Advanced Cancer Classification Using AI and Pattern Recognition Techniques

Sara Haddou Bouazza*, *Jihad* Haddou Bouazza

LAMIGEP, EMSI-Marrakech, Morocco

Abstract. Accurate cancer classification is essential for early detection and effective treatment, yet the complexity of gene expression presents significant challenges. In this study, we explored how combining multiple feature selection methods with various classifiers enhances the identification of marker genes for four cancers: leukemia, lung, lymphoma, and ovarian cancer. We applied feature selection techniques such as the F Test, Signal-to-Noise Ratio (SNR), T-test, ReliefF, Correlation Coefficient, Mutual Information, and minimum redundancy maximum relevance, along with classifiers including K-Nearest Neighbors, Support Vector Machines, Linear Discriminant Analysis, Decision Tree Classifiers, and Naive Bayes. Our results demonstrate that the SNR method consistently achieved the highest accuracy in gene selection, particularly when paired with K-means clustering. Remarkably, leukemia was classified with 100% accuracy using only four genes, lung cancer, and lymphoma with 100% and 97% accuracy, respectively, using three genes, and ovarian cancer with 100% accuracy using just one gene. These findings highlight the potential of minimal gene sets for highly precise cancer classification.

1 Introduction

The development of biotechnology has transformed genomics, particularly the capacity to accurately evaluate DNA gene expression [1]. The method starts with collecting RNA from biological tissues and converting it to complementary DNA (cDNA) [2]. Scientists may now analyze thousands of genes at once through technologies like microarray analysis and next-generation sequencing (NGS) [2]. These expression levels are critical for understanding cellular functioning, disease states, and other biological processes [3].

However, evaluating gene expression data presents considerable hurdles due to the nature of the data generated. This produces an expression matrix including a large number of genes (attributes) assessed in a small number of samples (patients) [4].

Addressing these challenges requires advanced data mining and AI techniques. Feature selection methods are particularly effective in identifying key factors that differentiate between health conditions or diseases. By focusing on a few important genes, scientists can enhance model accuracy and clarity, leading to more reliable and actionable biological insights.

2 Material and methods

This section provides a detailed overview of the materials and methods used in our research, emphasizing the importance of cancer research and the advanced techniques employed to enhance classification accuracy.

2.1 The Importance of Cancer Research

Cancer remains one of the leading causes of death worldwide, with various types posing significant health risks, such as leukemia, lung cancer, lymphoma, and ovarian cancer [5, 6, 7, 8] (Table 1). Each cancer type has unique genetic and molecular characteristics, underscoring the necessity of understanding gene expression patterns specific to each type. [9].

Table 1. Cancer details

Datasets	Number of genes	Number of samples
Leukemia	7129	72 (47 ALL, 25 AML)
Lung Cancer	12533	181 (31 MPM, 150 ADCA)
Lymphoma Cancer	7070	77 (58 DLBCL, 19 FL)
Ovarian Cancer	15154	253 (91 normal, 162 with cancer)

2.2 Feature Selection Methods in Cancer Classification

This section presents the various feature selection methods employed in cancer classification, highlighting their importance and application.

In cancer classification, selecting the right features is critical due to the complexity of gene and protein data [10]. Feature selection methods aim to identify the most

* Corresponding author: sara.hb.sara@gmail.com

informative features, enhancing model performance, interpretability, and generalization to new data [11, 12]. F Test - The F test (Eq.1) checks if the average values of two groups are significantly different [13].

$$F(x) = \frac{(\overline{x_1} - \overline{x_2})^2}{(s_1^2 + s_2^2)} \quad (1)$$

Where $\overline{x_y}$ and s_y^2 are the mean and standard deviation of the attribute for class y = 1, 2.

- Signal-to-Noise Ratio (SNR) - SNR compares the difference in average values of a feature between two types of data to the variation within each type (Eq.2) [14].

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

Where $\overline{x_{yJ}}$ is the mean of attribute j and s_{yj} is its standard deviation for classes y = 1, 2.

- T-Test - Similar to the F test, the T test checks if the average values of a feature are noticeably different between two groups (Eq.3) [15].

$$t(x) = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad (3)$$

Where n_y is the size, $\overline{x_y}$ is the mean and s^2_y is the variance of classes y = 1, 2.

- ReliefF - ReliefF rates features by how well they differentiate between similar instances (Eq.4). ReliefF handles noisy data and multiple classes but requires significant computing power [16].

$$Wd = wd - \sum_{j=1}^{K} \frac{diff(x_i, d_i, hits_j)}{m*y} + \sum_{c \neq class(x_i)} \frac{(p(y))}{1 - p(class(x_i))} \sum_{j=1}^{k} \frac{diff(x_i, d_i, misses_j)}{m*y} \quad (4)$$

$$diff(x_i, d_i, x_j) = \frac{|x_{id} - x_{jd}|}{max(d) - min(d)}$$

Max(d) (resp. min(d)) designates the maximum (resp. minimum) value that the characteristic designated by the index d can take on all of the data. x_{id} is the value of the d_{th} characteristic of data xi.

- Correlation Coefficient (CC) - The CC method selects features based on their correlation with class labels (Eq.5) [17].

$$t = \frac{Cov(X,Y)}{\sigma_X + \sigma_Y} \quad (5)$$

Cov (X,Y) is the covariance between X and Y and σ is the standard deviation of X and Y.

- Mutual Information (MI) - MI measures the amount of information one variable provides about another (Eq.6) [18].

$$IM(Y, X) = H(Y) - H(Y / X) \quad (6)$$

Where H(Y/X) measures the amount of information contained in attribute X when class Y is provided.

- Minimum Redundancy Maximum Relevance (mrMR) - mrMR selects features that are highly relevant to class labels and minimally redundant to each other (Eq.7). MrMR balances relevance and redundancy but can be computationally intensive with large datasets [19].

$$Redondance\ (i) = \frac{1}{|F|^2} \sum_{i,j \in F} I(i,j)$$

$$Pertinence\ (i) = \frac{1}{|F|^2} \sum_{i,j \in F} I(i,Y)$$

$$Score(i) = \frac{Pertinence\ (i)}{Redondance\ (i)} \quad (7)$$

With F and |F| represent, respectively, the set of attributes and its size, I (i,j) is the mutual information between the i^{th} and the j^{th} attribute, I (i,Y) is the mutual information between the i^{th} characteristic and the class vector (Y).

2.3 Clustering Combined with Filter Methods

In addition to individual feature selection techniques, combining clustering with filter methods enhances the selection process. Clustering groups similar features, allowing the selection of key features from each group. This approach identifies feature sets that best represent the main data patterns

K-Means Clustering: K-means clustering divides features into groups based on similarity. By grouping similar features, k-means reduces redundancy and simplifies the selection of representative features. Features within the same group provide similar information [20].

Filter Techniques (F Test, SNR, T-Test, ReliefF, CC, MI, mrMR): Each filter technique offers a unique evaluation of feature importance. After clustering, these techniques can be applied within each cluster to select the most important features. This ensures that selected features are both significant and diverse. This combined approach improves model performance, and robustness to new data, and provides deeper insights into the biological significance of selected features.

2.4 Cancer Classification Methods

Identifying different types of cancer from medical information is crucial for accurate diagnosis and appropriate treatment decisions. Various machine-learning algorithms have been effectively employed to classify cancer types using image data, genetic information, or patient records. Here, we discuss key techniques used in cancer classification:

- K-Nearest Neighbors (KNN): KNN is a simple and effective classification method. It assigns a new case to the group most common among its nearest neighbors in feature space [21].

- Support Vector Machines (SVM): SVM is a powerful tool for classification and regression. It finds the optimal hyperplane in a high-dimensional space that separates different classes [22].

- Linear Discriminant Analysis (LDA): LDA is a statistical technique for dimensionality reduction and classification. It combines features to maximize the separation between two or more classes [23].

- Decision Trees (DTC): are non-parametric methods for classification and regression. They iteratively split data based on feature values to create homogenous groups [24].

- Naive Bayes (NB): Naive Bayes classifiers are probabilistic models assuming feature independence. Despite the simplicity of this assumption [25].

2.5 Classification accuracy

We measure the classification accuracy to test different methods and look at how accurate they are to find the best approach for cancer classification (Eq.8) [26].

$$A_{ccuracy} = 100\,{(T_P + T_N)}/{(T_N + T_N + F_N + F_P)} \quad (8)$$

TP: Positive samples correctly recognized. TN: Negative samples correctly recognized. FP: Negative samples wrongly identified as positive. FN: Positive samples wrongly identified as negative.

3 Results

This section summarizes the findings of our study, which investigated the application of various classifiers combined with feature selection methods for cancer classification. The performance was evaluated based on accuracy and the number of genes used. Experiments were conducted on a standard laptop with an Intel® Core™ i5 CPU M 250 @ 2.4 GHz, 4 GB RAM, running Windows 10 (64-bit), using MATLAB R2023a for methodology implementation, data analysis, and classification tasks.

3.1 Application of Filter Selection + Classification Approach

In this subsection, we analyze the performance of several classifiers—KNN, SVM, LDA, DTC, and NB—when combined with different filter-based feature selection methods. These methods include F-test, SNR, T-test, ReliefF, CC, IM, and mrMR. The study spans various types of cancer, including leukemia, lung cancer, lymphoma cancer, and ovarian cancer.

The performance metrics for each classifier and feature selection method combination are summarized in Table 2. The table displays the highest accuracy (Max Acc %) achieved and the corresponding number of genes used (Nbr genes) for each type of cancer.

- Leukemia: The SNR and CC methods got the best accuracy (100%) when used with KNN, LDA, and NB. SNR used the fewest number of genes to get high accuracy, especially with LDA, where it only used four genes.

- Lung Cancer: SNR, T-test, ReliefF, and CC reached 97.3% accuracy using different types of classifiers. However, SNR reached the highest accuracy using only six genes for the KNN method.

- Lymphoma Cancer: SNR, ReliefF and CC methods had the best accuracies (95.6%) with different classifiers. The SNR method was very accurate and used fewer genes, especially for KNN, where it used only four genes.

- Ovarian Cancer: The SNR, ReliefF and CC methods had the best accuracy at 97.5% when used with KNN, SVM, and LDA. The SNR method was very effective and needed only 30 genes to achieve high accuracy with KNN.

Table 2. Comparing Filter-Based Feature Selection and Supervised Classification Techniques for cancer databases

		KNN		SVM		LDA		DTC		NB	
		Max Acc	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*
Leukemia	*F test*	97.05	69	84.2	59	97.05	93	58.82	8	73.52	2
	SNR	100	13	97.05	4	100	9	97.05	3	97.05	5
	T-test	97.05	75	97.05	2	97.05	66	91.17	13	58.82	95
	ReliefF	97.05	41	97.05	2	97.05	69	94.11	11	94.11	5
	CC	100	50	97.05	3	100	93	97.05	4	97.05	6
	IM	76.41	56	84.2	5	91.1	10	76.4	86	91.1	28
	mrMR	97.05	11	88.2	30	85.2	40	64.7	33	88.2	12
Lung Cancer	*F test*	83.2	82	88.5	6	67.7	53	66.4	3	84.5	5
	SNR	97.3	6	97.3	33	97.3	64	97.3	10	96.6	19
	T-test	96.6	13	97.3	17	97.3	17	90.6	11	96.6	4
	ReliefF	97.3	21	96,6	9	97.3	80	97.3	25	90.6	21
	CC	97.3	28	96,6	36	97.3	82	90.6	31	90.6	29
	IM	83.2	*10*	88.5	*5*	96.6	*24*	83.2	*7*	66.4	*19*
	MrMR	90.6	62	88.5	18	83.5	23	90.6	5	90.6	25
Lymphoma Cancer	*F test*	86.9	1	78.2	29	78.2	1	82.6	1	82.6	4
	SNR	95.6	4	95.6	32	95.6	24	91.3	3	91.3	3
	T-test	91.3	3	86.9	13	91.3	4	86.9	13	82.6	3
	ReliefF	95.6	86	91.3	20	95.6	93	91.3	12	86.9	2
	CC	95.6	13	91.30	39	95.6	97	91.3	8	91.3	55
	IM	86.9	10	86.9	15	52.1	50	91.3	13	91.3	29
	MrMR	86.9	15	86.9	10	91.3	5	86.9	10	82.6	13
Ovarian Cancer	*F test*	71.8	54	67.5	89	64.3	92	64.3	34	63.1	3
	SNR	97.5	30	97.5	39	96.8	37	96.8	35	91.1	21
	T-test	97.5	3199	96.8	12	84.2	19	91.1	25	91.1	41
	ReliefF	97.5	32	96.8	36	96.8	31	97.5	31	85.2	22
	CC	97.5	76	96.8	46	97.5	40	97.5	35	91.1	11
	IM	76.41	16	84.2	15	91.1	12	76.4	26	91.1	18
	MrMR	91.1	11	88.2	3	85.2	4	64.7	33	88.2	2

3.2 Application of Clustering + Filter Selection + Classification Approach

In this subsection, we investigate how integrating K-means clustering with filter selection methods improves classification accuracy across different cancer types. By incorporating K-means into the feature selection process, we aim to enhance accuracy and reduce gene selection variability. Table 3 displays the results of combining K-means with various feature selection methods and classifiers, highlighting the maximum accuracy achieved and the corresponding number of

selected genes. The number of clusters is k=3 for leukemia, k=4 for lung cancer, and k=3 for both lymphoma and ovarian cancers.

Table 3. Enhancing Classifier Accuracy with K-means Clustering

		KNN		SVM		LDA		DTC		NB	
		Max Acc	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*	*Max Acc*	*Nbr genes*
Leukemia	*K-means + F test*	97.05	6	97.05	5	97.05	13	91.1	6	91.1	12
	K-means + SNR	100	4	100	7	100	15	97.05	2	97.05	10
	K-means + T-test	100	12	100	11	97.05	12	97.05	13	91.1	35
	K-means + ReliefF	100	8	100	15	100	21	97.05	16	97.05	13
	K-means + CC	100	19	100	12	100	35	97.05	2	97.05	5
	K-means + IM	91.1	18	91.1	5	94.1	5	94.1	28	94.1	15
	K-means + mrMR	97.05	5	91.1	16	94.1	20	91.1	23	91.1	8
Lung Cancer	*K-means + F test*	90.6	12	90.6	15	88.5	28	83.2	12	90.6	18
	K-means + SNR	100	3	99.3	10	99.3	14	97.3	2	97.3	10
	K-means + T-test	97.3	11	97.3	7	99.3	12	96.6	5	96.6	12
	K-means + ReliefF	99.3	4	97.3	11	99.3	28	97.3	12	96.6	11
	K-means + CC	99.3	5	99.3	12	97.3	19	96.6	17	96.6	15
	K-means + IM	96.6	9	90.6	5	97.3	20	90.6	13	66.4	10
	K-means + mrMR	90.6	11	90.6	12	88.5	13	96.6	15	96.6	21
Lymphoma Cancer	*K-means + F test*	91.3	21	86.9	32	86.9	12	91.3	14	91.3	15
	K-means + SNR	97	3	97	10	97	12	95.6	12	95.6	22
	K-means + T-test	95.6	13	91.3	3	91.3	2	91.3	10	86.9	13
	K-means + ReliefF	97	12	95.6	10	97	17	95.6	2	91.3	13
	K-means + CC	97	8	95.6	4	97	22	95.6	1	91.3	12
	K-means + IM	91.3	7	95.6	7	91.3	4	91.3	2	91.3	3
	K-means + mrMR	91.3	13	91.3	23	95.6	16	91.3	3	91.3	7
Ovarian Cancer	*K-means + F test*	88.2	14	71.8	12	88.2	16	67.5	4	71.8	10
	K-means + SNR	99.3	5	100	4	97,5	11	97,5	4	98.1	9
	K-means + T-test	97,5	11	97,5	7	91.1	19	96.8	5	97.5	11
	K-means + ReliefF	99.3	12	99.3	16	98.1	31	98.7	1	96.8	2
	K-means + CC	99.3	6	99.3	13	98.7	22	98.1	4	98.7	12
	K-means + IM	91.1	16	91.1	15	91.1	12	88.2	6	96.8	8
	K-means + mrMR	96.8	11	91.1	3	91.1	4	88.2	13	91.1	2

- In a Leukemia cancer study, the K-means algorithm combined with SNR achieved perfect accuracy (100%) for KNN, SVM, and LDA while using a small number of features (4, 7, and 15 respectively). Additionally, K-means paired with T-test, ReliefF, and CC demonstrated high accuracy, reaching 100% for some classifiers. However, these methods typically require more features compared to SNR.

- For Lung Cancer classification, Using K-means clustering combined with SNR resulted in a perfect accuracy (100%) when paired with KNN, and 99.3% with SVM, and LDA algorithms, requiring only a few attributes (3, 10, and 14, respectively). The combinations of K-means with ReliefF and K-means with CC also performed exceptionally well, achieving up to 99.3% accuracy, though they generally required more attributes compared to SNR.

- When dealing with Lymphoma Cancer, using K-means combined with SNR or ReliefF achieved a high accuracy rate of 97% when paired with KNN, SVM, and LDA classifiers, requiring only a few attributes (3 to 12). K-means combined with the CC also performed well, but it needed more attributes (8 to 22) to reach the same level of accuracy.

- For classifying Ovarian Cancer, the K-means step significantly enhances the results. Specifically, using K-means with SNR reaches a perfect accuracy. Both ReliefF and CC methods worked exceptionally well, reaching an accuracy of 99.3% with KNN and SVM. These methods required several features, ranging from 4 to 16.

Incorporating K-means clustering greatly boosted the accuracy of cancer classification and helped identify key genes:

- Leukemia: 100% accuracy with KNN using 4 genes selected by K-means + SNR: M23197, M27891, AJ000480, and L09209_s_at.

- Lung: 100% accuracy with KNN using 3 genes selected by K-means + SNR: X93036, D87436, and K03195.

- Lymphoma: 97% accuracy with KNN using 3 genes: Z21966_at, D87119_at, and U28386_at.

- Ovarian: 100% accuracy with SVM using 4 genes: MZ49.784115, MZ3546.2884, MZ4362.0866, and MZ9159.3641.

3.3 Comparative Analysis of Classifier Accuracy Enhancement Techniques

In our study on leukemia cancer classification, we employed K-means clustering combined with various feature selection techniques, achieving notable accuracy improvements across classifiers. Specifically, the combination of K-means and SNR achieved a perfect 100% accuracy with KNN, SVM, and LDA, while DTC and NB reached 97.05%. In comparison, Mallick et al. (2023) [27] achieved 98.2% accuracy using artificial neural networks (ANN) with feature selection, while Nirmalakumari et al. (2023) [28] and Susmi et al. (2016) [29] both achieved 98.5% using genetic algorithms with SVM. Our results demonstrate either higher or comparable accuracy, particularly with K-means and SNR, highlighting the effectiveness of our hybrid approach.

In our study on lung cancer classification, the combination of K-means and SNR achieved near-perfect accuracy, with 100% using KNN and 99.3% using SVM and LDA. Other combinations also demonstrated high accuracy rates. In comparison, Fathi et al. (2021) [30] reported 95% accuracy using a hybrid approach with PCC and DT, while Olaniran and Abdullah (2020) [31] achieved 97.98% with a hybrid VB method. Astuti et al. (2021) [32] achieved 100% accuracy with both SVM and RF. Our results,

particularly with K-means and SNR, align with these high accuracies, demonstrating the robustness of our approach.

Lymphoma Cancer: Our study found that the combination of K-means and SNR achieved the highest accuracy of 97% with KNN and LDA, while other classifiers also performed well with accuracies around 95.6%. Comparatively, Olaniran and Abdullah (2020) [31] reported an accuracy of 94.92% using a hybrid VB approach. Rezaee et al. (2022) [33] achieved 97% accuracy using a deep neural network and ensemble gene selection. Nirmalakumari et al. (2023) [34] reported an accuracy of 94.95% with the Elephant Search classifier. Our results are competitive, particularly with K-means and SNR, indicating the efficacy of our hybrid methods.

Ovarian Cancer: In our study, the combination of K-means and Signal-to-Noise Ratio (SNR) achieved the highest accuracy of 100% with SVM and 99.3% with KNN. Other combinations also showed high accuracy rates, with K-means and ReliefF achieving 99.3% accuracy with both KNN and SVM. In comparison, Prabhakar and Lee (2020) [35] achieved a high classification accuracy of 99.48% using SVM-RBF with GBCO. Al-Murad and Hossain (2021) [36] reported an accuracy of 99.20% using ANN with ENORA and CSE for feature selection. Akmal and Fitrianah (2024) [37] achieved 100% accuracy using ABC for feature selection. Our results, particularly with K-means and SNR, demonstrate competitive accuracy, highlighting the effectiveness of our approach.

4 Conclusion

This study highlights the transformative power of selecting the right features to build accurate and reliable cancer classification models. By focusing on four different types of cancer, we discovered that the SNR method, especially when combined with k-means clustering, consistently provided the highest accuracy and the most relevant gene markers. For example, in leukemia, this approach achieved a perfect 100% accuracy by selecting just 13 critical genes.

The optimal results for each cancer type using K-means clustering and the corresponding number of selected genes are as follows: leukemia can be classified with 100% accuracy using just four genes; lung cancer also achieves 100% accuracy with three genes, while lymphoma cancer reaches 97% accuracy using the same number of genes. Ovarian cancer achieves perfect classification accuracy of 100% with only 15 genes.

References

1. Elden, R. H., Ghonim, V. F., Hadhoud, M. M., & Al-Atabany, W. (2023). Transcriptomic marker screening for evaluating the mortality rate of pediatric sepsis based on Henry gas solubility optimization. Alexandria Engineering Journal, 68, 693-707.
2. Elloumi, M., Ahmad, M. A., Samak, A. H., Al-Sharafi, A. M., Kihara, D., & Taloba, A. I. (2022). Error correction algorithms in non-null aspheric testing next generation sequencing data. Alexandria Engineering Journal, 61(12), 9819-9829.
3. Akgül, A., Khoshnaw, S. H., & Rasool, H. M. (2020). Minimizing cell signalling pathway elements using lumping parameters. Alexandria Engineering Journal, 59(4), 2161-2169.
4. Esmaeili, Y., Bidram, E., Bigham, A., Atari, M., Azadani, R. N., Tavakoli, M., ... & Zarrabi, A. (2023). Exploring the evolution of tissue engineering strategies over the past decade: From cell-based strategies to gene-activated matrix. Alexandria Engineering Journal, 81, 137-169.
5. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.
6. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, RichardsWG, Sugarbaker DJ, Bueno R: Translation of microarray data into clinically relevant can-cer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. CancerRes. 2002, 62: 4963-4967
7. M. A. Shipp, K. N. Ross, P. Tamayo 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, 2002
8. E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn,and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359:572–577, 2002
9. Alshmrani, G. M. M., Ni, Q., Jiang, R., Pervaiz, H., & Elshennawy, N. M. (2023). A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. Alexandria Engineering Journal, 64, 923-935.
10. Liao, X., Li, K., Gan, Z., Pu, Y., Qian, G., & Zheng, X. (2024). Prognostic prediction of ovarian cancer based on hierarchical sampling & fine-grained recognition convolution neural network. Alexandria Engineering Journal, 102, 264-278.
11. Alzahrani, A. S., Shah, R. A., Qian, Y., & Ali, M. (2020). A novel method for feature learning and network intrusion classification. Alexandria Engineering Journal, 59(3), 1159-1169.
12. Althobaiti, T., Althobaiti, S., & Selim, M. M. (2024). An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making. Alexandria Engineering Journal, 94, 311-324.

13. Faris, M., Mahmud, M. N., Salleh, M. F. M., & Alsharaa, B. (2023). A differential evolution-based algorithm with maturity extension for feature selection in intrusion detection system. Alexandria Engineering Journal, 81, 178-192.
14. Mahalakshmi, D. M., & Sumathi, S. (2019). Brain tumour segmentation strategies utilizing mean shift clustering and content based active contour segmentation. ICTACT J. Image Video Process, 9(4), 2002-2008.
15. Chandran, V., & Mohapatra, P. (2023). Enhanced opposition-based grey wolf optimizer for global optimization and engineering design problems. Alexandria Engineering Journal, 76, 429-467.
16. Gavisiddappa, G., Mahadevappa, S., & Patil, C. (2020). Multimodal biometric authentication system using modified ReliefF feature selection and multi support vector machine. International Journal of Intelligent Engineering and Systems, 13(1), 1-12.
17. Rahadian, H., Bandong, S., Widyotriatmo, A., & Joelianto, E. (2023). Image encoding selection based on Pearson correlation coefficient for time series anomaly detection. Alexandria Engineering Journal, 82, 304-322.
18. Shaheen, M., Naheed, N., & Ahsan, A. (2023). Relevance-diversity algorithm for feature selection and modified Bayes for prediction. Alexandria Engineering Journal, 66, 329-342.
19. Aljawarneh, M., Hamdaoui, R., Zouinkhi, A., Alangari, S., & Abdelkrim, M. N. (2024). Energy optimization for wireless sensor network using minimum redundancy maximum relevance feature selection and classification techniques. PeerJ Computer Science, 10, e1997.
20. Abo-Elnaga, Y., & Nasr, S. (2022). K-means cluster interactive algorithm-based evolutionary approach for solving bilevel multi-objective programming problems. Alexandria Engineering Journal, 61(1), 811-827.
21. Saroğlu, H. E., Shayea, I., Saoud, B., Azmi, M. H., El-Saleh, A. A., Saad, S. A., & Alnakhli, M. (2024). Machine learning, IoT and 5G technologies for breast cancer studies: A review. Alexandria Engineering Journal, 89, 210-223.
22. Roshani, M., Phan, G. T., Ali, P. J. M., Roshani, G. H., Hanus, R., Duong, T., ... & Kalmoun, E. M. (2021). Evaluation of flow pattern recognition and void fraction measurement in two phase flow independent of oil pipeline's scale layer thickness. Alexandria Engineering Journal, 60(1), 1955-1966.
23. Wijaya, I. G. P. S., Widiartha, I. B. K., Bimantoro, F., & Septiadi, A. (2019). Buildings cracks classification using zoning and invariant moment features and quadratic discriminant analysis classifier. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 158-168.
24. Omer, N., Samak, A. H., Taloba, A. I., & Abd El-Aziz, R. M. (2023). A novel optimized probabilistic neural network approach for intrusion detection and categorization. Alexandria Engineering Journal, 72, 351-361.
25. Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401-3409.
26. Afify, H. A. (2011). Evaluation of change detection techniques for monitoring land-cover changes: A case study in new Burg El-Arab area. Alexandria engineering journal, 50(2), 187-195.
27. Mallick, P. K., Mohapatra, S. K., Chae, G. S., & Mohanty, M. N. (2023). Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing, 27(3), 1103-1110.
28. Nirmalakumari, K., Rajaguru, H., & Rajkumar, P. (2023, April). Leukemia cancer classification using extrusive genes from microarray data. In AIP Conference Proceedings (Vol. 2725, No. 1). AIP Publishing.
29. Jacophine Susmi, S., Khanna Nehemiah, H., Kannan, A., & Christopher, J. (2016). Relevant Gene Selection and Classification of Leukemia Gene Expression Data. In Emerging Research in Computing, Information, Communication and Applications: ERCICA 2015, Volume 3 (pp. 503-510). Springer Singapore.
30. Fathi, H., AlSalman, H., Gumaei, A., Manhrawy, I. I., Hussien, A. G., & El-Kafrawy, P. (2021). Research Article An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data.
31. Olaniran, O. R., & Abdullah, M. A. A. (2020, March). Subset selection in high-dimensional genomic data using hybrid variational Bayes and bootstrap priors. In Journal of Physics: Conference Series (Vol. 1489, No. 1, p. 012030). IOP Publishing.
32. Astuti, W. (2021, August). Comparative analysis of support vector machine (SVM) and random forest (RF) classification for cancer detection using microarray. In 2021 9th International Conference on Information and Communication Technology (ICoICT) (pp. 650-656). IEEE.
33. Rezaee, K., Jeon, G., Khosravi, M. R., Attar, H. H., & Sabzevari, A. (2022). Deep learning-based microarray cancer classification and ensemble gene selection approach. IET Systems Biology, 16(3-4), 120-131.