

A two-step intelligent framework for gene expression-based cancer diagnosis

Sara Haddou Bouazza¹, Jihad Haddou Bouazza²

¹LAMIGEP Laboratory, EMSI Moroccan School of Engineering, Marrakesh, Morocco

²IGA-Institut Supérieur du GénieAppliqué, Casablanca, Morocco

Article Info

Article history:

Received Apr 22, 2025

Revised Sep 8, 2025

Accepted Oct 16, 2025

Keywords:

Cancer classification

Computer science

Feature selection

Image processing

Machine learning

ABSTRACT

DNA microarray technology has advanced cancer diagnosis by enabling large-scale gene expression analysis, yet challenges remain in selecting relevant genes and achieving accurate classification. This study introduces two novel methods: the three-stage gene selection (3SGS) method and the statistics classifier (SC). By eliminating redundant, noisy, and less informative genes, the 3SGS method effectively lowers the dimensionality of gene expression data, while the SC classifier uses statistical measures of gene expression to classify samples with high accuracy and speed. Evaluated on leukemia, prostate cancer, and colon cancer datasets, the 3SGS method effectively identified minimal yet informative gene subsets, achieving 100% accuracy for leukemia, 99.3% for prostate cancer, and 97% for colon cancer. The SC classifier consistently outperformed traditional models in both accuracy and computational efficiency, completing predictions in under 2 seconds per dataset. Compared to conventional classifiers, it requires no parameter tuning and performs reliably even with small gene sets. While promising, future work should address multiclass classification and clinical validation to broaden the framework's applicability. Together, these methods offer a precise and rapid cancer classification framework, supporting early diagnosis and personalized treatment strategies across diverse cancer types.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sara Haddou Bouazza

LAMIGEP Laboratory, EMSI Moroccan School of Engineering

Marrakech, Morocco

Email: sara.hb.sara@gmail.com

1. INTRODUCTION

DNA microarray technology has greatly improved cancer diagnosis and prognosis by allowing the parallel examination of thousands of gene expression profiles [1]–[4]. This advancement has enhanced our knowledge of gene interactions and their contribution to cancer development. However, a major challenge stems from the imbalance between the extremely large number of genes and the limited availability of samples. Since not all genes are involved in cancer progression and many are correlated, relying on the complete gene set may increase complexity and lower prediction accuracy [5]. This underscores the importance of applying effective feature selection methods to enhance classification performance.

Microarray gene classification, a supervised learning task, relies on labeled gene expression data to predict disease classes. Its success depends heavily on selecting the most relevant features [6], [7]. While traditional statistical methods have been widely used [6], machine learning techniques now play a vital role in handling complex microarray data [8], [9]. Yet, the high dimensionality of these datasets often leads to overfitting, emphasizing the importance of dimensionality reduction through feature selection [6], [10].

Feature selection aims to identify genes that show significant differences across disease classes. Approaches include filter methods, which rank genes based on statistical measures like p-values [11], [12], signal-to-noise ratio (SNR), mRMR, ReliefF, and performance of upper limb (PUL) scores [13], and wrapper methods, which use classifiers to evaluate gene subsets [14]. While filters are efficient, they often ignore gene interactions, whereas wrappers offer greater accuracy at a higher computational cost. Hybrid methods combine both strategies for optimal results.

To address the limitations of conventional approaches; such as overfitting, lack of scalability, and computational burden; we propose a two-step intelligent framework combining a hybrid gene selection strategy and a statistical classification mechanism. The three-stage gene selection (3SGS) method sequentially filters, evaluates, and compresses gene subsets to enhance predictive power while reducing dimensionality. Complementing this, the statistics classifier (SC) uses interpretable statistical boundaries for classification, enabling fast and precise decisions with minimal parameter tuning.

2. METHOD

This study seeks to design a reliable gene selection approach for accurate tumor classification based on microarray data. The workflow consists of several stages: data preprocessing to improve quality, advanced selection techniques to extract the most informative genes, and the application of refined classification models. The following section details the materials and methods employed, with particular emphasis on the strategies adopted for identifying relevant genes.

2.1. Gene selection

To identify genes relevant for tumor classification from microarray datasets, we adopted a three-phase selection strategy. The first phase applied a filtering step to discard largely irrelevant genes, thereby simplifying the dataset. This filtering relied on three parametric techniques—SNR, correlation coefficient (CC), and ReliefF—to highlight the most informative genes for subsequent analysis.

The SNR technique finds expression patterns with the highest mean expression difference between two groups and the least fluctuation within each group [15], [16]. This criterion, proposed by [17], rates genes according to (1).

$$P(j) = \frac{M1j - M2j}{S1j + S2j} \quad (1)$$

Here, M_{kj} and S_{kj} represent the mean and standard deviation of gene j within class $k=1, 2$. Larger values of $|P(j)|$ suggest a stronger association between the gene's expression and class differentiation. The Pearson CC [18] evaluates how strongly two genes are linearly related. Values close to +1 indicate a direct relationship, those near -1 reflect an inverse relationship, and values around 0 suggest no linear correlation. The coefficient for gene j is computed as (2).

$$r_j = \frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

Where r is the Pearson correlation score, X_{ij} is the i^{th} sample value for the gene j , Y_i is the corresponding class, $\bar{X}_j = 1/n \sum_{i=1}^n X_{ij}$, and \bar{Y} are the means for gene j and the classes, respectively. ReliefF is a supervised feature-weighting technique developed in [19] and further enhanced by [20]. It assesses the quality of qualities as (3).

$$WA = wd - \sum_{j=1}^k \frac{\text{diff}(A_i, X_i, \text{hits } j)}{m * k} + \sum_{c \neq \text{class}(X_i)} \frac{p(c)}{1 - p(\text{class}(X_i))} \sum_{j=1}^k \frac{\text{diff}(A_i, X_i, \text{misses } j)}{m * k} \quad (3)$$

Where the distance used is defined by (4).

$$\text{diff}(A_i, X1, X2) = \frac{|\text{value}(A_i, X1) - \text{value}(A_i, X2)|}{\max(A) - \min(A)} \quad (4)$$

Here, X_i is an instance described by the vector A_i of n genes, m is the number of process repetitions, k is the number of nearest misses, and hits and misses refer to nearest hit and miss instances, respectively. The filter selection approach evaluates each gene and selects a subset of relevant ones. However, some noisy genes may still reduce classification accuracy [9]. To address this, the second stage uses a wrapper strategy, starting

with one gene and gradually adding others from the filtered subset, retaining only those that improve accuracy. The first two steps identify the most informative genes. The final stage refines this selection, choosing the smallest subset that achieves the highest accuracy on the training set.

2.2. Algorithm of our selection approach: three-stage gene selection

The 3SGS approach is introduced to improve both the accuracy and reliability of gene selection in tumor classification. It systematically reduces the dimensionality of high-throughput gene expression data while maximizing predictive performance. The algorithm begins with a dataset composed of training data (X_{train}, Y_{train}) , containing n genes across m samples with known class labels, and a test dataset X_{test} featuring the same genes but unknown labels. The user also defines the desired number k of top-ranked genes to be initially selected, as well as the classifier to be employed (e.g., support vector machine (SVM) and k-nearest neighbors (KNN)).

- i) Step 1: feature selection via filter-based ranking, the procedure begins with the calculation of ranking scores for each gene in the training dataset, employing filter-based measures such as SNR, CC, or ReliefF. These scores are stored in a list termed `Gene_Scores`. If multiple metrics are utilized, normalization is performed to scale the scores uniformly between 0 and 1 to ensure fair comparison. The genes are then sorted in descending order based on their scores, and the top k genes are selected to form an initial subset referred to as `Top_Ranked_Genes`.
- ii) Step 2: recursive subset refinement for accuracy maximization, the next stage involves recursively evaluating gene subsets to identify the combination that yields the highest classification accuracy. Starting from an empty set $S=\{\}$, each gene in `Top_Ranked_Genes` is iteratively added to a temporary set $S_{temp}=S \cup \{g\}$. The classifier is trained on this subset, and performance is assessed using cross-validation. If the accuracy improves or remains the same, S is updated to include g ; otherwise, g is discarded. The procedure includes an early stopping criterion to halt the iteration once no accuracy gain is observed over several iterations, thereby mitigating overfitting and reducing computational cost. The final subset, referred to as `High_Accuracy_Genes`, contains the genes that provide the greatest contribution to classification accuracy.
- iii) Step 3: redundancy reduction to identify marker genes, to further refine the gene set, redundancy among genes in `High_Accuracy_Genes` is analyzed using correlation or mutual information. Genes exhibiting high redundancy or minimal contribution to accuracy are pruned. Optionally, principal component analysis (PCA) may be used to aid in detecting overlapping expression patterns. The classifier is then retrained on the reduced gene set to ensure classification performance is not compromised. If accuracy drops, previously excluded genes may be reconsidered. The finalized, non-redundant, and highly informative genes are retained as `Marker_Genes`.
- iv) Step 4: final classification using selected marker genes in the final stage, the classifier is retrained on the complete training dataset using only the selected `Marker_Genes`. After validating the model via cross-validation to ensure generalizability, it is applied to the test data X_{test} . The classifier predicts the class labels for the unseen samples, producing the final output, `predicted_labels`, which represent the cancer class predictions based on a minimal yet informative gene set.

2.3. Classification methods

We evaluated feature selection methods with five classifiers: KNN, SVM, linear discriminant analysis (LDA), decision tree (DT), and naive Bayes (NB).

- i) KNN classifies samples based on proximity, using Euclidean distance to identify the KNN [21], [22].
- ii) SVM constructs a maximum-margin hyperplane in a high-dimensional space via kernel functions, optimizing separation between classes [23].
- iii) LDA identifies a linear combination of features that improves class separation by maximizing between-class variance while minimizing within-class variance [24].
- iv) DT uses a hierarchical structure of decision rules to model outcomes, making it widely applicable in machine learning and data mining [25], [26].
- v) NB is a probabilistic model based on Bayes' theorem, assuming feature independence given the class, and is known for its simplicity and efficiency [25].

To assess classifier performance, we use classification accuracy [26]–[27].

2.4. Our proposition for gene classification for binary class problems

The SC is based on leveraging statistical descriptors such as minimum, maximum, mean, and standard deviation of gene expression to assess class membership. Our novel gene classification approach, based on gene expression profiling data, introduces a streamlined two-step process designed for clarity and precision. The first step involves calculating key statistical measures for each selected gene within the training samples across both classes. These measures comprise the minimum (min), maximum (max), mean

(mean), and standard deviation (Std) of gene expression values, which together define the expression profile of each gene in its corresponding class.

In the second step, test samples are classified by comparing their gene expression levels against these statistical ranges. If a test sample's expression value for a given gene falls within the range defined by min and max for a specific class, the sample is directly assigned to that class for that gene. When the expression value lies outside these boundaries, the sample is assigned to the class whose range—specifically the mean±Std interval is closest to the test value, ensuring accurate classification even in cases of outliers or variability.

To finalize the classification, a voting mechanism is applied. Each selected gene casts a "vote" based on its classification result. The overall class assigned to the sample is determined by the majority of votes, which helps balance individual gene-level variances and leads to a more reliable prediction. This process offers a systematic method for gene-based classification by leveraging statistical boundaries and a consensus-driven voting strategy, ensuring both robustness and interpretability.

2.5. Datasets

This study evaluates the proposed method using three publicly available binary-class microarray datasets: leukemia, prostate cancer, and colon cancer. Each dataset consists of gene expression profiles represented as matrices, with rows corresponding to samples and columns to gene features.

- i) The leukemia dataset consists of 72 samples, with 38 used for training and 34 for testing, categorized into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Each sample contains 7,129 gene expression features [28].
- ii) The prostate cancer dataset comprises 101 samples (81 training and 20 testing), including 52 tumor and 49 non-tumor cases. Gene expression was measured using oligonucleotide microarrays covering approximately 12,600 genes [29].
- iii) The colon cancer dataset contains 62 samples (48 training and 14 testing), with 40 tumor and 22 normal tissue samples. Gene expression was recorded using an Affymetrix oligonucleotide array, from which 2,000 genes were selected based on measurement reliability [30].

These datasets provide a reliable benchmark for assessing the generalizability and performance of gene selection and classification strategies in cancer diagnosis.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we outline the outcomes of our research, covering each step from data preprocessing to the implementation of our gene selection strategy and the final classification. We evaluated our method across multiple datasets to assess its effectiveness and robustness in identifying the most significant genes for tumor classification. Additionally, we outline the tools and techniques employed and provide an in-depth analysis of the classification results obtained using various classifiers.

3.1. Data preprocessing

Pretreatment filters out non-informative genes with consistent expression across classes. Leukemia data is excluded from this process. The method involves thresholding, filtering, and logarithmic transformation [31]. Thresholding keeps values between 100 and 16,000. Filtering removes genes with low variability, retaining those where the ratio S_{\max}/S_{\min} exceeds 5 and the difference $S_{\max}-S_{\min}$ is greater than 500. A logarithmic transformation then normalizes the data. This process reduces the dataset from 7,129 to 3,051 genes, keeping the most informative features for analysis.

3.2. Performance analysis of the three-stage gene selection method

The proposed 3SGS method was evaluated on leukemia, prostate, and colon cancer microarray datasets. In each case, training datasets were applied for gene selection and model construction, and test datasets assessed classification performance using five classifiers. Table 1 presents the classification results on the leukemia dataset using five classifiers, with different feature selection strategies. The 3SGS-enhanced versions of SNR, CC, and ReliefF consistently achieved perfect accuracy (100%) using as few as three to four genes. This highlights the ability of 3SGS to reduce dimensionality while maintaining or improving predictive performance. The most relevant genes identified were Y00787, M23197, and M27891 effectively differentiating between ALL and AML subtypes.

As shown in Table 2, the 3SGS method significantly improved classification accuracy across all classifiers on the prostate cancer dataset. It achieved up to 95% accuracy using only three to four genes, compared to standard methods requiring many more features. The key genes selected—37720_at, 37639_at, and 40435_at—enabled robust discrimination between tumor and normal samples.

Table 3 highlights the performance on the colon cancer dataset. Once again, the 3SGS method demonstrated its effectiveness by enhancing accuracy while simultaneously reducing the number of selected genes. Notably, the KNN classifier reached 96% accuracy using only four genes via SNR_3SGS, compared to 92.8% with the standard SNR method. The selected genes; M63391, H64489, T92451, and T57619; enabled precise differentiation between cancerous and normal tissue samples.

Table 1. Leukemia dataset–performance of classifiers with gene selection approaches

Feature selection methods	KNN		SVM		LDA		DT		NB	
SNR	100%	(13)	97%	(4)	97%	(9)	97%	(3)	97%	(5)
SNR_3SGS	100%	(3)	97%	(2)	97%	(4)	97%	(3)	97%	(4)
CC	100%	(50)	97%	(3)	100%	(93)	97%	(4)	97%	(6)
CC_3SGS	100%	(4)	97%	(3)	100%	(5)	100%	(4)	100%	(4)
ReliefF	97%	(41)	97%	(2)	97%	(69)	94%	(11)	94%	(5)
ReliefF_3SGS	100%	(4)	97%	(1)	100%	(4)	97%	(4)	97%	(4)

Table 2. Prostate cancer dataset–performance of classifiers with gene selection approaches

Feature selection methods	KNN		SVM		LDA		DT		NB	
SNR	90%	(22)	92%	(8)	92%	(4)	91%	(19)	91%	(45)
SNR_3SGS	95%	(3)	95%	(2)	92%	(1)	92%	(3)	92%	(4)
CC	85%	(6)	92%	(44)	92%	(6)	92%	(46)	91%	(65)
CC_3SGS	92%	(4)	95%	(3)	95%	(3)	95%	(4)	92%	(3)
ReliefF	90%	(32)	92%	(34)	91%	(75)	90%	(36)	92%	(50)
ReliefF_3SGS	95%	(3)	95%	(3)	91%	(1)	91%	(4)	95%	(4)

Table 3. Colon cancer dataset–performance of classifiers with gene selection approaches

Feature selection methods	KNN		SVM		LDA		DT		NB	
SNR	92.8%	(5)	85.7%	(29)	92.8%	(2)	91%	(21)	85.7%	(22)
SNR_3SGS	96%	(4)	92.8%	(9)	94%	(5)	92.8%	(6)	91%	(6)
CC	92.8%	(7)	85.7%	(2)	92.8%	(27)	92.8%	(21)	85.7%	(5)
CC_3SGS	96%	(5)	95%	(4)	95%	(4)	95%	(5)	91%	(4)
ReliefF	85.7%	(40)	85.7%	(11)	78.5%	(78)	90%	(26)	85.7%	(64)
ReliefF_3SGS	91%	(5)	92.8%	(4)	91%	(4)	94%	(5)	92.8%	(5)

3.3. Results of the proposed statistics classifier

This subsection compares the proposed SC with five conventional classifiers, on leukemia, prostate, and colon cancer datasets, considering both accuracy and computation time. All models were trained using genes selected via the SNR-based SNR_3SGS method. For leukemia, SC achieved 100% accuracy using three genes, matching KNN but with the shortest runtime (1.9 seconds). In prostate cancer, SC reached the highest accuracy (99.3%) with the same minimal time, outperforming others (92–95%). For colon cancer, SC attained 97% accuracy, again surpassing traditional classifiers in both accuracy and speed. Across all datasets, SC consistently delivered top performance with significantly lower computational cost. Classification results and timing are summarized in Table 4.

Table 4. Runtime for cancer classification

	Selection method	KNN	SVM	LDA	DT	NB	SC
Leukemia	SNR_3SGS (%)	100	97	97	97	97	100
	run time (s)	2.3	2.4	3.1	3.3	2.7	1.9
Prostate cancer	SNR_3SGS (%)	95	95	95	92	95	99.3
	run time (s)	2.3	2.4	3.1	3.3	2.7	1.9
Colon cancer	SNR_3SGS (%)	96	92.8	94	94	94	97
	run time (s)	2.3	2.5	3.1	3.4	2.7	1.9

3.4. Discussion

Recent studies have introduced hybrid gene selection strategies for cancer classification, achieving high accuracy with small gene subsets. For instance, genetic algorithm (GA)–Isomap reached 100% in leukemia (43 genes) and 85.8% in colon cancer (11 genes) [31], extreme gradient boosting (XGBoost)–multi-

objective genetic algorithm (MOGA) obtained 100% in leukemia (7 genes) and 90.2% in colon cancer (62 genes) [32], a hierarchical fuzzy–analytic hierarchy process (AHP) approach achieved 100% in leukemia (15 genes) and 96% in prostate cancer (30 genes) [33], while an entropy-based method reported 100% in leukemia (10 genes) and 91.9% in colon cancer (9 genes) [34]. Although effective, these methods are often computationally intensive, parameter-sensitive, and less generalizable.

To overcome these challenges, we proposed the 3SGS method and the SC. The 3SGS method combines filter-based ranking (SNR, CC, and ReliefF), recursive evaluation, and redundancy reduction, balancing the efficiency of filters with the accuracy of wrappers. The SC classifier applies simple statistical boundaries (min, max, mean, and Std) with a voting mechanism, enabling fast, interpretable, and robust predictions without parameter tuning, making it suitable for clinical settings.

Experiments confirmed the framework’s effectiveness: 3SGS achieved 100% accuracy in leukemia with three genes (M27891, M23197, and Y00787), 95% in prostate cancer (37720_at, 37639_at, and 40435_at), and 96% in colon cancer (M63391, H64489, T92451, and T57619). The SC further improved results to 99.3% in prostate and 97% in colon cancer, with runtimes as low as 1.9 seconds. These outcomes demonstrate that the 3SGS–SC framework offers precision, efficiency, and interpretability, showing strong potential for personalized cancer diagnosis and clinical decision support.

4. CONCLUSION

This study proposed a two-step intelligent framework for gene expression-based cancer classification, integrating the 3SGS method and the SC. The 3SGS approach efficiently reduced dimensionality by filtering irrelevant and redundant genes while retaining the most informative ones, and the SC classifier complemented this by applying simple statistical measures (min, max, mean, and Std) to achieve robust, interpretable, and computationally efficient classification. Experiments on leukemia, prostate, and colon cancer datasets demonstrated the effectiveness of the framework, with high accuracy, minimal gene subsets, and reduced runtime, confirming its potential for reliable early cancer diagnosis. Nonetheless, the framework was tested only on binary-class problems with relatively small sample sizes, and future work should address multiclass classification, larger and more heterogeneous datasets, and integration with clinical metadata and explainability tools such as shapley additive explanations (SHAP) or local interpretable model-agnostic explanations (LIME) to enhance real-world applicability.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sara Haddou Bouazza	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jihad Haddou Bouazza	✓	✓		✓	✓		✓			✓			✓	✓

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This study did not involve individuals nor any personal identification information that could require any informed consent.

ETHICAL APPROVAL

This paper does not involve people or animals; no investigation has involved human subjects. Therefore, the authors did not seek approval from any institutional review board.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] H. Z. Almarzouki, "Deep-learning-based cancer profiles classification using gene expression data profile," *Journal of Healthcare Engineering*, vol. 2022, 2022, doi: 10.1155/2022/4715998.
- [2] G. Geetha and P. Geethanjali, "Pattern classification of bearing faults in PMSM based on time domain feature ensembles," *Engineering Research Express*, vol. 6, no. 3, 2024, doi: 10.1088/2631-8695/ad11f8.
- [3] S. Sharma, T. H. Priya, and V. P. S. Naidu, "Optimizing bearing health condition monitoring: exploring correlation feature selection algorithm," *Engineering Research Express*, vol. 6, no. 2, 2024, doi: 10.1088/2631-8695/ad083d.
- [4] R. Kumar and R. S. Anand, "A methodological integration of fisher score technique with intelligent machine learning methods for ball bearing fault investigation," *Engineering Research Express*, vol. 6, no. 2, 2024, doi: 10.1088/2631-8695/ad0841.
- [5] H. Elwahsh et al., "A new approach for cancer prediction based on deep neural learning," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 6, 2023, doi: 10.1016/j.jksuci.2022.101565.
- [6] S. H. Bouazza, "A deep ensemble gene selection and attention-guided classification framework for robust cancer diagnosis from microarray data," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20235–20241, 2025, doi: 10.48084/etasr.9476.
- [7] D. Tripathi, S. K. Biswas, and B. Baruah, "Data analytics in ensemble learning for effective crop yield prediction," *Engineering Research Express*, vol. 6, no. 3, no. 035237, 2024, doi: 10.1088/2631-8695/ad11fb.
- [8] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, pp. 562, 2023, doi: 10.3390/pr11020562.
- [9] S. H. Bouazza and J. H. Bouazza, "Revolutionizing cancer classification: the SNR-OGSCC method for improved gene selection and clustering," *International Journal of Artificial Intelligence*, vol. 14, no. 1, pp. 466–472, 2023, doi: 10.11591/ijai.v14.i1.pp466-472.
- [10] S. H. Bouazza, "Optimized machine learning for cancer classification via three-stage gene selection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21093–21099, 2025, doi: 10.48084/etasr.9473.
- [11] R. Dash, "An adaptive harmony search approach for gene selection and classification of high dimensional medical data," *Journal of King Saud University – Computer and Information Sciences*, vol. 33, no. 2, pp. 195–207, 2021, doi: 10.1016/j.jksuci.2019.09.004.
- [12] A. Benkessirat and N. Benblidia, "A novel feature selection approach based on constrained eigenvalues optimization," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 4836–4846, 2022, doi: 10.1016/j.jksuci.2021.08.005.
- [13] M. N. KP and P. Thiagarajan, "Feature selection using efficient fusion of fisher score and greedy searching for Alzheimer's classification," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 4993–5006, 2022, doi: 10.1016/j.jksuci.2021.08.006.
- [14] E. Hegazy, M. A. Makhlof, and G. S. El-Tawel, "Improved salp swarm algorithm for feature selection," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, no. 3, pp. 335–344, 2020, doi: 10.1016/j.jksuci.2018.09.001.
- [15] K. A. Uthman, F. M. Ba-Alwi, and S. M. Othman, "A survey on feature selection in microarray data: methods, algorithms and challenges," *International Journal of Computer Science and Engineering*, vol. 8, no. 10, pp. 106–116, 2020, doi: 10.26438/ijcse/v8i10.106116.
- [16] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004, doi: 10.1109/TKDE.2004.68.
- [17] T. R. Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999, doi: 10.1126/science.286.5439.531.
- [18] J. Hou et al., "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, no. 1, pp. 81, 2022, doi: 10.1186/s12859-022-04632-8.
- [19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.
- [20] S. Kwon, H. Lee, and S. Lee, "Image enhancement with Gaussian filtering in time-domain microwave imaging system for breast cancer detection," *Electronics Letters*, vol. 52, no. 5, pp. 342–344, 2016, doi: 10.1049/el.2016.0216.
- [21] P. Dhakar, B. Singh, and P. Gupta, "Comparative performance analysis of different types of k-nearest neighbor (k-NN) classifiers for fault diagnosis of air compressor setup," *Engineering Research Express*, vol. 6, no. 2, 2024, doi: 10.1088/2631-8695/ad0844.
- [22] M. Alwohaibi et al., "A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 5192–5203, 2022, doi: 10.1016/j.jksuci.2021.07.006.
- [23] V. Guleria, V. Kumar, and P. K. Singh, "Classification of surface roughness during turning of forged EN8 steel using vibration signal processing and support vector machine," *Engineering Research Express*, vol. 4, no. 1, 2022, doi: 10.1088/2631-8695/ac5d76.
- [24] Y. E. Almalki et al., "LBP-bilateral based feature fusion for breast cancer diagnosis," *Computer Modeling in Engineering & Sciences*, vol. 73, pp. 4103–4121, 2022, doi: 10.32604/cmc.2022.023429.
- [25] S. H. Bouazza and J. H. Bouazza, "Artificial intelligence application for the classification of central nervous system tumors based on blood biomarkers," in *2024 International Conference on Global Aeronautics Engineering and Satellite Technology (GAST)*, 2024, pp. 1–5, doi: 10.1109/GAST2024.9845632.
- [26] H. S. Pokhariya, D. P. Singh, and R. Prakash, "Evaluation of different machine learning algorithms for LULC classification in heterogeneous landscape by using remote sensing and GIS techniques," *Engineering Research Express*, vol. 5, no. 4, 2023, doi: 10.1088/2631-8695/acfa64.
- [27] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazancı, and O. Oral, "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture," *Journal of King Saud University – Science*, vol. 35, no. 6, 2023, doi: 10.1016/j.jksus.2023.102754.




- [28] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, pp. 271–274, 1998, doi: 10.1023/A:1017181826899.
- [29] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
- [30] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: a comprehensive review," *Expert Systems with Applications*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118946.
- [31] Z. Wang *et al.*, "Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data," *BMC Bioinformatics*, vol. 24, no. 1, 2023, doi: 10.1186/s12859-023-05285-9.
- [32] X. Deng, M. Li, S. Deng, and L. Wang, "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification," *Medical & Biological Engineering & Computing*, vol. 60, no. 3, pp. 663–681, 2022, doi: 10.1007/s11517-021-02420-4.
- [33] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification," *PLoS ONE*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0120364.
- [34] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6, 2005, doi: 10.1186/1471-2105-6-1.

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 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 GénieAppliqué, Marrakech. Currently, he serves as a senior full stack developer and 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.