

A hybrid inductive model for gene expression data processing using spectral clustering

Sergii Babichev^{1,2,*†}, Oleg Yarema^{3,†} and Ihor Liakh^{4,†}

¹ Kherson State University, 27, University street, 73000, Kherson, Ukraine

² Jan Evangelista Purkyně University in Ústí nad Labem, Pasteurova, 15, 400 96, Ústí nad Labem, Czech Republic

³ Ivan Franko National University, 1, Universytetska str. 79000, Lviv, Ukraine

⁴ Uzhhorod National University, 14, University street, 88000, Uzhhorod, Ukraine

Abstract

One of the key directions in modern bioinformatics is the development of systems for diagnosing various diseases using gene expression data. Clustering gene expression profiles is a critical step in disease diagnosis systems. In this study, we propose a hybrid inductive model for clustering gene expression profiles using the spectral clustering algorithm. The implementation of this model aims to reduce reproducibility errors by serializing the data processing flow and optimizing clustering based on both internal and external quality criteria. The model is presented as a block diagram, and its practical implementation has demonstrated the high effectiveness of the proposed approach. The model's performance was evaluated using a convolutional neural network. The experimental dataset consisted of gene expression values assigned to the identified clusters. The simulation results indicate that the highest classification accuracy was achieved with a three-cluster structure, which corresponded to the highest balance between internal and external clustering quality criteria. These findings create opportunities for enhancing existing gene expression clustering models through more precise tuning of clustering algorithm hyperparameters, guided by the principles of inductive methods for analyzing complex systems.

Keywords

Gene expression data, spectral clustering, internal and external clustering quality criteria, convolution neural network (CNN), classification accuracy

1. Introduction

Gene expression (GE) data are a crucial element of modern research in bioinformatics and genomics. They enable the investigation of gene functional activity under various conditions and developmental stages while also aiding in the discovery of molecular mechanisms underlying biological processes. This, in turn, provides a foundation for developing and refining personalized medicine systems through accurate analysis and processing of GE data in diagnostic models, reconstruction, simulation, and validation of gene regulatory network (GRN) models [1]. As demonstrated by the analysis of contemporary GE data [2], the human genome consists of tens of thousands of genes, with around 25,000 of them active. The activity (expression) of these genes is governed by various processes that dictate an organism's functioning. Thus, identifying the subset of genes that directly determine the state of the organism remains one of the pressing challenges in bioinformatics, and as of now, it does not have a definitive solution.


A significant number of scientific studies are currently focusing on processing GE data to identify co-expressed genes through cluster analysis [3-6]. These studies aim to refine clustering techniques to more accurately group genes with similar expression patterns, which can reveal functional relationships and regulatory mechanisms within the genome. The results of such

IDDm'24: 7th International Conference on Informatics & Data-Driven Medicine, November 14 - 16, 2024, Birmingham, UK

* Corresponding author.

† These authors contributed equally.

✉ sergii.babichev@ujep.cz (S. Babichev); oleg.yarema@lnu.edu.ua (O. Yarema); ihor.lyah@uzhnu.edu.ua (I. Liakh)

 0000-0001-6797-1467 (S. Babichev); 0000-0003-3736-4820 (O. Yarema); 0000-0001-5417-9403 (I. Liakh)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

analyses are essential for advancing our understanding of gene networks and improving predictive models for various biological conditions. Thus, in [3], the authors focus on improving the process of identifying subsets of co-expressed genes by leveraging advanced cluster analysis techniques. The proposed approach enhances the quality of GE data imputation by exploiting multiple clustering solutions, enabling more accurate grouping of genes with similar expression patterns. This method significantly contributes to the allocation of gene subsets with shared functional activity, offering a robust tool for bioinformatics research. Study [4] introduces a Cluster Decomposition-based Anomaly Detection method, known as scCAD, to improve the identification of co-expressed genes in single-cell GE data. By iteratively refining clusters based on differential signals, scCAD enhances the detection of rare cell types that are often missed by traditional clustering methods. Benchmarking on 25 datasets shows scCAD's superiority in identifying rare cell types and disease-related immune subtypes, providing valuable insights into complex biological processes. In [5], the authors emphasize the importance of clustering in optimizing the analysis of single-cell chromatin accessibility (scATAC-seq) and multi-omic datasets. They benchmark eight feature engineering pipelines across various data processing stages, assessing their ability to discover and differentiate cell types based on clustering performance. SnapATAC and SnapATAC2 are highlighted as the most effective methods for datasets with complex cell-type structures, proving critical in extracting meaningful insights from high-dimensional and noisy data. Study [6] discusses the challenges of developing effective clustering algorithms for spatial transcriptomics (ST) data, focusing on defining spatially coherent regions within tissue slices and integrating multiple slices from different sources. The authors systematically benchmark a range of state-of-the-art clustering, alignment, and integration methods using diverse datasets, evaluating their performance with eight metrics related to spatial accuracy and contiguity. Based on these results, the study provides detailed recommendations for selecting the most suitable methods for specific datasets and offers guidance for future method development in ST data analysis.

However, it should be noted that the successful application of cluster analyze the data to identify and form subsets of co-expressed GE profiles for disease diagnosis systems faces several limitations and unresolved challenges. Despite significant progress in refining clustering techniques to better group genes with similar expression patterns, certain obstacles persist. A key limitation is the difficulty in identifying rare or subtle gene expressions, especially when data is noisy or high-dimensional, such as in scRNA-seq or spatial transcriptomics. While methods like scCAD and SnapATAC2 have advanced in this area, they still rely on iterative refinement and sophisticated benchmarks and may overlook rare gene sets or struggle with large, complex datasets.

Another unsolved issue is the challenge of integrating multiple tissue samples or datasets, particularly in spatial transcriptomics and multi-omics studies, where spatial coherence and alignment are critical but difficult to achieve. Existing methods often lack scalability or struggle with generalizing across diverse data sources. Furthermore, many studies highlight the lack of comprehensive benchmarks, limiting the ability to systematically compare and improve clustering algorithms.

In sum, while current research has made strides in improving gene expression clustering, developing more robust, scalable, and generalizable methods remains a pressing need to ensure the accurate formation of co-expressed gene subsets for reliable disease diagnosis based on GE data.

The performance of the spectral clustering algorithm for GE data clustering has shown promise due to its ability to effectively handle complex, non-linear relationships within high-dimensional datasets [7,8]. In this study, we continue the research presented in [9,10] and propose a hybrid inductive model that utilizes spectral clustering to form subsets of co-expressed genes, enhancing the ability to detect subtle patterns in gene expression profiles. Spectral clustering operates by transforming data into a lower-dimensional space, where traditional clustering techniques can be applied more efficiently, thus overcoming limitations of other algorithms that may struggle with high-dimensionality and noise inherent in gene expression data.

The hybrid approach combines spectral clustering with inductive methods of complex system analysis to further improve accuracy in grouping co-expressed genes, leveraging the algorithm's

strength in identifying clusters of varying shapes and sizes. By applying spectral clustering to gene expression data, we achieve better delineation of gene subsets that are often difficult to separate using standard techniques. This model has the potential to significantly enhance disease diagnosis systems by improving the precision and scalability of clustering in complex biological datasets.

2. Materials and Methods

Spectral clustering (SC) is a modern technique that helps identify clusters with arbitrary shapes by leveraging similarity matrices between the studied objects [11-13]. Compared to traditional clustering methods, such as k-means, hierarchical agglomerative, and divisive approaches, spectral clustering provides several significant advantages. It often delivers superior results in terms of clustering quality and is also relatively easy to implement, utilizing standard linear algebra operations efficiently. Unlike many conventional algorithms, spectral clustering does not rely on the absolute positions of objects in space; instead, it focuses on analyzing the affinities between them, which makes it especially effective for grouping complex structures. The typical implementation of spectral clustering follows a sequence of key steps:

1. **Constructing the Similarity Graph.** A similarity graph $G = (V, E)$ is an undirected graph comprising a set of nodes $V = \{v_1, v_n\}$ (the objects being studied) and a set of edges $E = s_{ij}$, which connect nodes i and j and define the measure of proximity between them. Two nodes are considered connected if the similarity value s_{ij} between the corresponding objects (nodes of the graph) exceeds a certain threshold, and the edge is assigned a weight w_{ij} . In this scenario, the clustering task can be formalized as follows: the graph structure should be constructed so that the edges between different groups (clusters) have very low weights, indicating that objects in different clusters are as dissimilar as possible. Conversely, edges between nodes within the same group should have high weights, signifying that objects within the same cluster are as similar as possible. Constructing the similarity graph involves calculating a similarity matrix using an appropriate proximity metric based on the characteristics of the objects being studied. For instance, when clustering gene expression profiles, a hybrid modified metric based on maximizing mutual information and Pearson correlation is used. In conclusion, the similarity graph is an undirected weighted graph where the strength of connection between nodes is determined by the weight of the edge connecting them. The degree of a node is defined as the sum of the weights of the edges connecting this node to its neighbors:

$$d_i = \sum_{j=1}^m w_{ij} \quad (1)$$

where m is the number of nodes directly connected to node i . Note that if two nodes are not directly connected, the weight of the edge between them is zero. Based on node degrees, the degree matrix D is formed, which is a diagonal matrix with the degrees of the nodes d_1, d_m on the main diagonal. This process creates the conditions for cluster formation by initializing a threshold coefficient that limits the number of connections with non-zero weights. All components in a subset of objects A are considered connected if the weights of direct or indirect connections between all nodes in A are greater than zero, and the weights between nodes in A and those in other subsets are zero. Depending on how the set of objects and the corresponding similarity matrix are transformed into a similarity graph, the following types of graphs can be identified:

- ϵ -neighborhood graph: This type connects all points (object identifiers) whose pairwise distances are smaller than a predefined ϵ -neighborhood. Since distances between all pairs

are measured on the same scale (no larger than ϵ), the graph is typically unweighted and does not require additional information regarding the strength (weight) of the connections.

- k-nearest neighbors graph: In this type of graph, node i is connected to node j if j is one of the k-nearest neighbors of i . The weight of the edges is initialized based on the similarity matrix, making this a weighted graph.
 - Fully connected graph: This type connects all nodes with positive connection strengths based on the similarity matrix. Local ϵ -neighborhoods are formed using appropriate similarity functions, such as a Gaussian similarity function.
2. **Constructing the Laplacian Matrix and Computing Eigenvectors.** The Laplacian matrix, derived from the graph's Laplacian, is a central component in spectral clustering. For this process, we assume that the graph G is undirected and weighted, with its weight matrix denoted as W . The eigenvectors of the similarity matrix can be either normalized or unnormalized. The eigenvalues of W are sorted in ascending order, and the first k eigenvectors correspond to the smallest k eigenvalues. The Laplacian matrix can be computed using either normalized or unnormalized values.
 3. **Cluster Formation Using the k-Nearest Neighbors Method.** In this method, the clustering structure is determined by applying the k-nearest neighbors algorithm. The algorithm assigns each node to a cluster based on its proximity to the nearest neighbors.

2.1. Step-by-step procedure for implementing the SC algorithm

Assume that the experimental data consists of n objects (points in an m -dimensional space), where the distances between all pairs of points are defined by a similarity matrix. Depending on the method used to construct the similarity graph and compute the Laplacian matrix, several step-by-step procedures form the basis of the SC algorithm.

1. *SC Algorithm based on the unnormalized Laplacian matrix.*

Input: Similarity matrix $W \in R^{n \times n}$, number of clusters k .

Steps:

- Build the similarity graph using the values of the similarity matrix W to initialize the weights of the corresponding edges.
- Calculate the unnormalized Laplacian matrix L .
- Calculate the first k eigenvectors of L : u_1, u_k . Form matrix $U \in R^{n \times k}$, where each column represents an eigenvector u_1, u_k .
- For each $i = 1, n$, extract vector $y_i \in R^k$, corresponding to the i -th row of matrix U .
- Clustering the points, corresponding to the vectors $y_i \in R^k$, using the k-means algorithm to form clusters C_1, C_k .

Output: Clusters A_1, A_k , where $A_i = \{j/y_j \in C_i\}$ contains the points in the i -th cluster.

2. *SC Algorithm based on the normalized Laplacian using the Shi and Malik method.*

Input: Similarity matrix $W \in R^{n \times n}$, number of clusters k .

Steps:

- Build the similarity graph using the values of the similarity matrix W to initialize the edge weights.
- Calculate the normalized Laplacian matrix L .

- Calculate the first k eigenvectors of L , corresponding to the equation $Lu = \lambda Du$, where λ is the eigenvalue corresponding to eigenvector u . Form matrix $U \in R^{n \times k}$, where each column represents an eigenvector u_1, u_k .
- For each $i = 1, n$, extract vector $y_i \in R^k$, corresponding to the i -th row of matrix U .
- Clustering the points, corresponding to the vectors $y_i \in R^k$, using the k -means algorithm to form clusters C_1, C_k .

Output: Clusters A_1, A_k , where $A_i = \{j/y_j \in C_i\}$ contains the points in the i -th cluster.

3. *SC Algorithm based on the normalized Laplacian using the Ng, Jordan, and Weiss method.*

Input: Similarity matrix $W \in R^{n \times n}$, number of clusters k .

Steps:

- Build the similarity graph initializing edge weights with the values from matrix W .
- Calculate the normalized Laplacian matrix L_{sym} .
- Calculate the first k eigenvectors of L_{sym} : u_1, u_k . Form matrix $U \in R^{n \times k}$, where each column represents an eigenvector u_1, u_k .
- Normalize the rows of matrix U to form matrix $T \in R^{n \times k}$, according to the equation:

$$t_{ij} = \frac{u_{ij}}{\sqrt{\sum_{j=1}^k u_{ij}^2}} \quad (2)$$

- For each $i = 1, n$, extract vector $y_i \in R^k$, corresponding to the i -th row of matrix T .
- Clustering the points, which is associated with the vectors $y_i \in R^k$, using the k -means algorithm to form clusters C_1, C_k .

Output: Clusters A_1, A_k , where $A_i = \{j/y_j \in C_i\}$ contains the points in the i -th cluster.

It is important to note that, in all cases, the results of the algorithm depend on the method used to construct the similarity matrix (i.e., how object proximity is measured) and the desired number of clusters. However, in many instances, the number of clusters cannot be predetermined, making it necessary to apply various clustering methods alongside quantitative criteria to evaluate clustering quality. The choice of proximity metric depends on the type of data. For the gene expression profiles analyzed in the simulation, a modified hybrid metric is used, combining a mutual information maximization criterion with Pearson's consistency criterion [14]. The number of clusters is determined using methods based on an objective inductive clustering technique

2.2. Hybrid inductive model for clustering GE profiles using the SC algorithm

The practical implementation of the step-by-step procedure for GE profiles clustering using the SC algorithm comprises the following phases:

Stage I. Dataset Preparation, Model Initialization

- 1.1. Form the GE matrix X , where $X \in R^{n \times m}$. Here, m and n are the amount of genes and samples, respectively.

- 1.2. Construct a measure to assess the similarity of GE profiles.
- 1.3. Develop functions to calculate various type of criteria (internal, external, balance) for evaluating the quality of GE profiles clustering.
- 1.4. Split the GE profiles into two comparable groups A and B .
- 1.5. Calculate the distance matrices for the GE profiles allocated in the comparable groups.
- 1.6. Set the range of possible clusters quantity, k_{min} and k_{max} .

Stage II. Clustering of GE Data and Quality Evaluation

- 2.1. Initialize the number of clusters $k = k_{min}$.
- 2.2. Perform grouping of GE data in the subsets A and B .
- 2.3. Calculate the internal and corresponding external quality criteria.
- 2.4. When k is less than k_{max} , increment the cluster count by one and repeat step 2.2. If not, proceed to compute the balance criterion using the internal and external metrics obtained.
- 2.5. Evaluate the results and identify the optimal clustering that maximizes the balance criterion.

Stage III. GE Data Classification

- 3.1. Create subsets of GE data from the identified clusters to be used as input for a convolutional neural network (CNN).
- 3.2. Utilize a CNN on the GE data allocated within formed clusters, assess classification performance metrics.
- 3.3. Evaluate the findings and generate subsets of co-expressed GE profiles.

3. Simulation, Results and Discussion

The modeling was executed using GE data from the GSE19188 dataset [15], which involved patients undergoing lung cancer research. The data, obtained from the Gene Expression Omnibus (GEO) [16], includes DNA analysis results from 156 patients using DNA microarray technology. Of these, 65 were determined to be healthy, whereas 91 were diagnosed with cancer. After filtering out low-expressed genes, the dataset matrix was reduced to a size of (156×10,000). Based on previous research [14], we used the WB-index [17] and the PBM criterion [18] as internal clustering quality metrics. In this case, the most effective clustering occurs when the WB-index is minimized and the PBM-index is maximized. The external quality index was determined by the normalized difference of the respective internal measures, computed on subsets A and B . The balance criterion was accessed using Harrington method in accordance to technique, described in detail in [9,10]. Figure 2 depicts the simulation results. The modeling process involved varying the number of clusters between 2 and 10.

As observed, the internal and external measures of clustering performance can sometimes conflict with each other, highlighting the importance of calculating the balance measure, which incorporates both internal and corresponding external metrics. Its maximum value is achieved when the gene expression profiles are grouped into three clusters (Figure 2d). The internal WB-index indicates that the best clustering solution involves three clusters for subset A and two for subset B (Figure 2a). When using the internal PBM index, the optimal clustering for both subsets aligns with a three-cluster structure (Figure 2b). For the external metrics, the most effective clustering is a three-cluster configuration when applying the WB-index and a four-cluster structure when using the PBM criterion (Figure 2c).

The next step in implementing the algorithm, whose structural flowchart is shown in Figure 1, involves applying a CNN to the GE data within the identified groups. To validate the previous findings on the effectiveness of clustering quality criteria, structures containing 2, 3, and 4 clusters were examined. The experimental data consisted of 10,000 gene expression profiles from 156 lung cancer patients. The modeling results are presented in Table 1.

These findings demonstrate that a three-cluster configuration offers the best performance regarding classification accuracy and the loss function during neural network training. It's worth

mentioning that classification accuracy stays consistently high in all cases, due to the CNN's effectiveness with this data type and its resilience to noise. The classification accuracy was assessed on a test subset of data that was not used during the training phase of the neural network. Notably, for the three-cluster structure, a perfect classification accuracy of 100% was attained for the third cluster, which contains 4,964 genes, with the lowest loss function value. In the remaining clusters of this structure, 38 out of 39 objects in the test subset were accurately classified. These findings provide a strong foundation for improving diagnostic objectivity in complex diseases, allowing for balanced decision-making based on classification results from different gene expression clusters through the application of an alternative voting method.

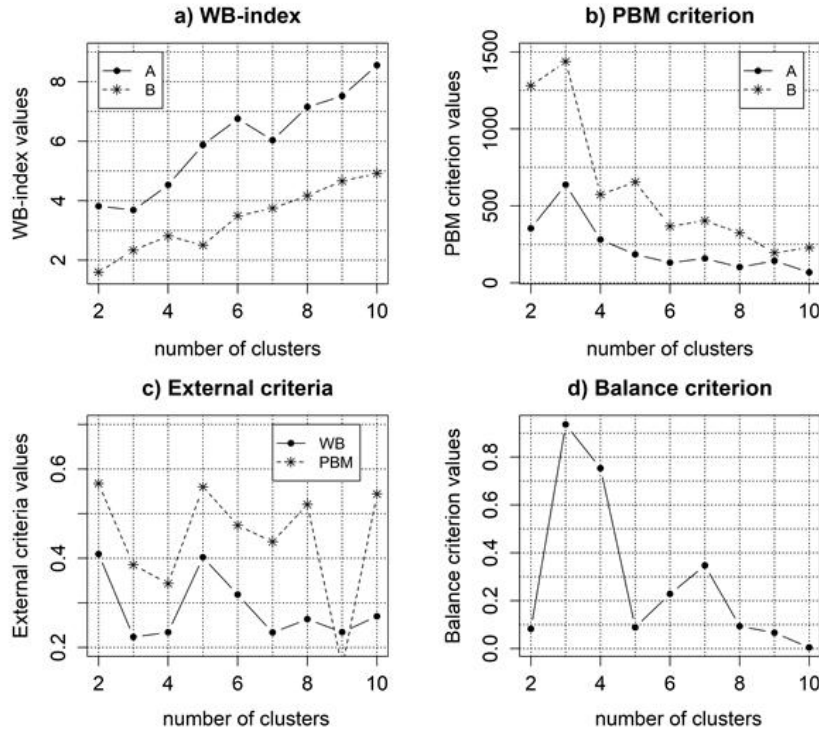


Figure 2: The modeling results demonstrate the practical execution of the hierarchical SC algorithm grounded in inductive approaches for analyzing complex systems

Table 1

The modeling findings for the application of the proposed model in processing GE data

Structure and results		CL 1	CL 2	CL 3	CL 4
Two-group clustering	Gene count	4074	4926	–	–
	Accuracy, %	95	97	–	–
	Loss	0.254	0.067	–	–
Three-group clustering	Gene count	2487	2549	4964	–
	Accuracy, %	97	97	100	–
	Loss	0.141	0.123	0.058	–
Four-group clustering	Gene count	1615	2779	4715	891
	Accuracy, %	97	97	97	95
	Loss	0.169	0.142	0.189	0.295

4. Conclusions

The hybrid inductive model for clustering gene expression profiles using spectral clustering has demonstrated high effectiveness in identifying co-expressed gene subsets. Through a series of

modeling experiments, we observed that the three-cluster structure consistently provided optimal performance, particularly in terms of classification accuracy and minimizing the loss function during CNN training. This method allowed for the efficient handling of high-dimensional and noisy data, which is often characteristic of gene expression datasets.

Our results validate the balance criterion as a robust metric for evaluating clustering quality, as it harmonizes internal and external clustering measures. Furthermore, the application of CNNs to gene expression data within clusters showed impressive accuracy, achieving perfect classification in some cases, confirming the potential of this combined approach for disease diagnosis and gene analysis.

This study opens new avenues for the practical application of hybrid models in the medical field, particularly in the diagnosis of complex diseases. The model's robustness to noise and its ability to produce reliable clustering outcomes highlight its potential for enhancing diagnostic objectivity in clinical settings. Future research could focus on refining the model by experimenting with different clustering techniques and expanding the approach to other disease types and datasets

5. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly exclusively for grammar and spelling checks, as well as for paraphrasing and rewording. After utilizing these services, the authors thoroughly reviewed and edited the content as needed and take full responsibility for the publication's final content.

References

- [1] H. Lodish, A. Berk, C.A. Kaiser, et al. *Molecular Cell Biology*, 9th edition. W.H. Freeman, 2021.
- [2] The Cancer Genome Atlas Program (TCGA). National Cancer Institution. Center for Cancer Genomics, 2024, July, 27, URL: <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>
- [3] S. Yosboon, N. Iam-On, T. Boongoen, P. Keerin, K. Kirimasthong. Optimised multiple data partitions for cluster-wise imputation of missing values in gene expression data, *Expert Systems with Applications* 257 (2024) 125040. doi: 10.1016/j.eswa.2024.125040.
- [4] Y. Xu, S. Wang, Q. Feng, et al. scCAD: Cluster decomposition-based anomaly detection for rare cell identification in single-cell expression data, *Nature Communications* 15 (1) (2024) 7561. doi: 10.1038/s41467-024-51891-9.
- [5] S. Luo, P.-L. Germain, M.D. Robinson, F. von Meyenn. Benchmarking computational methods for single-cell chromatin data analysis, *Genome Biology* 25(1) (2024) 225. doi: 10.1186/s13059-024-03356-x
- [6] Y. Hu, M. Xie, Y. Li, et al. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics, *Genome Biology* 25(1) (2024), 212. doi: 10.1186/s13059-024-03361-0.
- [7] I. Sakata, Y. Kawahara. Enhancing spectral analysis in nonlinear dynamics with pseudo eigenfunctions from continuous spectra, *Scientific Reports* 14(1) (2024) 19276. doi: 10.1038/s41598-024-69837-y
- [8] Y. Liu, X. Lin, Y. Chen, R. Cheng. Multi-order graph clustering with adaptive node-level weight learning, *Pattern Recognition* 156 (2024) 110843. doi: 10.1016/j.patcog.2024.110843
- [9] S. Babichev, L. Yasinska-Damri, I. Liakh, I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques, *Applied Sciences(Switzerland)* 13(10) (2023) 6022. doi: 10.3390/app13106022
- [10] S. Babichev, L. Yasinska-Damri, I. Liakh, J. Škvor. Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network, *Applied Sciences(Switzerland)* 12(22) (2022) 11795. doi: 10.3390/app122211795.

- [11] M. Romero, O. Ramírez, J. Finke, C. Rocha. Supervised Gene Function Prediction Using Spectral Clustering on Gene Co-expression Networks, *Studies in Computational Intelligence* 1016 (2022) 652–663. doi: 10.1007/978-3-030-93413-2_54.
- [12] K. Yu, W. Xie, L. Wang, S. Zhang, W. Li. Determination of biomarkers from microarray data using graph neural network and spectral clustering. *Scientific Reports*, 2021, 11(1), art. no. 23828. DOI: 10.1038/s41598-021-03316-6.
- [13] J. Liu, S. Ge, Y. Cheng, X. Wang. Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction, *Frontiers in Genetics* 12 (2021) 718915. doi: 10.3389/fgene.2021.718915.
- [14] S. Babichev, L. Yasinska-Damri, I. Liakh, B. Durnyak. Comparison analysis of gene expression profiles proximity metrics, *Symmetry* 13(10) (2021) 1812. doi: 10.3390/sym13101812.
- [15] J. Hou, J. Aerts, B. den Hamer, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction, *PLoS ONE* 5 (2010) e10312. doi: 10.1371/journal.pone.0010312.
- [16] Gene Expression Omnibus. 2024, July, 20. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi>
- [17] Q. Zhao, P. Fränti. WB-index: A sum-of-squares based index for cluster validity, *Data and Knowledge Engineering* 92 (2014) 77–89. doi: 10.1016/j.datak.2014.07.008.
- [18] J. Rojas-Thomas, M. Santos, M. Mora, N. Duro. Performance analysis of clustering internal validation indexes with asymmetric clusters, *IEEE Latin America Transactions* 17(5) (2019) 8891949, 807–814. doi: 10.1109/TLA.2019.8891949