feature selection for microarray data via distributed parallel algorithms. *Future Generation Computer Systems*. 2019;100:952–981. Available from: https://doi.org/10.1016/j.future.2019.02.030.
27) Pashaei E, Pashaei E, Aydin N. Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization. *Genomics*. 2019;111(4):669–686. Available from: https://dx.doi.org/10.1016/j.ygeno.2018.04.004.
28) Momenzadeh M, Sehhati M, Rabbani H. A novel feature selection method for microarray data classification based on hidden Markov model. *Journal of Biomedical Informatics*. 2019;95. Available from: https://dx.doi.org/10.1016/j.jbi.2019.103213.
29) Gharehchopogh FS, Khaze SR, Maleki I. A New Approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network Algorithms. *Indian Journal of Science and Technology*. 2015;8(3):237–246. Available from: https://doi.org/10.17485/ijst/2015/v8i3/59570.
30) Dash R, Dash R, Rautray R. An evolutionary framework based microarray gene selection and classification approach using binary shuffled frog leaping algorithm. 2019. Available from: https://doi.org/10.1016/j.jksuci.2019.04.002.
31) Potharaju SP, Sreedevi M. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. *Clinical Epidemiology and Global Health*. 2019;7(2):171–176. Available from: https://dx.doi.org/10.1016/j.cegh.2018.04.001. doi:10.1016/j.cegh.2018.04.001.
32) Gangavarapu T, Patil N. A novel filter-wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets. *Applied Soft Computing*. 2019;81. Available from: https://doi.org/10.1016/j.asoc.2019.105538.
33) Bonab MS, Ghaffari A, Gharehchopogh FS, Alemi P. A wrapper-based feature selection for improving performance of intrusion detection systems. *International Journal of Communication Systems*. 2020;33(12). Available from: https://dx.doi.org/10.1002/dac.4434.
34) Rahnema N, Gharehchopogh FS. An improved artificial bee colony algorithm based on whale optimization algorithm for data clustering. *Multimedia Tools and Applications*. 2020;79(43-44):32169–32194. Available from: https://dx.doi.org/10.1007/s11042-020-09639-2.