Revolutionizing cancer classification: the snr-ogscc method for improved gene selection and clustering

Sara Haddou Bouazza, Jihad Haddou Bouazza
LAMIGEP, EMSI, Marrakech, Morocco

Article Info

Article history:

Received Sep 9, 2024
Revised Oct 19, 2024
Accepted Oct 23, 2024

Keywords:

Artificial intelligence
Cancer classification
Computer science
Feature selection
Machine learning

ABSTRACT

This study presents the signal-to-noise ratio optimized gene selection and clustering for cancer classification (SNR-OGSCC) methodology, aimed at enhancing classification accuracy while reducing the dimensionality of gene expression data across various cancer types. Implemented on a standard computational setup, the SNR-OGSCC method combines advanced filtering, clustering, and machine learning techniques, demonstrating significant improvements in classification accuracy on seven cancer datasets: leukemia, colon cancer, prostate cancer, lung cancer, lymphoma, central nervous system (CNS) tumors, and ovarian cancer. Notably, our approach achieved perfect accuracies of 100% for leukemia, lung cancer, and ovarian cancer, with high accuracies of 98.4% for colon cancer, 99.1% for prostate cancer, 98.3% for lymphoma, and 99.7% for CNS tumors, while requiring as few as 4–5 genes for effective classification. These findings highlight the efficiency and robustness of the SNR-OGSCC methodology, suggesting its potential to identify meaningful biomarkers and improve personalized cancer treatment strategies. Further validation with larger datasets and biological experiments is essential to confirm its applicability in clinical settings.

Corresponding Author:

Sara Haddou Bouazza
LAMIGEP, EMSI
Marrakech, Morocco
Email: sara.hb.sara@gmail.com

1. INTRODUCTION

Cancer remains one of the leading causes of morbidity and mortality worldwide, necessitating the urgent need for advanced diagnostic tools and methodologies. Traditional diagnostic methods often rely on histopathological examination and imaging techniques, which can be time-consuming and subjective. With the rise of genomic technologies [1], gene expression profiling has emerged as a promising avenue for enhancing cancer diagnosis and treatment strategies [2], [3]. This study focuses on optimizing gene selection and clustering methodologies to improve classification accuracy in cancer detection, specifically through the introduction of the signal-to-noise ratio optimized gene selection and clustering for cancer classification (SNR-OGSCC) approach.

Despite significant advancements in machine learning and bioinformatics, existing methods for gene selection and classification often face challenges related to high dimensionality and noise in gene expression data [4]. Many traditional approaches, such as the signal-to-noise ratio (SNR) method, have been employed to filter relevant genes [5], yet they may not adequately account for the complexities inherent in cancer datasets. Previous studies have demonstrated the effectiveness of various classification techniques [6], however, there remains a gap in methodologies that integrate advanced filtering, clustering, and machine learning to enhance both accuracy and interpretability.

Current literature reveals several methodologies with varying success rates. For example, research has shown that artificial neural networks and genetic algorithms can achieve high classification accuracies in leukemia cancer datasets, with reported accuracies of up to 98.5% [7]. Similarly, studies on colon and prostate cancer have employed techniques such as random forest and deep learning models, achieving accuracies of 95.16% [8] and 97.19% [9], respectively. While these results are promising, they often come at the cost of requiring extensive computational resources and large gene sets [10]. Our research aims to bridge these gaps by developing a methodology that not only improves classification accuracy but also reduces the number of genes needed for effective cancer diagnosis.

This study addresses a critical need in cancer diagnostics by presenting the SNR-OGSCC methodology. Our approach not only enhances classification accuracy but also effectively reduces the dimensionality of gene expression data. The subsequent sections of this paper will provide a detailed methodology, present comprehensive results across various cancer datasets, and engage in a discussion that highlights the implications of our findings. By applying this methodology, we aim to provide compelling evidence that SNR-OGSCC serves as a powerful tool for researchers and clinicians, ultimately paving the way for more precise and efficient cancer diagnostics and personalized treatment strategies.

2. METHOD

In this study, we propose an innovative methodology called SNR-OGSCC that tackles the complexity and high dimensionality of cancer classification using gene expression data. Our approach integrates advanced filtering, clustering, and classification techniques, with an emphasis on improving classification accuracy as the evaluation metric. By focusing on the identification of the most relevant genes, the SNR-OGSCC methodology aims to enhance the overall classification performance [11]–[13] across multiple types of cancer, leading to more precise predictions and better treatment strategies.

2.1. Gene selection and filtering process

The gene selection process forms the foundation of our methodology, as it directly influences the performance of the classification models [14], [15]. We employ the SNR as the primary filtering technique to identify the most informative genes from each dataset. SNR compares the mean difference in expression levels between cancerous and normal samples to the variation within each group. The formula used is as (1) [16]:

$$P(x) = \frac{\overline{x_{1J}} - \overline{x_{2J}}}{s_{1j} + s_{2j}} \tag{1}$$

Where $\overline{x_{yJ}}$ is the mean of attribute j and s_{yj} is its standard deviation for classes y=1, 2. We select the top 15% of genes based on their SNR scores, which ensures that only the most significant features, capable of maximizing the separation between cancerous and normal samples, are retained for further analysis.

2.2. Clustering for dimensionality reduction

Clustering is a key component in our methodology, designed to group similar genes and further reduce the dimensionality of the dataset [17], [18]. This step is crucial in enhancing classification accuracy by focusing on the most relevant features while minimizing noise. We utilize k-means as the primary clustering method [19], combined with the density-based spatial clustering of applications with noise (DBSCAN) for improved robustness [20]. To determine the optimal number of clusters for k-means, we use the elbow method [21] alongside the silhouette score, which assesses cluster compactness and separation. Post-k-means, DBSCAN is applied to refine the clustering process by identifying dense regions within the dataset.

2.3. Representative gene selection from clusters

Once the clusters are established, we select representative genes from each cluster based on their SNR scores. In the case of k-means clusters, the gene with the highest SNR score is chosen, whereas for DBSCAN, we select the most central gene in relation to the cluster centroid. This process ensures that the final set of genes is not only biologically significant but also optimal for the subsequent classification steps, preserving the integrity of the data.

2.4. Classification methods

For the classification of the gene sets, we apply three powerful machine learning algorithms: k-nearest neighbors (KNN), support vector machine (SVM), and linear discriminant analysis (LDA). Each classifier is tailored to handle high-dimensional gene expression data, leveraging different mathematical approaches. KNN uses Euclidean distance to classify samples based on the nearest neighbors [22], SVM employs a radial basis function (RBF) kernel to capture non-linear relationships [23], and LDA focuses on maximizing class separability by modeling normally distributed data [24]. This multi-classifier strategy ensures robust performance across various datasets.

2.5. Model evaluation

We rigorously evaluate the performance of the classifiers using stratified k-fold cross-validation [25]. This approach ensures that the class distributions in the training and testing sets are preserved across all folds, allowing for consistent and fair assessments of classification accuracy. By using accuracy as the sole evaluation metric, we maintain a clear and objective measure of how effectively the models distinguish between cancerous and non-cancerous samples. The classification accuracy is calculated as (2) [26]:

$$\mathrm{A_{ccuracy}} = 100 \, {}^{(\mathrm{T_P} + \mathrm{T_N})}\!\Big/_{(\mathrm{T_N} + \mathrm{T_N} + \mathrm{F_N} + \mathrm{F_P})} \tag{2}$$

True positive (TP): positive samples correctly recognized. True negative (TN): negative samples correctly recognized. False positive (FP): negative samples wrongly identified as positive. False negative (FN): positive samples wrongly identified as negative.

2.6. Parameters for cancer datasets

Our SNR-OGSCC methodology is applied to multiple cancer datasets, such as leukemia [27], colon cancer [28], prostate cancer [29], lung cancer [30], lymphoma [31], central nervous system (CNS) tumors [32], and ovarian cancer [33] each with different numbers of genes and samples. For each dataset, we use the same filtering and clustering techniques, with consistent classifier parameters Table 1. This uniform approach ensures comparability and reproducibility across the different types of cancer data, as shown in the parameters table.

Table 1. Proposed parameters for cancer datasets

Cancer type	Number of genes	Number of samples	K (k-means)	Classifier parameters

Leukemia	7,129	72 (47 ALL, 25 AML)	3	KNN: k=3;
				SVM: C=10, γ=0.01;
				LDA: ncomponents=1;

Colon cancer	6,500	62 (22 normal, 40 with cancer)	4	KNN: k=5;
				SVM: C=5, γ=0.01;
				LDA: ncomponents=1;

Prostate cancer	12,600	102 (52 normal, 50 with cancer)	5	KNN: k=5;
				SVM: C=15, γ=0.1;
				LDA: ncomponents=1;

Lung cancer	12,533	181 (31 MPM, 150 ADCA)	4	KNN: k=5;
				SVM: C=20, γ=0.05;
				LDA: ncomponents=1;

Lymphoma	7,070	77 (58 DLBCL, 19 FL)	3	KNN: k=5;
				SVM: C=10, γ=0.01;
				LDA: ncomponents=1;

Cns tumors	7,129	60 (39 survivors, 21 deceased)	4	KNN: k=3;
				SVM: C=5, γ=0.1;
				LDA: ncomponents=1;

Ovarian cancer	15,154	253 (91 normal, 162 with cancer)	5	KNN: k=5;
				SVM: C=15, γ=0.05;
				LDA: ncomponents=1;

3. RESULTS AND DISCUSSION

This section presents the findings of the SNR-OGSCC methodology. We discuss the experimental setup, results from the new filtering approach, interpretation of these results, limitations of the study, and conclude with the implications of our findings.

3.1. Experimental setup

The experiments were conducted on a standard laptop with an Intel® Core™ i5 CPU M 250 @ 2.4 GHz dual-core processor, 4 GB of RAM, running Windows 10 (64-bit). The MATLAB R2023a software was utilized to implement the SNR-OGSCC methodology, perform data analysis, and carry out classification tasks. Despite limited computational power, the system successfully processed the datasets, which ranged in size from 6,500 to over 15,000 genes and included sample sizes between 60 and 253. This setup allowed the study to balance computational efficiency with classification accuracy, proving that even with basic hardware; the proposed method could be implemented effectively.

3.2. Results corresponding to the new filtering approach

This study evaluated the performance of the SNR-OGSCC across seven cancer datasets. The method demonstrated substantial improvements in classification accuracy while significantly reducing the number of genes used for each cancer type. In particular, SNR-OGSCC consistently outperformed the baseline SNR method and the SNR+ clustering approach by optimizing gene selection through advanced clustering and filtering techniques. Table 2 summarizes the classification accuracies and the number of genes selected by each method (SNR, SNR+ clustering, and SNR-OGSCC) for the various cancer datasets.

The SNR-OGSCC method consistently reduced the number of genes needed for accurate classification while maintaining or improving classification accuracy across all datasets. This demonstrates the strength of integrating advanced clustering techniques, which significantly contribute to dimensionality reduction and selection of the most biologically relevant genes. By refining the gene selection process and optimizing classifiers, the method balances efficiency and performance, making it suitable for practical applications in gene expression analysis for cancer diagnosis.

Table 2. Classification accuracies and gene selection for various cancer datasets

Dataset	Feature selection method	KNN		SVM		LDA	
		Acc	Nbr genes	Acc	Nbr genes	Acc	Nbr genes
Leukemia	SNR	97.05	13	97.05	26	91.1	19
	SNR+ clustering	100	6	100	7	100	15
	SNR-OGSCC	100	4	100	5	100	4
Colon cancer	SNR	92.8	5	85.7	29	92.8	2
	SNR+ clustering	96	6	91.1	4	94	8
	SNR-OGSCC	98.4	5	97.2	7	98.4	5
Prostate cancer	SNR	90	22	92	8	92	4
	SNR+ clustering	98,2	2	92	3	92	2
	SNR-OGSCC	99.1	11	98.2	7	98.2	7
Lung cancer	SNR	97.3	6	97.3	33	97.3	64
	SNR+ clustering	100	13	99.3	10	99.3	14
	SNR-OGSCC	100	5	100	7	99.3	7
Lymphoma cancer	SNR	95.6	4	95.6	32	95.6	24
	SNR+ clustering	97	3	97	10	97	12
	SNR-OGSCC	98.3	5	98.3	9	97	6
CNS tumors	SNR	76.7	6	65.1	21	69.7	28
	SNR+ clustering	98.6	7	92.3	13	84	18
	SNR-OGSCC	99.7	4	98.6	12	99.7	9
Ovarian cancer	SNR	97.5	30	97.5	39	96.8	37
	SNR+ clustering	99.3	5	100	5	97,5	11
	SNR-OGSCC	100	4	100	4	99.3	9

3.3. Interpretation of results

The SNR-OGSCC method proved highly effective in improving classification accuracy across a diverse set of cancer datasets while significantly reducing the number of genes selected. This reduction in the number of genes directly impacts the interpretability of results and the computational load required for analysis. For example, in the leukemia dataset, only 4-5 genes were needed to achieve 100% accuracy, highlighting the method's capacity to retain only the most relevant features. This improved accuracy across multiple classifiers, including KNN, SVM, and LDA, reinforces the robustness of the proposed methodology. Moreover, the consistent performance of SNR-OGSCC across different cancer types demonstrates its flexibility and adaptability, making it a valuable tool for both researchers and clinicians.

The use of clustering to enhance gene selection efficiency is particularly important in complex datasets where high dimensionality can obscure meaningful patterns. The SNR-OGSCC method not only improves classification results but also facilitates the identification of potential biomarkers, which could have significant implications for personalized medicine. The high classification accuracy, even with a reduced set of genes, suggests that the method can help isolate biologically significant markers with greater precision.

3.4. Limitations

Despite the promising results, the SNR-OGSCC method has several limitations. First, the relatively small sample sizes in certain datasets, such as CNS tumors, could limit the generalizability of the findings. Applying the method to larger datasets would provide more robust evidence of its efficacy across a broader range of cancer types. Additionally, while the method performed well on standard hardware, larger and more complex datasets may require higher computational power, which could challenge the scalability of the approach.

Another limitation is the need for biological validation of the selected gene sets. Although the SNR-OGSCC method successfully identifies the most informative genes for classification, further experimental work is needed to confirm their biological relevance. Validation studies would be critical before these gene sets can be used as biomarkers in clinical settings. Additionally, the method’s reliance on machine learning algorithms tailored to high-dimensional data, such as SVM and KNN, could present challenges in certain applications, particularly where data distributions deviate from model assumptions.

3.5. Discussion

This study aimed to enhance cancer classification accuracy by employing the SNR-OGSCC methodology. While prior research has explored various machine learning and statistical techniques to classify cancer types, they often fall short in balancing classification accuracy with the dimensionality of gene expression data. For instance, Mallick *et al.* [34] and Nirmalakumari *et al.* [35] achieved high classification accuracies of 98.2% and 98.5% for leukemia, respectively, yet they do not specifically address the need for reduced gene sets to facilitate interpretability and clinical applicability.

Our findings indicate that the SNR-OGSCC method successfully achieved 100% accuracy across multiple cancer datasets, including leukemia, colon cancer, and lung cancer, with significantly reduced numbers of genes selected. For instance, in the leukemia dataset, only 4-5 genes were necessary to achieve optimal performance, contrasting with Mallick *et al.* [34], who required 13 genes to reach 98.2% accuracy. This demonstrates the method's capacity to isolate the most relevant features effectively.

When comparing our results with those of Shafi *et al.* [8], who reported an accuracy of 95.16% for colon cancer using random forest techniques, our SNR-OGSCC method surpassed this accuracy at 98.4% while employing a similar number of genes. This trend continued across other cancer types. For instance, while Fathi *et al.* [36] achieved 95% accuracy for lung cancer, our method provided perfect classification (100%) with fewer gene inputs. Similarly, studies like Rajaguru *et al.* [9] and Alshareef *et al.* [37] reached accuracies of 96.46% and 97.19% for prostate cancer, respectively; our methodology achieved 99.1% accuracy using a minimal gene subset.

In the lymphoma cancer dataset, Olaniran and Abdullah [38] reported an accuracy of 94.92% with a hybrid variational bayes (VB) approach, while our method achieved 98.3%, indicating a substantial improvement. Similarly, for CNS tumors, Painuli *et al.* [39] obtained an impressive accuracy of 99.6% using an logistic regression (LR)-based model, but our approach maintained high performance (99.7%) while simplifying the gene set. In ovarian cancer, Prabhakar and Lee [40] achieved a high classification accuracy of 99.48% using SVM-RBF with genetic bee colony optimization (GBCO). However, our methodology not only matched this performance but also consistently required fewer genes for classification, demonstrating its efficiency and effectiveness.

Despite the encouraging results, this study's limitations should be noted. The relatively small sample sizes in datasets, such as CNS tumors, where Painuli *et al.* [39] reported an accuracy of 99.6%, may affect the generalizability of our findings. Furthermore, the study's focus on classification accuracy may overlook other essential factors, such as precision and recall, which are vital for clinical relevance. The requirement for biological validation of the selected genes remains another challenge, as noted in various studies, including those by Prabhakar and Lee [40], which emphasize the need for further experimental confirmation of computational predictions.

Future research should focus on validating the SNR-OGSCC methodology across larger, more diverse datasets to confirm its efficacy and reliability in different clinical settings. Investigating the biological significance of the selected genes could also enhance the method’s applicability in personalized medicine. Additionally, exploring the integration of other machine learning algorithms, as seen in [41], [42], could further optimize classification accuracy and robustness.

In summary, our findings provide compelling evidence that the SNR-OGSCC methodology significantly enhances cancer classification accuracy while reducing the dimensionality of gene expression data. This approach addresses critical gaps in the existing literature by facilitating the identification of biologically relevant markers with potential clinical significance. The ability to achieve high classification accuracy with fewer genes positions SNR-OGSCC as a valuable tool in cancer diagnosis and treatment, warranting further exploration and validation in future studies.

4. CONCLUSION

The SNR-OGSCC methodology demonstrates significant advancements in the field of cancer diagnostics through enhanced classification accuracy and reduced dimensionality in gene expression data. Our study revealed that SNR-OGSCC consistently outperformed traditional methods across multiple cancer types, achieving remarkable accuracies, including 100% for leukemia, 98.4% for colon cancer, 99.1% for prostate cancer, and 100% for lung cancer. Additionally, it attained 98.3% for lymphoma cancer, 99.7% for CNS tumors, and 100% for ovarian cancer, while requiring significantly fewer genes for effective classification—

e.g., as low as 4 genes for leukemia and ovarian cancer. These improvements highlight the strength of integrating advanced gene selection and clustering techniques, enabling better separation between cancerous and non-cancerous samples. The results indicate that by optimizing gene selection through clustering, we can not only improve the robustness of cancer classification but also facilitate the identification of biologically relevant biomarkers. For instance, in the leukemia dataset, our method required only 4–5 genes to achieve perfect accuracy, demonstrating its capacity to retain only the most pertinent features. The potential implications of our findings extend beyond improved diagnostic capabilities. The SNR-OGSCC method paves the way for personalized treatment strategies by enabling the identification of specific genetic markers associated with different cancer types. As such, our methodology serves as a vital tool for researchers and clinicians striving to enhance cancer detection and treatment. However, to fully realize the potential of the SNR-OGSCC approach, further validation is required. Future studies should focus on applying this methodology to larger datasets and diverse cancer types to assess its generalizability and robustness. Additionally, biological validation of the selected gene sets will be essential to confirm their relevance in clinical settings. Ultimately, this research contributes to the ongoing effort to refine cancer classification techniques and emphasizes the importance of integrating computational methods with biological insights for improved patient outcomes.

REFERENCES

[1] R. H. Elden, V. F. Ghonim, M. M. A. Hadhoud, and W. Al-Atabany, “Transcriptomic marker screening for evaluating the mortality rate of pediatric sepsis based on Henry gas solubility optimization,” *Alexandria Engineering Journal*, vol. 68, pp. 693–707, Apr. 2023, doi: 10.1016/j.aej.2022.12.027.

[2] M. Elloumi, M. A. Ahmad, A. H. Samak, A. M. Al-Sharafi, D. Kihara, and A. I. Taloba, “Error correction algorithms in non-null aspheric testing next generation sequencing data,” *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 9819–9829, Dec. 2022, doi: 10.1016/j.aej.2022.03.041.

[3] A. Akgül, S. H. A. Khoshnaw, and H. M. Rasool, “Minimizing cell signalling pathway elements using lumping parameters,” *Alexandria Engineering Journal*, vol. 59, no. 4, pp. 2161–2169, Aug. 2020, doi: 10.1016/j.aej.2020.01.041.

[4] Y. Esmaeili *et al.*, “Exploring the evolution of tissue engineering strategies over the past decade: from cell-based strategies to gene-activated matrix,” *Alexandria Engineering Journal*, vol. 81, pp. 137–169, Oct. 2023, doi: 10.1016/j.aej.2023.08.080.

[5] M. Shaheen, N. Naheed, and A. Ahsan, “Relevance-diversity algorithm for feature selection and modified Bayes for prediction,” *Alexandria Engineering Journal*, vol. 66, pp. 329–342, Mar. 2023, doi: 10.1016/j.aej.2022.11.002.

[6] L. Zhang, L. Li, M. Tang, Y. Huan, X. Zhang, and X. Zhe, “A new approach to diagnosing prostate cancer through magnetic resonance imaging,” *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 897–904, Feb. 2021, doi: 10.1016/j.aej.2020.10.018.

[7] S. J. Susmi, H. K. Nehemiah, A. Kannan, and J. Christopher, “Relevant gene selection and classification of leukemia gene expression data,” in *Emerging Research in Computing, Information, Communication and Applications*, Singapore: Springer, 2016, pp. 503–510, doi: 10.1007/978-981-10-0287-8_47.

[8] A. S. M. Shafi, M. M. I. Molla, J. J. Jui, and M. M. Rahman, “Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques,” *SN Applied Sciences*, vol. 2, no. 7, Jul. 2020, doi: 10.1007/s42452-020-3051-2.

[9] K. Nirmalakumari, H. Rajaguru, and P. Rajkumar, “Microarray prostate cancer classification using eminent genes,” in *2021 Smart Technologies, Communication and Robotics (STCR)*, Oct. 2021, pp. 1–5, doi: 10.1109/STCR51658.2021.9588811.

[10] E. Badr, S. Almotairi, M. A. Salam, and H. Ahmed, “New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis,” *Alexandria Engineering Journal*, vol. 61, no. 3, pp. 2520–2534, Mar. 2022, doi: 10.1016/j.aej.2021.07.024.

[11] A. S. Alzahrani, R. A. Shah, Y. Qian, and M. Ali, “A novel method for feature learning and network intrusion classification,” *Alexandria Engineering Journal*, vol. 59, no. 3, pp. 1159–1169, Jun. 2020, doi: 10.1016/j.aej.2020.01.021.

[12] T. Althobaiti, S. Althobaiti, and M. M. Selim, “An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making,” *Alexandria Engineering Journal*, vol. 94, pp. 311–324, May 2024, doi: 10.1016/j.aej.2024.03.044.

[13] F. Noman *et al.*, “Multistep short-term wind speed prediction using nonlinear auto-regressive neural network with exogenous variable selection,” *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1221–1229, Feb. 2021, doi: 10.1016/j.aej.2020.10.045.

[14] N. M. Sallam, A. I. Saleh, H. Arafat Ali, and M. M. Abdelsalam, “An efficient EGWO algorithm as feature selection for B-ALL diagnoses and its subtypes classification using peripheral blood smear images,” *Alexandria Engineering Journal*, vol. 68, pp. 39–66, Apr. 2023, doi: 10.1016/j.aej.2023.01.004.

[15] A. M. Khalid, W. Said, M. Elmezain, and K. M. Hosny, “A new binary object-oriented programming optimization algorithm for solving high-dimensional feature selection problem,” *Alexandria Engineering Journal*, vol. 85, pp. 72–85, Dec. 2023, doi: 10.1016/j.aej.2023.11.021.

[16] D. M. Mahalakshmi and S. Sumathi, “Brain tumour segmentation strategies utilizing mean shift clustering and content based active contour segmentation,” *ICTACT Journal on Image and Video Processing*, vol. 9, no. 4, pp. 2002–2008, May 2019, doi: 10.21917/ijivp.2019.0284.

[17] A. S. Negm, O. A. Hassan, and A. H. Kandil, “A decision support system for acute leukaemia classification based on digital microscopic images,” *Alexandria Engineering Journal*, vol. 57, no. 4, pp. 2319–2332, Dec. 2018, doi: 10.1016/j.aej.2017.08.025.

[18] E. E. Nithila and S. S. Kumar, “Segmentation of lung nodule in CT data using active contour model and fuzzy C-mean clustering,” *Alexandria Engineering Journal*, vol. 55, no. 3, pp. 2583–2588, 2016, doi: 10.1016/j.aej.2016.06.002.

[19] Y. Abo-Elnaga and S. Nasr, “K-means cluster interactive algorithm-based evolutionary approach for solving bilevel multi-objective programming problems,” *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 811–827, Jan. 2022, doi: 10.1016/j.aej.2021.04.098.

[20] H. Wu, Y. Chen, W. Zhu, Z. Cai, A. A. Heidari, and H. Chen, “Feature selection in high-dimensional data: an enhanced RIME optimization with information entropy pruning and DBSCAN clustering,” *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 9, pp. 4211–4254, Sep. 2024, doi: 10.1007/s13042-024-02143-1.

[21] E. Guven, “Decision of the optimal rank of a nonnegative matrix factorization model for gene expression data sets utilizing the unit invariant knee method: development and evaluation of the elbow method for rank selection,” *JMIR Bioinformatics and Biotechnology*, vol. 4, Jun. 2023, doi: 10.2196/43665.

[22] H. E. Saroğlu *et al.*, “Machine learning, IoT and 5G technologies for breast cancer studies: a review,” *Alexandria Engineering Journal*, vol. 89, pp. 210–223, Feb. 2024, doi: 10.1016/j.aej.2024.01.043.

[23] M. Roshani *et al.*, “Evaluation of flow pattern recognition and void fraction measurement in two phase flow independent of oil pipeline’s scale layer thickness,” *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1955–1966, Feb. 2021, doi: 10.1016/j.aej.2020.11.043.

[24] I. G. P. S. Wijaya, I. B. K. Widiartha, F. Bimantoro, and A. W. Septiadi, “Buildings cracks classification using zoning and invariant moment features and quadratic discriminant analysis classifier,” *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, Dec. 2019, doi: 10.24843/LKJITI.2019.v10.i03.p04.

[25] T. Mahesh, A. Kaladevi, J. Balajee, V. Vivek, M. Prabu, and V. Muthukumaran, “An efficient ensemble method using k-fold cross validation for the early detection of benign and malignant breast cancer,” *International Journal of Integrated Engineering*, vol. 14, no. 7, Dec. 2022, doi: 10.30880/ijie.2022.14.07.015.

[26] H. A. Afify, “Evaluation of change detection techniques for monitoring land-cover changes: a case study in new Burg El-Arab area,” *Alexandria Engineering Journal*, vol. 50, no. 2, pp. 187–195, Jun. 2011, doi: 10.1016/j.aej.2011.06.001.

[27] T. R. Golub *et al.*, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999, doi: 10.1126/science.286.5439.531.

[28] C. Park and S.-B. Cho, “Evolutionary ensemble classifier for lymphoma and colon cancer classification,” in *The 2003 Congress on Evolutionary Computation, 2003. CEC ’03.*, vol. 4, pp. 2378–2385, doi: 10.1109/CEC.2003.1299385.

[29] A. H. Chen, Y.-W. Tsau, and C.-H. Lin, “Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles,” *BMC Genomics*, vol. 11, no. 1, 2010, doi: 10.1186/1471-2164-11-274.

[30] G. J. Gordon *et al.*, “Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma,” *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[31] M. A. Shipp *et al.*, “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning,” *Nature Medicine*, vol. 8, no. 1, pp. 68–74, Jan. 2002, doi: 10.1038/nm0102-68.

[32] J.-Y. Yeh, “Applying data mining techniques for cancerclassification on gene expression data,” *Cybernetics and Systems*, vol. 39, no. 6, pp. 583–602, Aug. 2008, doi: 10.1080/01969720802188292.

[33] E. F. Petricoin *et al.*, “Use of proteomic patterns in serum to identify ovarian cancer,” *The Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002, doi: 10.1016/S0140-6736(02)07746-2.

[34] P. K. Mallick, S. K. Mohapatra, G.-S. Chae, and M. N. Mohanty, “Convergent learning–based model for leukemia classification from gene expression,” *Personal and Ubiquitous Computing*, vol. 27, no. 3, pp. 1103–1110, Jun. 2023, doi: 10.1007/s00779-020-01467-3.

[35] K. Nirmalakumari, H. Rajaguru, and P. Rajkumar, “Leukemia cancer classification using extrusive genes from microarray data,” *Second International Conference on Circuits, Signals, Systems and Securities (ICCSSS - 2022)*, vol. 2725, no. 1, 2023, doi: 10.1063/5.0125232.

[36] H. Fathi, H. AlSalman, A. Gumaei, I. I. M. Manhrawy, A. G. Hussien, and P. El-Kafrawy, “An efficient cancer classification model using microarray and high-dimensional data,” *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/7231126.

[37] A. M. Alshareef *et al.*, “Optimal deep learning enabled prostate cancer detection using microarray gene expression,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–12, Mar. 2022, doi: 10.1155/2022/7364704.

[38] O. R. Olaniran and M. A. A. Abdullah, “Subset selection in high-dimensional genomic data using hybrid variational bayes and bootstrap priors,” *Journal of Physics: Conference Series*, vol. 1489, no. 1, Mar. 2020, doi: 10.1088/1742-6596/1489/1/012030.

[39] D. Painuli, S. Bhardwaj, and U. Kose, “Optimized diagnosis of central nervous system (CNS) cancer using gene expression microarray & machine learning (ML) methods,” *European Chemical Bulletin*, vol. 12, no. 10, pp. 9757-9771, 2023.

[40] S. K. Prabhakar and S.-W. Lee, “An integrated approach for ovarian cancer classification with the application of stochastic optimization,” *IEEE Access*, vol. 8, pp. 127866–127882, 2020, doi: 10.1109/ACCESS.2020.3006154.

[41] A. Al-Murad and M. F. Hossain, “An integrated feature selection method for neural network to classify ovarian cancer,” in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, Jul. 2021, pp. 1–6, doi: 10.1109/ACMI53878.2021.9528106.

[42] M. Z. H. Akmal, and D. Fitrianah, “Exploring feature selection for microarray classification,” *International Journal of Computing and Digital Systems*, vol. 16, no. 1, 2024.

BIOGRAPHIES OF AUTHORS

Sara Haddou Bouazza holds a Doctorate in Electrical Engineering and Informatics, as well as a Master's in Electrical Engineering from Cadi Ayyad University, Marrakech. She also completed her Bachelor's in Physical Sciences. Currently, she is a professor and researcher at the LAMIGEP laboratory, EMSI Marrakech. Her research includes AI techniques for cancer classification, gene expression analysis, and security challenges in IoT environments. She has published numerous papers, including recent work on leukemia classification and AI in CNS tumors. She can be contacted at email: sara.hb.sara@gmail.com.

Jihad Haddou Bouazza is an engineer specializing in software engineering and image processing from IGA Institut Supérieur du Génie Appliqué, Marrakech. Currently, he serves as a senior full stack developer & tech lead at Nexular Corp. He is certified in Python, machine learning, and as a certified network security specialist (CNSS). His research includes pattern recognition using artificial intelligence, with a publication presented at the GAST24 congress. He can be contacted at email: haddou.jihad@gmail.com.