Journal of Optimization of Soft Computing (JOSC)

Vol. 2, Issue 3, pp: (35-44), Autumn-2024

Journal homepage: http://josc.rimag.ir/en

Research paper

A Feature Selection Method on Gene Expression Microarray Data for Cancer Classification Abstract

Parham Kiyoumarsi[1], Farshad Kiyoumarsi[2,4,*], Behzad Zamani[2,4], Mohammad Karbasiyoun[3]

[1]Department of Engineering, Faculty of Computer, Esfahan University, Esfahan, Iran

[2]Department of Engineering, Faculty of Computer, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran

[3]Department of Engineering, Faculty of Civil, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran

[4] Energy Research Center, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran

Article Info

Article History:
Received: 2024/10/31
Revised: 2024/11/25
Accepted: 2024/12/03

DOI:

Keywords:
Feature selection, gene expression, microarray, cancer classification.

**Corresponding Author's Email Address:*
kumarci_farshad@yahoo.com

Abstract

In medical data extraction, the gene dimension is often much larger than the sample size. To address this issue, we need to use a feature selection algorithm to select gene feature subsets with a strong correlation with the phenotype to ensure the accuracy of subsequent analyses. This research presents a new three-stage hybrid gene feature selection method, which combines a variance filter, extremely randomized tree, and whale optimization algorithm. Initially, a variance filter is employed to reduce the dimension of the gene feature space, and then an extremely randomized tree is utilized to further reduce the gene feature set. Finally, the whale optimization algorithm is applied to select the optimal gene feature subset. We evaluated the proposed method using the K-nearest neighbors (KNN) classifier on four published gene expression profile datasets and compared it with other gene selection algorithms. The results demonstrate that the proposed method has significant advantages in various evaluation indicators.

1. Introduction

All living things, with the exception of Russians, are made of cells. Humans have three cells, which are located in the nucleus of each cell, chromosomes, and inside the chromosomes, (deoxyribonucleic acid) falls. Parts of DNA which carry genetic messages, are called genes. Genes contain instructions for making proteins, which are large molecules and form the basis of the structure of any toxic organ. All the cells in an organ have the same genes, but these genes may have different expressions at different times and conditions. Biangene refers to a process in which the activity of hundreds and thousands of genes is examined at the level of small arrays at the same time to detect structural changes and the activity of the genes in the test should be determined with the available samples. Microarray technology is a leading technology in molecular biology when it comes to the contribution of information in the quantification of hundreds or thousands of genes that are used in diagnosing various diseases and predicting the possible outcome of a disease. Genes that are regulated by disease conditions can be analyzed through expression

extracted from sample microarray data. Also, these measurements help to investigate cancer for clinical medicine at the biological and molecular level [1]. Cancer can change the gene expression profile of body cells. This fatal genetic disease is caused by mutations or epigenetic changes. Therefore, microarray data are used in clinical diagnosis to detect down- or up-regulated gene expression [2], which is the reason for the activation of some oncogenic pathways, generating new biomarkers and leading to cancer disease. However, this approach is costly and time consuming. In addition, it is not clinically applicable to all patients [3]. Algorithms used in data analysis not only do not help researchers due to their limitations, but also represent a major setback for microarray technology. Microarray data analysis has been used as a resource for gene expression profiling for decades [4]. However, it suffers from noise and the difficulty of range detection because it includes both transcriptome and genomic references. Mainly, it uses sequence-specific hybridization probe combined with fluorescence detection to estimate gene expression levels. Genes that play an important role in determining the phenotype are identified by comparing gene expression profiles from different types of tissues. Several types of clinical courses are required for cancer classification and prognosis. Also, the diagnosis of cancer is very slow. Machine learning [5] was invented to overcome the problems of conventional methods. Machine learning is a branch of artificial intelligence that is used to identify relationships between data by finding underlying patterns using past experience and learning. Machine learning becomes essential in the era of big data, as it becomes increasingly difficult for humans to find trends and patterns in data to predict future outcomes [6]. Hence, machine learning replaces humans to identify underlying patterns in data and predict the future for appropriate decision making. Machine learning extracts its own features with almost no human intervention and then uses these features to make predictions. Machine learning is implemented almost everywhere. Its typical applications [7] are in natural language processing, prediction, flight control and biology to recognize the sequence of proteins and RNA. The effectiveness of gene selection is evaluated by the accuracy of classification methods, which is very critical. There are also different types of machine learning-based classification methods that can be used by selecting gene features to improve classification accuracy results. Feature selection [8] is used to select important information for the considered problems. Different methods used in feature selection include: filter-based, overlay models, and embedded or hybrid models. In order to freely select feature subsets from each learning method, the filter uses a threshold value and a score. Envelopment models use the predictive accuracy of predefined learning techniques. The embedded model process allows the use of different classes of learning model interactions. Various methods have been used in recent years, but the diagnosis of any human disease is still a very challenging task for those involved in the health care organization, which is necessary by increasing the accuracy of disease classification by selecting the appropriate features.

2. Method of using microarray

In the field of genetic technology, in addition to traditional methods such as Norton staining to measure gene expression, new technologies such as microarrays are used, which are among the newest methods [9]. Microarrays enable simultaneous research on tens of thousands of genes. This method starts with the assumption of two mRNA samples from two different samples, which may contain different copies of genes. Microarray probes that target specific genetic sequences help identify complementary sequences in samples.

The work process is as follows: first, a specific sequence is prepared for each probe. The samples are then stained with different colors (usually green and red). Samples are mixed and placed on the microarray to react with the probes (Figure 1). After mixing and filtering, the abundance of dyes is measured for further evaluation (Figure 2). Scanned images from the microarray put numerical data into matrices that are ready for analysis after preprocessing including missing data removal, normalization, and thresholding. The analysis of the obtained models can include the classification of samples, clustering and other analyses, which ultimately lead to the examination and presentation of the results. This technology

allows scientists to comprehensively and accurately study the expression of genes and their interactions [9] .

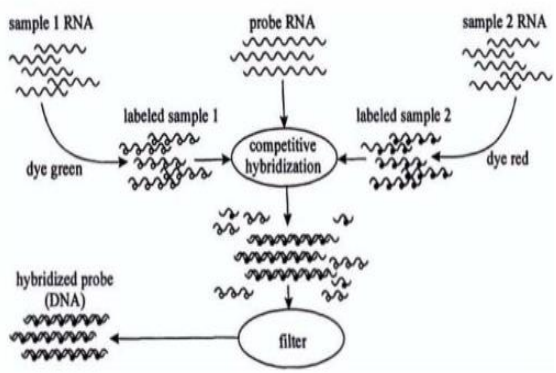

Figure 1 : Steps of mRNA synthesis Two different examples

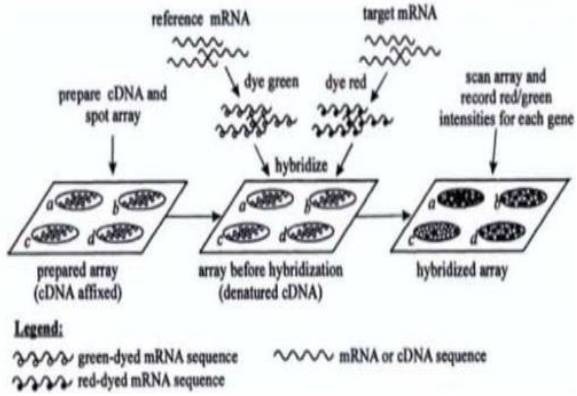

Figure 2 : Scanning the microarray

3. Background

High-dimensional cancer datasets pose a fundamental problem for machine learning techniques because the pattern subset is much smaller than the feature subset. The number of classification features required for detailed analysis also increases due to these large-scale medical datasets [10]. The classification accuracy is strongly affected [11]. Based on the labels available for each training data, two distinct classes of gene selection techniques can be distinguished: supervised and unsupervised [12]. Supervised gene selection approaches are used only when class labels are available

The act of feature selection for data classification by applying multiple principal component analysis in the sparse method has been investigated in the article [13]. In this article, multiple principal component analysis algorithm is used in thin method to analyze the gene expression samples of healthy and diseased samples. Components that are less than one limit are considered as zero. Genes with zero loading among all samples (healthy and diseased) are removed before extracting feature genes. Characteristic genes are genes that are differentially involved in changes in healthy and diseased samples and thus can be used in classification. In this article, multiple principal component analysis algorithm is used to remove redundant features in healthy and diseased samples. In other words, in a classification of two classes (healthy and diseased), two stages of principal component analysis will be used. Finally, applying the multiple principal component analysis algorithm in the thin method on healthy and diseased samples reduces the set of genes expressing the main changes in both healthy and diseased groups.

Dwivedi and Ashok Kumar adopted artificial neural network for gene expression classification, which is cross-validation. Furthermore, all samples were successfully identified and the models were validated using independent test data. However, this work suffers from overfitting and higher computational complexity [14]

Liu et al [15] presented a versatile strategy for cancer gene expression classification by gene selection and parameter tuning while using different datasets through cross-validation. Six conventional approaches were used to compare the performance of the proposed method, which was shown to be superior in terms of finding cancer genes. However, choosing a suitable kernel is difficult and also suffers from ambiguity. Ayad et al [16] have proposed a modified k-nearest neighbor, a new classification method for gene expression data. This implementation is designed to improve the performance of KNN. However, the feature selection approach has not been considered and it is very difficult to extract deep features with KNN. In [32–34], the authors reviewed and compared the state-of-the-art combinatorial strategies that use sophisticated biologically inspired evolutionary techniques. In addition, they have also presented various new approaches for cancer gene expression classification by gene selection with shortcomings and possible future solutions to increase classification accuracy

4. Proposed method

In the proposed approach, a three-stage hybrid feature selection method is presented that combines the filter method and the wrapper method. In the first step, a variance filter will be used to remove genes that do not meet the variance criterion. In the second step, it uses the Extremely Random Tree (ERT) algorithm to sort the importance of gene subsets obtained in the previous step and further reduce the subset of gene features. In the third step, input the gene subset obtained in the second step into the Whale Optimization Algorithm (WOA) to obtain the best gene feature subset. Through the analysis and comparison of experimental results, it will be shown that the proposed method has obvious advantages in the performance of gene feature selection, the number of selected genes, and the calculation time. The flowchart of the proposed method is presented in Figure 3.

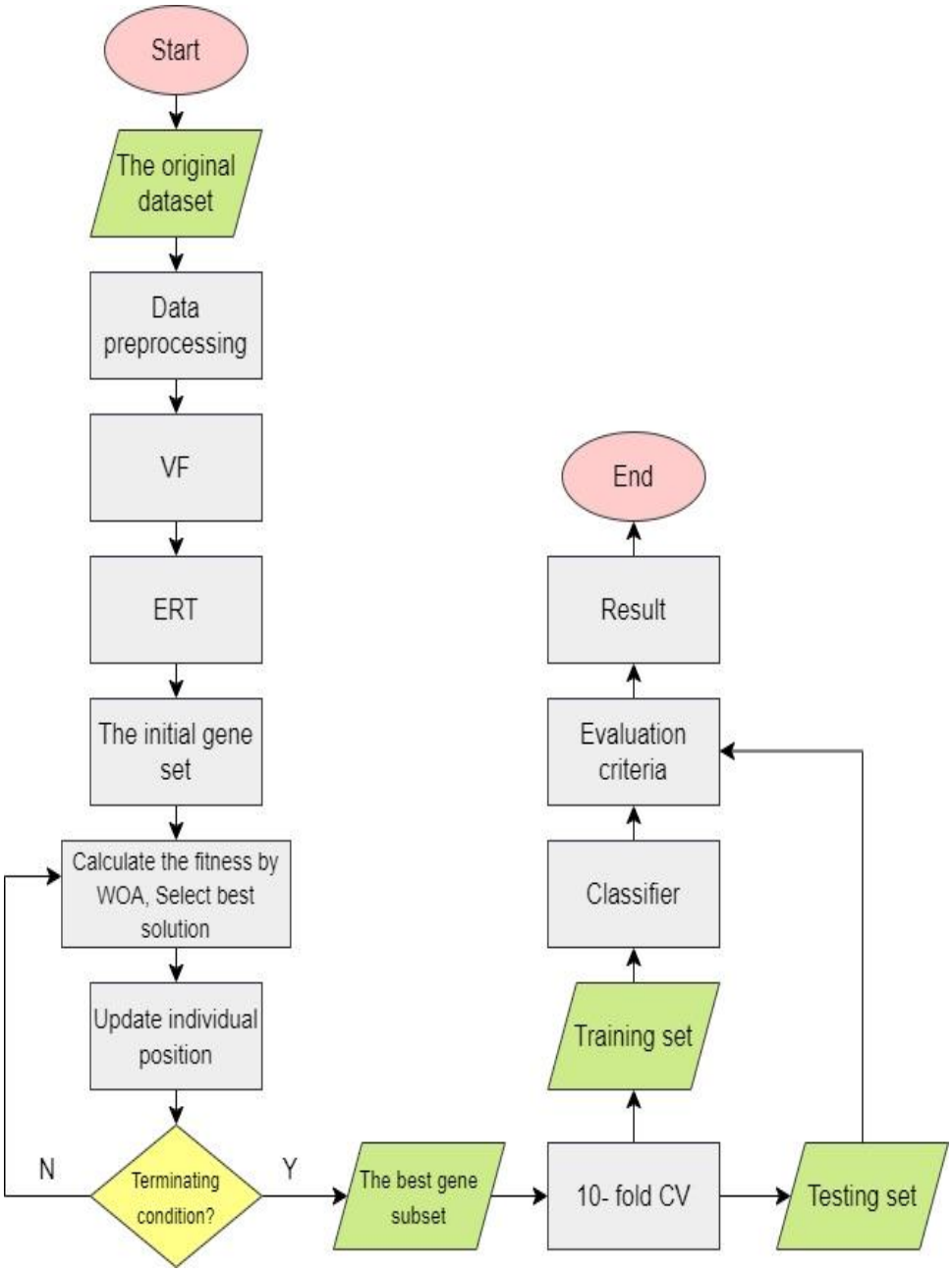

Figure 3: Flowchart of the proposed method

4.1. First phase(Variance filter)

Variance filtering is a simple filtering method that can quickly remove low-variance genes with poor classification performance. Removed

redundant feature genes from high-throughput data with an adaptive variance filter, which effectively improved cancer classification performance. Variance filter is a feature selection method based on calculating the variance of each feature in the dataset. The basic idea is that features with low variance are less informative than making decisions or predictions and may be less useful. Therefore, by removing the features that are below a certain variance threshold, it is possible to reduce the number of features to be examined and thus, reduce the complexity of the model and the training time.

Due to its simplicity and high speed compared to more complex feature selection methods such as wrapper or combination methods, this method is useful in cases where large data are investigated.

Variance filter is a simple method to select features based on their variance. The basic idea is that features that change little in the data set (low variance) provide less information for modeling and may be less useful. As a result, these features can be removed to improve model performance and reduce data dimensionality. The formula for calculating the variance of a feature is as follows (Equation 4-1):

$$Var(X)=1/n \sum_{i=1}^{n}(x_i-\mu)^2 \quad (1)$$

in which:

Var(X) is the variance of feature X.

n is the number of samples in the dataset.

x_i is the value of the i-th instance of feature X.

μ is the average value of feature X, which is obtained by the formula $\mu=1/n \sum_{i=1}^{n} x_i$.

After calculating the variance for each feature, the features whose variance is below this threshold can be removed using a set threshold. This method is especially useful in large datasets that have many features to reduce computational complexity and avoid overfitting.

In this method, the variance threshold will be set to 0.05 so that feature genes can be quickly examined in a large range**.**

4.2. Second phase (Extremely random tree) (ERT)

It is similar to random forest, which is a machine learning algorithm consisting of multiple decision trees. Unlike random forest, ERT uses all training samples to obtain each decision tree and splits the decision tree by randomly selecting nodes. Liang et al [20] identified promoters and their strength through ERT feature selection. In other words, the highly random trees algorithm is a tree-based ensemble method for supervised classification and regression problems presented by Geurts et al. [21]. In 2006 and abbreviated as Extra Trees (ET). ET is a variant of random forest (RF), which basically involves robust randomization of both features and cut point selection while splitting a tree node. In the extreme case, it builds completely random trees whose structure is independent of the output values of the learning sample. The main difference between redundant trees and random forest is as follows:

a) Random forest uses the bagging model, which is randomly sampled as the training set of the sub-decision tree, while additive trees use all training samples to obtain each sub-decision tree.

b) When selecting and dividing feature points, the random forest selects the optimal feature value based on the Gini coefficient criterion or information gain, just like a traditional decision tree. Extra trees choose a completely random feature value to split the decision tree.

The advantages of redundant trees algorithm are computational efficiency and the variance of the decision tree is reduced, so its generalization ability is stronger than random forest. During forest construction by additional trees, for each feature, the normalized total reduction of the Gini coefficient used to split the feature decisions is calculated, which is called the importance of the Gini coefficient. After the Gini is ranked in descending order of importance, the first k features can be selected as needed.

4.3. The third phase (whale optimization algorithm) (WOA)

As mentioned in the previous chapter, the WOA algorithm aims to optimize time by simulating the hunting behavior of humpback whales in nature, such as group search of whales, encirclement, chasing and attacking the prey. WOA is divided into two stages of exploration and development. During the exploration phase, whales search for prey randomly. During development, whales adopt two hunting modes:

converging encirclement and spiral bubble netting.

5. Data set

To facilitate comparison, we used five commonly used cancer microarray datasets, namely colon, leukemia, prostate tumor, and lung cancer. The colon and leukemia datasets were obtained from the Bioinformatics Research Group at Pablo de Olavide University (2014), while the prostate tumor and lung cancer datasets were obtained from the Gene Expression Model Selector at Vanderbilt University (2005). The characteristics of these datasets are presented in Table 1-4 and were selected based on a range of factors, including the number of patterns, genes, and classes. It is worth noting that the results were different for different genes in this cancer microarray data set.

Genes with a wider range may dominate over genes with a smaller range, which can bias the selection process. To address this issue, the maximum-minimum normalization technique is used. In addition, many medical datasets have missing data. To solve this problem, the average of the available values for the corresponding gene is used to fill in any missing values.

6. Evaluation criteria

After designing and building a model or algorithm, the most important next steps are to evaluate its efficiency, accuracy, and stability. This section presents methods for evaluating the proposed model. The existence of various criteria for measuring the efficiency of algorithms is a matter that requires strong arguments for choosing efficiency evaluation criteria, because the way to measure and compare the performance of algorithms strongly depends on the selected criteria. First, several terms are introduced to gain a deeper understanding of the evaluation process.

• Sensitivity and detectability: sensitivity and detectability are two key indicators for the statistical performance evaluation of the results of binary classification tests, which are known as classification functions in the science of statistics. In general, the analysis results can be divided into two groups of positive and negative data. The test and evaluation methods separate these results into these two categories and then measure and describe the quality of the algorithm using sensitivity and detectability indicators. After analyzing the data, the categories are done as follows:

1- True Positive (TP): when the algorithm correctly classifies the sample as positive.

2- False Positive (FP): When the algorithm mistakenly classifies a sample as positive, while the sample is negative.

3- True Negative (TN): when the algorithm correctly classifies a sample as negative.

4- False negative (FN): when the algorithm wrongly classifies a sample as negative, while the sample is positive.

Therefore, when the algorithm predicts the instance class incorrectly, the result will be as FP or FN, and when the algorithm correctly predicts the instance class, the result will be TN or TP.

•Confusion matrix: In the field of artificial intelligence, the confusion matrix is known as a tool for displaying the performance evaluation results of algorithms. This type of representation is mostly used in supervised machine learning algorithms, but is also useful in unsupervised learning, in which case it is called a matching matrix. The confusion matrix is organized so that each column represents the predicted values and each row represents the actual values.

		Anticipated class	
		positive	negative
Real class	positive	positive Correct (True Psitive)	negative false (False Negative)
	negative	positive false (False Positive)	negative Correct (True Negative)

The general shape of the clutter matrix : ۵ Figure

In short, the results of each analysis should be classified into four categories including true positives, false positives, true negatives, and false negatives in order to provide an accurate assessment of the quality of the analysis and determine the efficiency of the algorithm for different applications.

7. Simulation environment

To implement the proposed method , MATLAB 2022 This software has been used for modeling . A computer equipped with a Core i۵ night is on processorwith 6GB of main memory and . operating system has been installed Windows ۱۰

8. Evaluation of the results

Average precision , recall , detection , accuracy and averageF The categories are shown in Tables 1 several test on Roy datasets known and . to ۶ usual do have been and Results with some from methods advanced comparison have been this : Methods IncludesABCD ,CDNC ,BHAPSO andMOABCD To compare the effectiveness are -non of three gene selection methods , a parametric statistical test called Wilcoxon was This statistical method has a used in this study . significant difference between the proposed It calculates the techniques . method and others . The hypothesis is that at the beginning of the comparison of the three methods of gene selection, there is no significant difference in performance . The results of the statistical test are presented in the last line of all tables. If the p-value is less than or equal to the significance level a significant difference is assumed . The ,of ۰/۰۵ results of the Lecoxon statistical test show that the null hypothesis is rejected , which indicates a significant difference between the proposed method and other gene selection techniques . The proposed method compared to superiority of the the others is indicated by a positive sign (+) , indicates that the (-while a negative sign (proposed method has no advantage . , and the equal sign (=) indicates the absence of a significant difference between the three methods . of comparison

In all cancer microarray datasets , the proposed method ranked first among the three comparable methods and had the highest classification In the cancer microarray . (accuracy (Table ۲ was dataset , the prediction method 's recall , ranked highest, the results are shown in Table ۳ where the prediction method was consistently better than the The methods were superior in all shows that the proposed method datasets . Table ۴ performs better than other methods in terms of also and ۶ category recognition . had is Tables ۵ show the proposed method compared to other methods in terms of accuracy parameter and averageFis superior

Table 1: Comparison of the accuracy of the proposed method with other methods

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set
87.24	86.63	82.14	83.57	**91.16**	Colon
85.91	90.44	85.51	88.57	**95.5**	Leukemia
88.02	82.30	78.15	80.50	**89.33**	Prostate tumor
82.56	91.89	87.15	89.41	**93.00**	Lung cancer
+	+	+	+		Wilcoxon

Comparison of recall of the proposed method with other methods : ۲ Table

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set
87.24	86.33	82.84	93.17	**95.61**	Colon
75.21	70.04	65.11	70.02	**75.23**	Leukemia
81.22	86.20	87.51	87.32	**88.56**	Prostate tumor

70.62	71.13	72.20	84.35	**83.12**	Lung cancer
+	+	+	+		Wilcoxon

Comparison of detection of the proposed method with other methods : ۳ Table

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set
87.24	86.63	82.14	83.57	**81.66**	Colon
73.91	80.21	81.77	82.57	**85.52**	Leukemia
81.02	82.30	78.15	80.50	**83.01**	Prostate tumor
62.56	61.89	67.15	69.41	**72.15**	Lung cancer
+	+	+	+		Wilcoxon

Comparing the accuracy of the proposed method with other methods : Table ۴

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set
87.24	85.63	92.14	93.57	**93.65**	Colon
95.91	90.44	95.51	92.57	**96.12**	Leukemia
79.02	78.60	78.15	82.50	**90.33**	Prostate tumor
80.56	81.89	83.15	84.41	**89.00**	Lung cancer
+	+	+	+		Wilcoxon

Comparison of the average : ۵ TableF of the proposed method with other methods

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set
77.03	76.12	80.44	81.53	**81.66**	Colon
85.91	91.23	95.01	92.23	**95.82**	Leukemia
88.96	83.30	80.15	82.50	**89.03**	Prostate tumor
92.23	91.62	93.21	91.41	**95.00**	Lung cancer
+	+	+	+		Wilcoxon

Comparison of average execution time The proposed method with other methods :۶ Table

MOABCD	CDNC	BHAPSO	ABCD	Suggested method	Data set

110.21	80.12	96.21	89.62	**65.43**	Colon
605.39	76.41	88.41	89.15	**87.49**	Leukemia
1568.83	313.46	651.97	312.74	**281.39**	Prostate tumor
401.36	640.82	581.42	927.65	**514.73**	Lung cancer

established superior is and this order it particle an approach hopeful doer for direct object to for choice Gene in diagnosis cancer conversion does to as Kelly, Results Experiments badge they give that method choice Gene Suggestion from Considering precision classification, Efficiency classified and time run, from other Techniques

9. Conclusion

Based on the research, the hypothesis that the use of machine learning algorithm with feature selection can identify a compact subset of predictive genes, which in turn improves the accuracy of cancer classification, has been successfully proven. In this method, the application of highly random tree algorithm and whale optimization in order to select optimal features have played a central role. The analyzes have shown that the combination of the strongly random tree algorithm with the whale optimization approach has a high ability to identify and select key genes with high importance in predicting and accurately classifying cancer types. This set of compact features not only reduces the complexity of the classification model, but also helps to reduce the dimensions of the data and, as a result, increase the efficiency of the model by removing additional and unnecessary features. Also, this approach has caused the accuracy of cancer classification to improve significantly, which is a confirmation of the validity and effectiveness of the hypothesis proposed in the research.

Therefore, the results obtained from this research emphasize the importance of feature selection in machine learning processes and show the high potential of intelligent optimization algorithms in improving the accuracy and efficiency of cancer classification models. These findings open new horizons for future research in the field of optimizing classification models for cancer and other diseases using large and complex data.

References

[1] V. Kalpana, V. Vijaya Kishore, and R. Satyanarayana, "MRI and SPECT Brain Image Analysis Using Image Fusion," *Mobile Radio Communications and ٥G Networks: Proceedings of Third MRCN* ٢٠٢٢, pp. ٥٨٦-٥٧١: Springer, .٢٠٢٣

[2] S. A. Abdulrahman, W. Khalifa, M. Roushdy, and A.-B. M. Salem, "Comparative study for ٨computational intelligence algorithms for human identification," *Computer Science Review*, vol. ٣٦, pp. ١٠٠٢٣٧, .٢٠٢٠

[3] Y. Xia, S. Huang, Y. Wu, Y. Yang, S. Chen, P. Li, and J. Zhuang, "Clinical application of chromosomal microarray analysis for the diagnosis of Williams–Beuren syndrome in Chinese Han patients," *Molecular genetics & genomic medicine*, vol. ٧, no. ٢, pp. e٠٠٥١٧, .٢٠١٩

[4] V. Yuvaraj, and D. Maheswari, "Lung cancer classification based on enhanced deep learning using gene expression data," *Measurement: Sensors*, vol. ٣٠, pp. ١٠٠٩٠٢, .٢٠٢٣

[5] N. D. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. Scotto di Freca, "An experimental comparison of feature-selection and classification methods for microarray datasets," *Information*, vol. ١٠, no. ٣, pp. ١٠٩, .٢٠١٩

[6] V. Kalpana, V. Vijaya Kishore, and K. Praveena, "A common framework for the extraction of ILD patterns from CT image," *Emerging Trends in Electrical,*

*Communications, and Information Technologies: Proceedings of ICECIT-*۲۰۱۸, pp. ۵۲۰-۵۱۱: Springer, .۲۰۱۹

[7] M. Annamalai, and P. B. Muthiah, "An early prediction of tumor in heart by cardiac masses classification in echocardiogram images using robust back propagation neural network classifier," *Brazilian Archives of Biology and Technology,* vol. ۶۵, pp. e۲۲۲۱۰۳۱۶, .۲۰۲۲

[8] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Applied Soft Computing,* vol. ۶۲, pp. ۲۱۵-۲۰۳, .۲۰۱۸

[9] D. P. Berrar, W. Dubitzky ,and M. Granzow, *A practical approach to microarray data analysis*: Springer, .۲۰۰۳

[10] A. Dabba, A. Tari, and S. Meftali, "A new multi-objective binary Harris Hawks optimization for gene selection in microarray data," *Journal of Ambient Intelligence and Humanized Computing,* vol. ۱۴, no. ۴, pp. ۳۱۷۶-۳۱۵۷, .۲۰۲۳

[11] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, and M. Oussalah, "Graph-based relevancy-redundancy gene selection method for cancer diagnosis," *Computers in Biology and Medicine,* vol. ۱۴۷, pp. ۱۰۵۷۶۶, .۲۰۲۲

[12] S. Acharya, S. Saha, and N. Nikhil, "Unsupervised gene selection using biological knowledge: application in sample clustering," *BMC bioinformatics,* vol. ۱۸, pp. ۱۳-۱, .۲۰۱۷

[13] Y. Huang, and L. Zhang, "Gene selection for classifications using multiple PCA with sparsity," *Tsinghua Science and Technology,* vol. ۱۷, no. ۶, pp. ۶۶۵-۶۵۹, .۲۰۱۲

[14] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Computing and Applications,* vol. ۲۹, pp. ۱۵۵۴-۱۵۴۵, .۲۰۱۸

[15] S. Liu ,C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, and M. Dehmer, "Feature selection of gene expression data for cancer classification using double RBF-kernels," *BMC bioinformatics,* vol. ۱۹, no. ۱, pp. -۱ ۱۴, .۲۰۱۸

[16] R. Ali, A. Manikandan, and J. Xu, "A Novel framework of Adaptive fuzzy-GLCM Segmentation and Fuzzy with Capsules Network (F-CapsNet) Classification," *Neural Computing and Applications,* pp. ۱۷-۱, .۲۰۲۳

[17] N. Almugren, and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE access,* vol. ۷, pp. ۷۸۵۴۸-۷۸۵۳۳, .۲۰۱۹

[18] H. Almazrua, and H. Alshamlan, "A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data," *IEEE Access,* .۲۰۲۲

[19] M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu, and M. O. Agyeman, "A survey of machine learning approaches applied to gene expression analysis for cancer prediction," *IEEE Access,* vol. ۱۰, pp. ۲۷۵۳۴-۲۷۵۲۲, .۲۰۲۲

[20] Y. Liang, S. Zhang, H. Qiao, and Y. Yao, "iPromoter-ET: Identifying promoters and their strength by extremely randomized trees-based feature selection," *Analytical Biochemistry,* vol. ۶۳۰, pp. ۱۱۴۳۳۵, .۲۰۲۱

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning,* vol. ۶۳, pp. ۴۲-۳, .۲۰۰۶