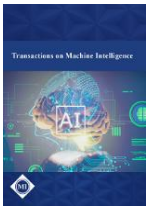
Transactions on Machine Intelligence 6(2) (2023) 66-75

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI



Investigation and Analysis of Gene Expression Using the Fusion Method of Feature Selection and Dynamic Neural Network Classification

F. Turki[1], A. Khadem[1], A.H. Jalali Aghchay[2,*]

[1] Department of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran
[2] Assistant Professor, Department of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran.

ARTICLE INFO

Article History:
Received 10 March 2023
Received in revised form 14 April 2023
Accepted 29 May 2023
Available online 8 June 2023

Keywords:
Machine Learning, Neural Network, Data Classification, Result Prediction, Biomechanics

ABSTRACT

The analysis of high-volume microarray data faces challenges such as limited sample size, computational complexity, and the risk of inappropriate gene selection. The scarcity of samples hampers computational analysis and classification complexity, while reducing the classification's ability to generalize and predict new samples. Moreover, datasets with a high gene-to-sample ratio raise concerns about the selection of relevant genes for accurate predictive models. Interpreting disease-causing genes becomes intricate as only a subset of genes offers a precise biological insight into the disease. To address these issues, a focus on a smaller set of gene expression data is crucial for a more effective understanding of informative genes. Hence, the primary objective in microarray data analysis is to significantly reduce the number of genes through discriminative gene selection, enhancing the precision of information contained in the data. This article conducts gene expression classification on various cancer types, including colon cancer, breast cancer, leukemia, prostate tumors, and DLBCL. Each cancer type is independently evaluated in the feature selection cycle and classified using varying numbers of features. This approach aims to overcome challenges in microarray data analysis and improve the accuracy and interpretability of gene expression classification.

1. INTRODUCTION

In contemporary times, extensive research efforts have been dedicated to the domain of Feature Selection and Classification. The ensuing examples serve to illustrate notable contributions in this area.

Zhang, H. (2021) presents a novel feature selection algorithm employing approximate conditional entropy based on fuzzy information granule. This approach significantly reduces gene dataset dimensions, leading to improved classification accuracy. Link to the article [1]. Rezaee et al. (2022) proposes a combination strategy for gene expression in various diseases. The study employs soft ensembling to identify effective genes and utilizes a novel deep neural network for classification, achieving high accuracy. Link to the article [2]. Gakii and Rimiru (2021) explore feature selection methods, including differential gene expression analysis, co-expression networks, and

* Corresponding Author: jalali@kntu.ac.ir
Assistant Professor, Department of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran

http://dx.doi.org/10.47176/TMI.2023.66

association rule mining. This study aims to identify common cancer-related genes and their potential shared biological functions. Link to the article [3]. Seryasat and Haddadnia (2018) evaluate a new ensemble learning framework for mass classification in mammograms. The research contributes to the advancement of mass classification techniques in breast cancer diagnosis [4].

The central goal of data classification is to objectively assign an input vector to a specific class. Binary classification, which involves distinguishing between two classes, is the simplest form, while multi-class classification, assigning inputs to more than two classes, represents a more complex scenario. Recent research underscores the efficacy of microarrays in cancer diagnosis [5].

1.1. overview of classifiers

Microarray data exhibit distinctive characteristics that distinguish them from other automatic classification methods, characterized by high dimensions and a limited number of samples. These traits can pose challenges in various machine learning methods, including overfitting and high dimensionality in classification. To tackle these issues, numerous methods have been proposed in the research literature, often relying on simple linear or non-linear models.

1.1.1. k-Nearest Neighbor (K-nn)

The mentioned algorithm falls under the category of supervised machine learning algorithms, applicable to both classification and predictive regression problems. However, it is commonly used in predictive classification challenges within the industry. The K-nearest neighbors (K-NN) algorithm is straightforward, lacking a specific learning stage, and utilizes all available data sets during classification. Notably, K-NN is a parameter-free learning algorithm, as it doesn't rely on assumptions about the initial data. To apply the algorithm, a dataset is required. For the initial stage of K-NN, the test data should be loaded alongside the training data. Subsequently, the value of "k," representing the closest data points, needs to be selected, and this integer can vary. Distance between each line of the training data and the test data is determined using one of the Euclidean methods, hamming distance, or Manhattan distance. The Euclidean method is typically employed for distance calculation. The lines are sorted in ascending order based on the distance value, and the algorithm assigns a class to the test point using the most common class of these lines. Figure 1 illustrates an instance of the K-NN algorithm with k values of 3 and 5.

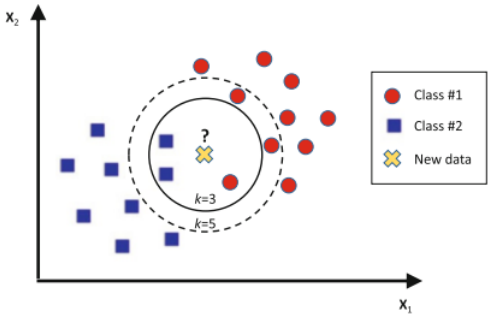
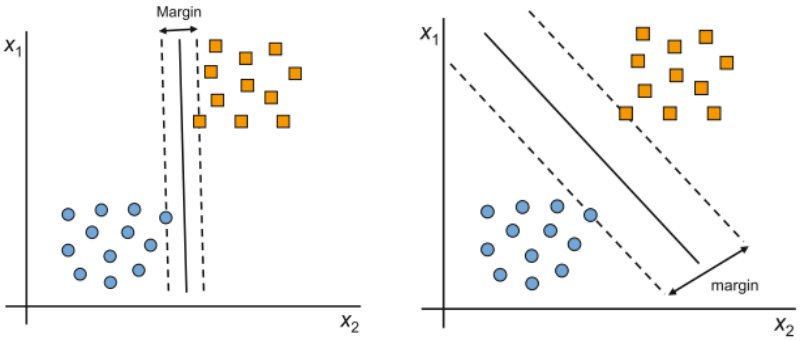


Fig. 1. An example for K-nn algorithm for 3 and 5 k values

The most important thing is to get the optimal value for k [6].

1.1.2. Support Vector Machines (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that segregates data samples depicted as points in space by employing a line or hyperplane. This segregation ensures that data points on the same side of the line are considered similar and grouped together. When new data samples are introduced into the same space, they are assigned to one of the existing categories based on their placement. SVM is effective in creating a clear boundary between different classes in the dataset, making it a valuable tool for classification tasks.



a b

Fig. 2. An example for SVM (a) with a small margin (b) with a large margin

Figure (b) has a larger margin than Figure (a). From a scientific point of view, the maximum margin of hyperplane copes with the problem of overfitting and has a good generalization capacity [7].

1.1.3. Decision Trees and random forests

The decision tree algorithm stands out as a widely utilized data mining technique, serving as a predictive model applicable to both regression and class models. When employed for classification purposes, it is denoted as a classification tree, while in the context of regression tasks, it is recognized as a regression decision tree [8-10]. In assessing the effectiveness of a classifier, precise measurement of its accuracy is crucial. It is recommended to gauge the accuracy of a classifier using dedicated test data. After constructing the model with training data, evaluating its performance on test data allows for an accurate assessment of its ability to assign class labels to samples.

1.1.4. Bayes theorem

The Naive Bayes algorithm emerges as a straightforward and effective classification technique in scenarios where classification challenges are encountered. Particularly suitable when dealing with situations featuring a limited number of variables but abundant observations, the straightforward Bayes algorithm is adept at discerning classifications. Fundamentally, the Bayes classification algorithm is rooted in Bayes' theorem. Leveraging this algorithm involves applying Bayes' theory, expressed by the following formula:

$$p(c / x) = \frac{p(x / c)p(c)}{p(x)} \tag{1}$$

In equation (1), $p(c/x)$ is posterior probability, $p(x/c)$ is likelihood, $p(c)$ is class prior probability and $p(x)$ is predictor prior probability.

2. EXPERIMENT PROCEDURE

The experimental datasets selected for analysis include diffuse large B-cell lymphomas (DLBCL) [11], leukemia [12], prostate cancer [13], colon cancer [14], and breast cancer [15]. The DLBCL dataset consists of 7070 genes from 77 samples, with 58 samples corresponding to DLBCL and the remaining to follicular lymphoma (FL). To distinguish between these lymphoma types, classification models are developed using gene expression data. The Leukemia dataset, derived from bone marrow and blood culture media, encompasses 72 samples (47 ALL and 25 AML) and measures gene expression across 7,129 genes. The third dataset involves 102 samples (50 normal and 52 prostate tissue) with 12,533 genes. For clarity, technical abbreviations will be elucidated upon their initial usage. All gene data have undergone quantile normalization.

Furthermore, the dataset for intestinal cancer comprises 2000 descriptive genes from 62 tissues (62 samples). Among these, 22 samples (35.5% of the total data) represent a healthy state, while 40 samples (64.5% of the total

data) represent a cancerous state. Lastly, the breast cancer dataset includes 132 tissue samples and 1926 expressed genes. Of these, 11 samples (8.3% of the total data) are healthy, and 122 samples (92% of the total data) are from cancerous tissue. Table 1 provides a succinct summary of the sample data, including the number of features, samples, and relevant classes.

Table 1. Sample data along with indicators such as number of features, number of samples and number of related classes

No. Dataset	Name of gene data	No. of samples	No. of features	No. of class	Missing values
1	DLBCL	77	7070	2	No
2	Prostate cancer	102	12533	2	No
3	Leukemia cancer	72	7129	2	No
4	Colon cancer	62	2000	2	No
5	Breast cancer	132	1926	2	No

2.1. The perspective of the method

Feature selection and classification are crucial machine learning techniques that contribute to a nuanced understanding and accurate discrimination of data. In this study, we focused on gene expression classification for five different datasets—colon cancer, breast cancer, leukemia, prostate tumors, and DLBCL. The feature selection process involved analyzing each dataset individually through various sets of features. For classification, we employed four types of neural networks: dynamic neural network, support vector machine network, Bayes theory, K nearest neighbor, and decision tree.

The feature selection was executed using majority voting, considering Relief, PCC, F-Score, and Term Variance methods. Our findings emphasize that the integration of features significantly impacts accuracy, highlighting the effectiveness of a fusion method in enhancing classification results.

2.1.1. Validation Method

The K-fold method is a robust validation model where the dataset is partitioned into K distinct parts. The modeling process iterates K times, employing each K-1 part for training and the remaining part for testing and validating the predictive model. This method ensures thorough validation by repeatedly cycling through different subsets of data. Technical term abbreviations are clarified upon their initial use, and the text strictly adheres to principles of objectivity, logical structure, conventional language, clarity, formal register, and grammatical correctness. The average prediction error is computed over each of the K steps, providing a reliable measure of the model's performance. This approach, applied to both feature selection and classification stages, ensures robustness by using random subsets of data and mitigates the impact of data distribution on the modeling process.

3. RESULTS

The neural network model emerges as the most suitable dynamic model for adapting to diverse data types, given its capacity for efficient adjustments through weight tuning. It's important to note that precise regulatory parameters aren't assumed to be universally effective in this model, except for factors like the number of neurons per layer and the number of layers, which vary based on the specific problem.

To harness the potential of the neural network model, a deliberate repetition process is employed for finding the optimal classification model with the highest accuracy. This process is purposeful and not based on random pattern discovery. Initially, fitting is carried out with a small number of layers and neurons per layer. Subsequently, both the number of neurons and layers are incrementally added. Table 2 showcases the configurations with a k-fold value of 5 in the classification process and quadruple neural networks.

Table 2. The settings with k fold equal to 5 in classification and quadruple neural networks

25	**21**	**14**	**12**	**7**	**5**	**4**	**3**	**2**	**No. of features**	
Yes	Yes	Yes	No	No	Yes	No	Yes	No	**Success compared to other classifiers**	
96%	95%	97.6%	96%	96%	97.3%	98%	96%	93%	Classification accuracy in 10 repetitions of networks	**CV=5**
95%	94%	99%	98%	98%	96%	97%	96%	95%	Classification accuracy in 20 repetitions of networks	
95%	94%	96%	95%	97%	98%	96%	95%	94%	Classification accuracy in 10 repetitions of networks	**CV=10**
95%	97%	98%	98%	98%	96%	97%	96%	94%	Classification accuracy in 20 repetitions of networks	
8220 4823 8545 7451 2718 8290 9138 6640 5460 5014 7531 5227 120 9949 6620 7574 6168 3997 6569 3200 5047 7229 7652 5815 8009	8220 4823 7372 7652 398 938 1514 3673 3774 4011 4447 4700 4986 5237 5648 7531 2714 3200 7346 7451	4823 7451 8765 8468 7515 2718 6821 7531 10130 5815 120 7652 9949	8220 4823 275 3059 3077 3699 4447 7061 7451 4700 5461 6144	3699 3077 6144 8220 10324 8536 7061	7451 4823 8765 3200 7515	4823 7451 7229 8545	3200 4447 4823	3200 4823	Selected features	

The study focused on gene sets associated with prostate disease and underwent seven repetitions. The results demonstrate the superiority of the dynamic neural network-based learning approach over other methods in gene classification, showcasing the effectiveness of the gene selector model in identifying the best subset of selected features. This superiority is evident not only in the substantial gap between the maximum and minimum points on their respective graphs but also in their lower average output accuracy. Unlike other methods that exhibit wider and larger ranges of feature selection changes, the proposed method covers a more confined area, as indicated by the smaller size of its box plots compared to those of other methods.

In essence, the proposed feature selection method yields higher accuracy levels and proves more suitable than other methods. This conclusion is supported by the consistently larger height of the box in the proposed method for feature selection and classification compared to other methods, emphasizing the range between quartiles or the box height. Additionally, the proposed method demonstrates fewer outlier values (Outlier AUC) in classification compared to other methods. While outliers are detected for each average accuracy level of the methods, it is observed that the dynamic neural network method has relatively fewer outliers than other methods, further reinforcing its efficacy for achieving optimal classification. Figure 3 illustrates the outcomes of applying three features and the performance of various classification models. From left to right, the models used include Bayes classifiers, SVM, decision tree, KNN, and dynamic neural network with a weighting method, all with 10-fold cross-validation in the first iteration.

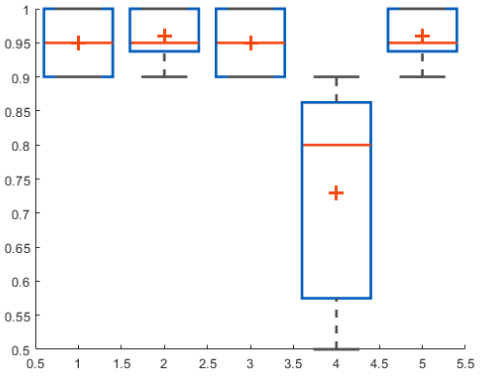


Fig. 3. The result of applying 3 features and performance of different classifications. From left to right, respectively, Bayes classifiers, SVM, decision tree, KNN and dynamic neural network by weighting method with CV=10 and first iteration

Figure 4. shows the result of applying 4 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, KNN and dynamic neural network by weighting method with CV=10 and second iteration.

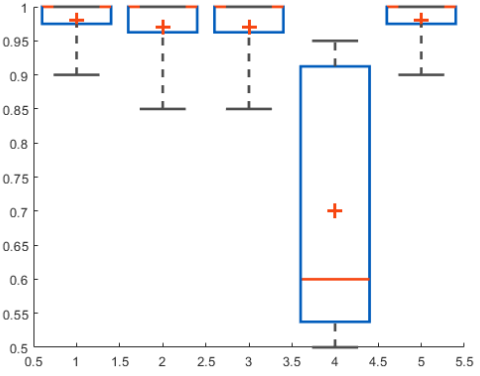


Fig. 4. The result of applying 4 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, KNN and dynamic neural network by weighting method with CV=10 and second iteration

Figure 5 shows the result of applying 5 features and performance of different classifications, which are, from left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and third iteration, respectively.

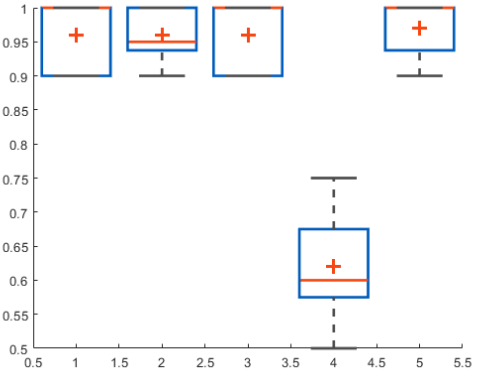


Fig. 5. The result of applying 5 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and third iteration

3.1. Intrusion Detection using bagging algorithm

Figure 6 is the result of applying 7 features and performance of different classifiers, from left to right, respectively, Bayesian classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and fourth iteration.

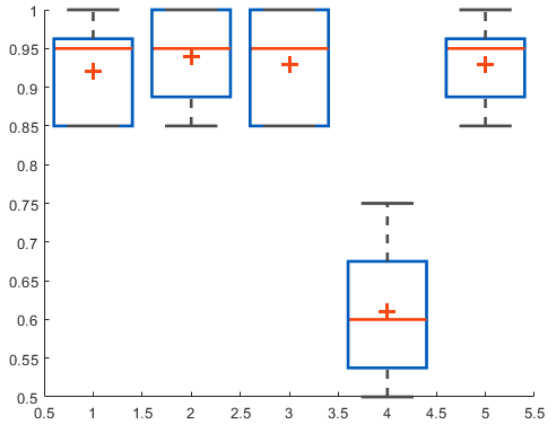


Fig. 6. The result of applying 7 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and fourth iteration

Figure 7 shows the result of applying 12 features and performance of different classifications from left to right, respectively, Bayesian classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and repetition the fifth.

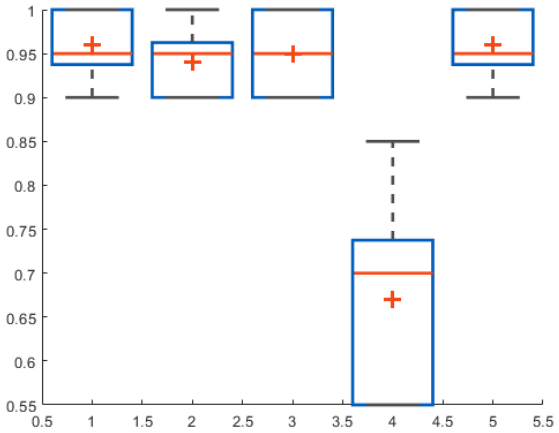


Fig. 7. The result of applying 12 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and fifth iteration

Figure 8 shows the result of applying 15 features and functions of different classifications. From left to right, respectively, Bayes classifiers, SVM, decision tree, KNN and dynamic neural network by weighting method with CV=10 and sixth iteration

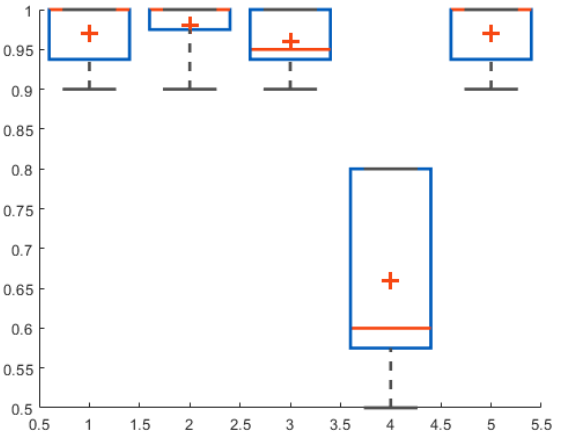


Fig. 8. The result of applying 15 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and sixth iteration

Figure 9 shows the result of applying 25 features and functions of different classifications. From left to right, respectively, Bayes classifiers, SVM, decision tree, KNN and dynamic neural network by weighting method with CV=10 and the seventh iteration.

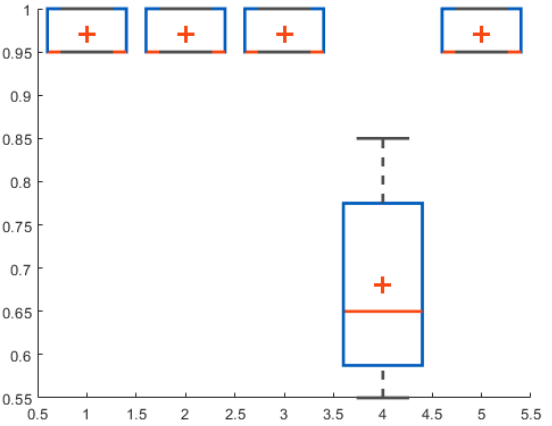


Fig. 9. The result of applying 25 features and performance of different classifications. From left to right, Bayes classifiers, SVM, decision tree, CNN and dynamic neural network by weighting method with CV=10 and seventh iteration

It can be seen in figure 10 that the NN method has the best accuracy and compliance. Finally, the description of the code written with MATLAB software is explained.

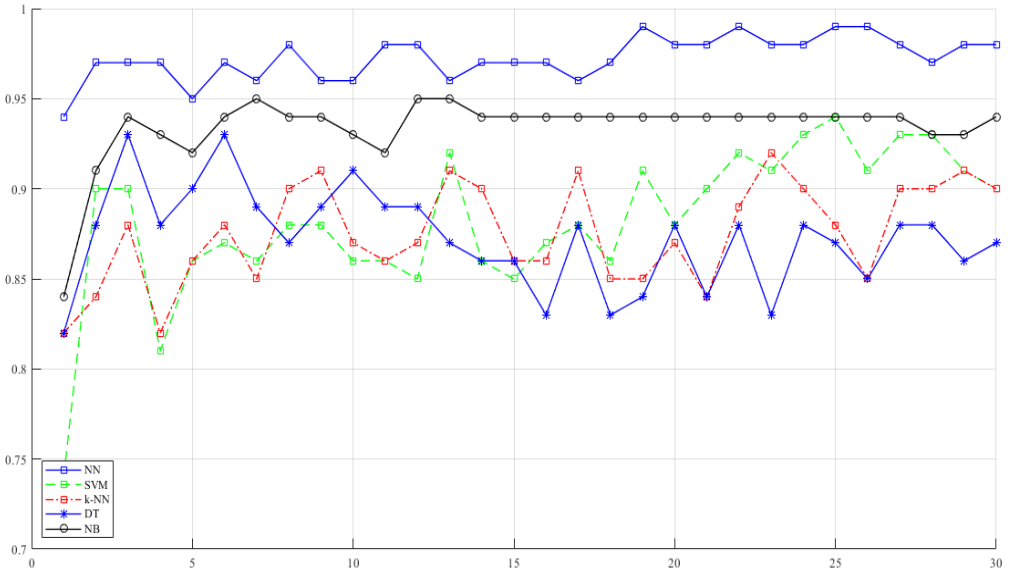


Fig. 10. Comparison between classification solutions

4. CONCLUSION

In this article:

Gene expression classification was conducted for five types of data, namely colon cancer, breast cancer, leukemia, prostate tumors, and DLBCL. Each data type was individually subjected to the feature selection cycle and classification with a variable number of features.

The classification step involved utilizing four categories of neural networks, including dynamic neural network, support vector machine network, Bayes theory, K nearest neighbor, and decision tree.

Feature selection was performed through majority voting using Relief, PCC, F-Score, and Term Variance methods.

For validation, the k-fold model, involving the division of data into k separate parts, was employed, with the NN (Neural Network) method.

The NN method demonstrated the highest accuracy and performance among the classification methods for all four types of data.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Zhang, H. (2021). Feature Selection Using Approximate Conditional Entropy Based on Fuzzy Information Granule for Gene Expression Data Classification. Frontiers in Genetics, 12, https://doi.org/10.3389/fgene.2021.631505.

[2] Rezaee, K., Jeon, G., Khosravi, M., Attar, H., & Sabzevari, A. (2022). Deep learning-based microarray cancer classification and ensemble gene selection approach. IET Systems Biology, 16, 120 - 131. https://doi.org/10.1049/syb2.12044.

[3] Gakii, C., & Rimiru, R. (2021). Identification of cancer related genes using feature selection and association rule mining. Informatics in Medicine Unlocked, 24, 100595. https://doi.org/10.1016/J.IMU.2021.100595.

[4] Seryasat, O. R., & Haddadnia, J. (2018). Evaluation of a new ensemble learning framework for mass classification in mammograms. Clinical breast cancer, 18(3), e407-e420. https://doi.org/10.1016/j.clbc.2017.05.009

[5] Gordon, A. D. (1999). Classification. CRC Press. https://doi.org/10.1201/9780367805302

[6] Boguslawski, L. (2004). Influence of pressure fluctuations distribution on local heat transfer on flat surface impinged by turbulent free jet. Proceedings of the ASME - ZSIS International Thermal Science Seminar II, Bled, Slovenia. https://doi.org/10.1615/ICHMT.2004.IntThermSciSemin.230

[7] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. Analyst, 135(2), 230-267. https://doi.org/10.1039/B918972F

[8] Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo.

[9] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1993). Classification and regression trees, wadsworth international group, belmont, ca, 1984. Case Description Feature Subset Correct Missed FA Misclass, 1, 1-3.

[10] Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18(60), 1-8.

[11] Monti, S., Savage, K. J., Kutok, J. L., Feuerhake, F., Kurtin, P., Mihm, M., & Shipp, M. A. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood, 105(5), 1851-1861. https://doi.org/10.1182/blood-2004-07-2947

[12] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537. https://doi.org/10.1126/science.286.5439.531

[13] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2), 203-209. https://doi.org/10.1016/S1535-6108(02)00030-2

[14] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 96(12), 6745-6750. https://doi.org/10.1073/pnas.96.12.6745

[15] Matamala, N., Vargas, M. T., González-Cámpora, R., Miñambres, R., Arias, J. I., Menéndez, P., Benítez, J. (2015). Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. Clinical Chemistry, 61(8), 1098-1106. https://doi.org/10.1373/clinchem.2015.238691