Cancer Classification Using Pattern Recognition and Computer Vision Techniques

Sara Haddou Bouazza*, *Jihad* Haddou Bouazza

LAMIGEP, EMSI-Marrakech, Morocco

Abstract. The rapid advancement of DNA microarray technology has significantly contributed to the classification of various cancers, particularly leukemia. However, the high-dimensional nature of gene expression data presents challenges such as data noise and irrelevant features, leading to reduced prediction accuracy. This study proposes a novel Hybrid Filter-Wrapper Gene Selection (HFWGS) method that integrates filter-based techniques (Signal-to-Noise Ratio, Correlation Coefficient, and ReliefF) with wrapper-based approaches to enhance feature selection for leukemia classification. Additionally, a Hybrid Statistical-Gene Voting (HSGV) approach was implemented to further refine classification accuracy. A comparative analysis of classifiers, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Linear Discriminant Analysis (LDA), demonstrated that the HFWGS method consistently improved classification performance, achieving 100% accuracy with a reduced subset of genes. The proposed methods provide an efficient framework for optimizing gene selection and improving diagnostic accuracy in leukemia, paving the way for more targeted therapeutic interventions.

1 Introduction

DNA microarray technology holds significant promise for cancer classification and prognosis prediction by analyzing expression levels of thousands of genes, providing insights into gene interactions [1, 2]. It aids in diagnosing diseases and exploring genetic factors, enhancing our understanding of the genetic causes of various anomalies [3, 4].

However, a key challenge is the imbalance between sample sizes and the large number of genes, as not all genes are relevant to cancer, and many have indirect interactions. Using the full gene set can reduce prediction accuracy [5]. Feature selection and classification are critical, interlinked challenges in microarray gene analysis. Effective feature selection isolates significant genes, enhancing classification accuracy through various statistical and computational approaches, including data mining and machine learning [8, 9, 10].

Gene selection methods include filter and wrapper approaches. Filter methods independently assess gene relevance using heuristic scoring and metrics like p-values, Signal-to-Noise Ratio (SNR), maximum relevance minimum redundancy, ReliefF, and PUL scores [11, 12, 13]. However, they may overlook gene interactions and struggle with optimal subset size. Wrapper methods evaluate gene utility with classifiers [14], while hybrid methods combine filters for pre-selection and wrappers for refining the final gene subset, improving accuracy.

2 Materials and methods

This study aims to enhance gene selection for more accurate tumor classification. We applied advanced techniques to identify relevant genes from a complex microarray dataset, with the procedures outlined below

2.1 Gene Selection Methods

Gene selection is vital in bioinformatics for identifying key features in high-dimensional datasets, improving classification accuracy by reducing noise. Methods like Signal-to-Noise Ratio (SNR), Correlation Coefficient (CC), and ReliefF assess gene-class relationships. However, effective selection must also consider noise. We propose a hybrid method that integrates filter-based and wrapper-based approaches to retain the most informative genes.

Signal to Noise Ratio (SNR): SNR measures each gene's discriminative power by evaluating the ratio of mean class difference to within-class variability [15, 16]. SNR for a feature G is computed as:

$$SNR = \frac{\mu_{class1} - \mu_{class2}}{\sigma_{class1} + \sigma_{class2}} \quad (1)$$

Where μ_{class1} and μ_{class2} are mean values of the feature in the two classes, and σ_{class1}, σ_{class2} are standard deviations within each class.

Correlation Coefficient (CC): Pearson's correlation coefficient measures the linear relationship between a

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

feature and target [17, 18]. For feature X and target Y, the CC formula is:

$$CC = \frac{n(\sum XY)-(\sum X)(\sum Y)}{\sqrt{[n\sum X^2-(\sum X)^2][n\sum Y^2-(\sum Y)^2]}} \quad (2)$$

Where n represents the sample size.

ReliefF: ReliefF is an algorithm that selects features by estimating feature importance based on how well features differentiate between instances near each other [19, 20]. ReliefF identifies its nearest neighbor from the same class (nearest hit) and a different class (nearest miss) for each instance in the dataset. Features that differentiate between classes (i.e., high difference between miss and hit) are given higher importance. The ReliefF Training Process is based on three steps:

1. Initialize: Start with a weight vector initialized to zero for all features.
2. Iterate over instances: For each instance I:
   - Find the nearest instance from the same class
   - Find the nearest instance from a different class
   - Update the weight vector for each feature f:

$$W(f) = W(f) - diff\ (I, hit, f) + diff\ (I, miss, f) \quad (3)$$

Where diff (I, J, f) is the difference between instances I and J on feature f.

3. Rank Features: Features are ranked based on their weight vector. Features with higher weights are considered more important.

Hybrid Filter-Wrapper Gene Selection with Advanced Techniques (HFWGS): In the filter-based selection approach, genes are ranked based on their relevance, and a subset of highly ranked genes (Subset 1) is selected. However, certain noisy genes, while seemingly relevant, may negatively affect classification accuracy when combined with other genes.

To address this issue, a wrapper selection strategy is applied. Starting with the most relevant gene from Subset 1, genes are added one by one, and classification accuracy is measured after each addition. If adding a gene improves accuracy, it is retained; otherwise, it is discarded. This step creates Subset 2, which contains only genes that contribute positively to accuracy.

In the final step, a minimal subset of genes is selected by removing genes from Subset 2 and checking if removing a gene maintains or improves accuracy. The final subset contains only the most essential genes for classification.

The HFWGS proposed architecture has an input of a set of features (genes) and their corresponding class labels. And as Output, A minimal subset of features that maximizes classification accuracy.

Algorithm Steps:

1. Filter Selection (Step 1):
   - Ensemble Ranking: Use filter method
   - Select Subset 1: Choose the top 100 genes from the ranked list based on the ensemble score.
2. Wrapper Selection (Step 2):
   - Initialize: Start with the gene having the highest score from Subset 1.
   - Greedy Forward-Backward Search:
     - Add Genes: Add one gene at a time from Subset 1 and measure classification accuracy using cross-validation (5-fold).
     - Backward Elimination: After adding each new gene, attempt to remove previously added genes to improve performance.
     - Stopping Criteria:
       - Stop adding genes when the improvement in accuracy is less than 0.01%.
       - Limit the number of genes in Subset 2 to a maximum of 50 genes.
3. Final Minimal Subset Selection (Step 3):
   - Refinement: For each gene in Subset 2, test whether removing the gene impacts accuracy:
     - If accuracy decreases, retain the gene.
     - If accuracy remains the same or improves, remove the gene.
   - Stopping Criteria:
     - Stop removing genes if the subset size drops below 10 genes, or if removing any more decreases accuracy.
4. Output Subset: Return the minimal subset of genes that maximizes classification accuracy.

2.2 Classification methods

In the realm of leukemia classification, where accurate differentiation between cancerous and non-cancerous instances is crucial, various machine learning algorithms have proven effective in achieving binary classification. Classification accuracy was considered the key evaluation metric for the classifiers.

K Nearest Neighbors (KNN): KNN is an instance-based learning algorithm that classifies instances based on the majority vote of their K-nearest neighbors [22].

The KNN Hyperparameters are:

- k: Number of neighbors k=3
- Distance Metric: Euclidean distance
- Weights: 'distance'

Support Vector Machine (SVM): SVM is a powerful classifier that seeks to find the optimal hyperplane that maximizes the margin between different classes [23]. The training process involves solving a convex optimization problem to determine this hyperplane. The SVM Hyperparameters are:

- Regularization Parameter (C): C=1
- Kernel: Linear
- Gamma: 'scale'

Linear Discriminant Analysis (LDA): LDA projects data into a lower-dimensional space to maximize class separability [24]. The training process involves calculating class means and within-class scatter to maximize the between-class scatter.

Decision Tree for Classification (DTC): DTC is a tree-based model where each internal node represents a decision based on a specific feature, and each leaf node corresponds to a class label [25]. The model recursively splits the dataset based on feature values to maximize information gain, typically using the Gini impurity measure.

The DTC Hyperparameters are:

- Max Depth: 10
- Min Samples Split: 5
- Criterion: 'gini'

Naïve Bayes (NB): NB is a probabilistic classifier grounded in Bayes' theorem, which assumes that features are independent [26]. The training process involves computing the probabilities of classes and the likelihood of features, followed by classification based on posterior probabilities.

The classification Accuracy formula is as follows [27]:

$$Acc = 100 * \frac{\mathrm{TP} + \mathrm{TN}}{TN + TP + FN + FP} \quad (4)$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Our novel Hybrid Statistical-Gene Voting approach operates in two steps: first, it employs statistical measures of selected genes, followed by a voting mechanism to classify samples.

Hybrid Statistical-Gene Voting Approach (HSGV): The architecture of HSGV takes gene expression profiles from training samples as input, where each gene has multiple expression levels across two distinct classes. The output is the classification of test samples into one of the two classes (class 1 or class 2) using a voting mechanism. The classification process consists of the following key steps:

- Statistical Value Calculation: Compute statistical measures for selected genes across both classes.
- Voting-Based Classification: Classify samples based on gene vote scores, with the final classification decision made through a majority vote.

The Hybrid Statistical-Gene Voting Approach involves three steps:

1. Step 1: Statistical Measure Calculation
   For each selected gene, compute the following statistical measures in the training samples for both classes:
   - Min: Minimum expression value.
   - Max: Maximum expression value.
   - Mean: Mean expression value.
   - Std: Standard deviation of expression values.
2. Step 2: Voting-Based Classification Using Enhanced Techniques
   - Gene Classification: For each test sample, compare the gene expression values to the statistical intervals (Mean ± Std) calculated for both classes.
     - If the expression value does not lie within the defined interval for either class, classify the sample into the nearest class.
   - Weighted Voting Schema: After all genes are classified, implement a weighted voting mechanism where each gene's vote is weighted by its accuracy or reliability. Genes with higher classification accuracy during training have a greater impact on the final vote.

2.3 Microarray Data Overview

The microarray data in this study is structured as an N×M matrix, where N represents the number of samples and M denotes the number of genes. Each entry corresponds to the expression level of a specific gene in a given sample.

For the Leukemia dataset, 7,129 genes are analyzed across 72 samples, which include both acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [28]. The gene expression levels are vital for distinguishing between these two types of leukemia, making the data critical for our classification efforts.

This dataset supports the application of HSGV, facilitating an in-depth analysis of gene expression patterns. We leverage this data to employ statistical measures and voting mechanisms for accurate classification. Integrating the HFWGS method ensures the selection of only the most informative genes, thereby improving the model's accuracy and reliability.

3 Experimental results

This section presents the obtained results of leukemia classification. Experiments were performed in a MATLAB environment on a laptop with an Intel(R) Core(TM) i5 CPU M 250 @ 2.4GHz processor. The running time of our applications varied between 30 seconds and one minute.

3.1 Results corresponding to the new HFWGS Approach

In this study, we evaluated different classifiers with various feature selection methods on leukemia data. The results indicate that the HFWGS Approach (FA) enhances classification accuracy while minimizing the number of selected genes.

Table 1 shows the classification accuracies for each classifier with different feature selection strategies. The SNR and CC methods achieved accuracies between 100% and 97% for 3 to 93 genes, with accuracy reaching 100% using just three to five genes when combined with HFWGS. The ReliefF method yielded accuracies from 94% to 97% for 2 to 69 genes, which could be improved to 100% with HFWGS using only four genes.

Table 1: Performance of Various Classifiers with Different Feature Selection Methods For Leukemia

		Classifiers				
		KNN	**SVM**	**LDA**	**DT**	**NB**
Feature Selection Methods	SNR	100% (13)	97% (4)	97% (9)	97% (3)	97% (5)
	SNR_ HFWGS	100% (3)	97% (2)	97% (4)	97% (3)	97% (4)
	CC	100% (50)	97% (3)	100% (93)	97% (4)	97% (6)
	CC_ HFWGS	100% (4)	97% (3)	100% (5)	100% (4)	100% (4)
	ReliefF	97% (41)	97% (2)	97% (69)	94% (11)	94% (5)
	ReliefF_ HFWGS	100% (4)	97% (1)	100% (4)	97% (4)	97% (4)

The results demonstrate that the HFWGS strategy consistently enhanced classification accuracy while reducing the number of selected genes, highlighting its effectiveness for gene selection in leukemia diagnostics. Table 2 summarizes the selected genes and their corresponding classification accuracies, showcasing how each classifier, when paired with the HFWGS method, assesses the importance of these genes. This table details the genes chosen by each method and their effectiveness in achieving high classification accuracy.

Table 2: Leukemia Accuracy of Different Classifiers Using Genes Selected by Various Feature Selection Methods

Classifier	Selection Method	Selected genes	Classification Accuracy (%)
K-NN	SNR-HFWGS	M27891,M23197, Y00787	100
	CC-HFWGS	M27891,M23197, U03642,Y00787	100
	ReliefF-HFWGS	M27891, M23197, L09209, Y00787	100
SVM	SNR-HFWGS	M27891, J05243	97
	CC-HFWGS	M27891,X62654,M19507	97
	ReliefF-HFWGS	M23197, J05243	97
LDA	SNR-HFWGS	M27891,M23197, Y00787, U03642	97
	CC-HFWGS	M27891,M23197, Y00787, J05243, M19507	100
	ReliefF-HFWGS	M27891,M23197, J05243, U03642	100
DT	SNR-HFWGS	M27891,M23197, U03642	97
	CC-HFWGS	M27891,M23197, Y00787, J05243	100
	ReliefF-HFWGS	M27891,M23197, M19507, L09209	97
NB	SNR-HFWGS	M27891,M23197, Y00787, J05243	97
	CC-HFWGS	M27891,M23197, Y00787, M19507	100
	ReliefF-HFWGS	M27891,M23197, U03642, J05243	97

This table emphasizes the importance of selected genes in achieving high classification accuracy across different classifiers. The consistent selection of genes like M27891, M23197, and Y00787 highlights their key role in differentiating tumor from non-tumor samples. After identifying key leukemia genes, we analyzed the frequency of each gene's selection in Figure 1.

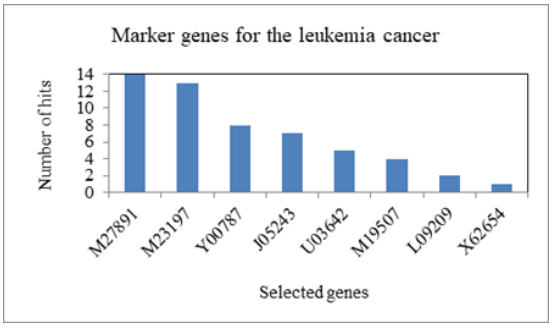

Fig. 1. Number of hits for each selected gene for leukemia

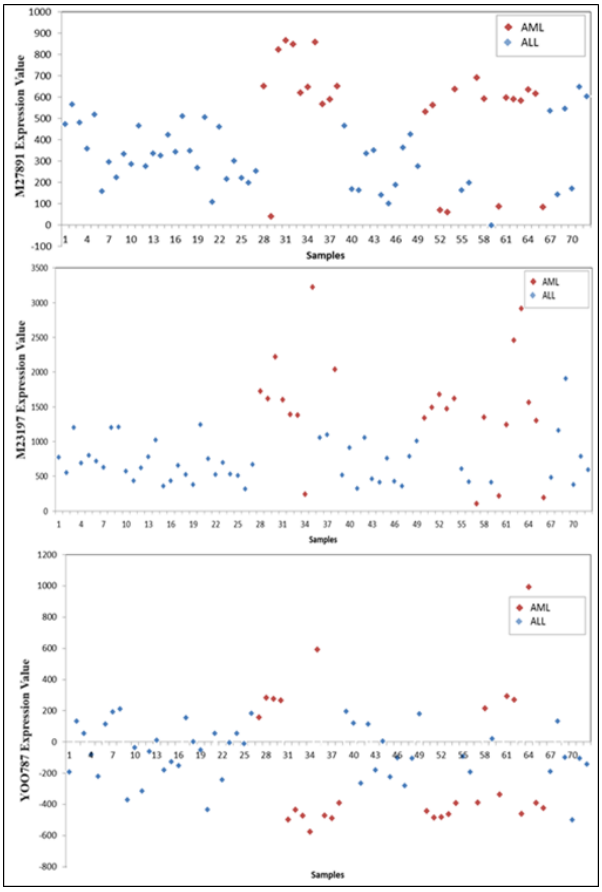

Fig. 2. Gene Expression Values for M27891, M23197, and Y00787 in ALL and AML Samples

After identifying the key genes, we examined their expression levels across 72 samples. Figure 2 illustrates the expression of the three major genes—M27891, M23197, and Y00787—across the ALL and AML classes. The x-axis represents the samples, while the y-axis shows the gene expression levels.

3.2 Results corresponding to the new HSGV

This subsection compares the performance of our HSGV method with several well-known classifiers. We began by using the HFWGS approach, focusing on the SNR filter for its proven effectiveness in achieving optimal classification. Following this gene selection, we trained six classifiers: KNN, SVM, LDA, DT, NB, and our proposed HSGV classifier. For leukemia classification, the SNR_HFWGS technique identified a subset of three key genes, as shown in Table 3.

Table 3: Cancer Classification Accuracy and Cost Time

		Classifiers					
Leukemia	**Selection method**	**KNN**	**SVM**	**LDA**	**DT**	**NB**	**HSGV**
	SNR_ HFWGS	100 %	97%	97%	97%	90%	100%
	cost time	2.3s	2.4s	3.1s	3.3s	2.7s	1.9s

The KNN classifier scored 100% accuracy, while the SVM, LDA, DT, and NB classifiers achieved 97%. Our

HSGV obtained the best accuracy of 100% in just 1.9 seconds.

3.3 Discussion

Recent studies in cancer classification using microarray data have introduced various gene selection techniques, each with unique capabilities. For instance, a 2021 study on the Isomap-GA method combines Isomap for nonlinear dimensionality reduction with a genetic algorithm (GA) for optimized gene selection, achieving 100% accuracy for leukemia with 43 genes [29]. Another 2022 research employed a hybrid approach using XGBoost and Multi-Objective Genetic Algorithm (MOGA), attaining 100% accuracy for leukemia with just 7 genes [30]. Additionally, a 2021 study focused on entropy-based gene selection, which utilized normalized mutual information and iterative selection to achieve 100% accuracy with 10 genes [31]. Collectively, these studies underscore the potential of various gene selection methods to achieve high classification accuracy while differing in the number of genes selected, thereby advancing cancer genomics through the integration of filter and wrapper methods alongside sophisticated learning algorithms and optimization techniques.

In our study, we focus on two main areas: Gene Selection and Classification. We proposed an HFWGS approach that enhances traditional filter algorithms by incorporating a noise filter to remove noise and a redundancy filter to eliminate overlapping genes. Our classification method aims to minimize gene usage while utilizing statistical measures (minimum, maximum, mean, standard deviation) for selected genes, which leads to majority voting for class decisions. Unlike the previously mentioned studies that primarily used filter techniques like SNR and ReliefF with wrapper methods, we employed classifiers such as KNN, SVM, LDA, DT, and NB. We also established the minimal number of necessary genes for classification by removing redundant genes based on expression profiles. Performance analysis on leukemia data demonstrated that our gene selection and classification methods improved both accuracy and processing time.

4 Conclusion

In this study, we introduced a novel hybrid gene selection framework combining filter and wrapper-based methods to enhance the classification of leukemia using DNA microarray data. The Hybrid Filter-Wrapper Gene Selection (HFWGS) approach consistently outperformed traditional gene selection methods, improving classification accuracy while minimizing the number of selected genes. Furthermore, the Hybrid Statistical-Gene Voting (HSGV) approach provided a robust classification mechanism, effectively differentiating between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).

The experimental results underscore the potential of the HFWGS method to optimize gene selection, leading to more accurate cancer diagnostics with fewer, but more informative, genes. The success of this approach in leukemia classification suggests its applicability to other cancers and complex diseases, offering a powerful tool for precision medicine. Future work will focus on extending this method to larger datasets and exploring its integration with deep learning models to further enhance predictive accuracy.

REFERENCES

1. Almarzouki, H. Z. (2022). Deep-learning-based cancer profiles classification using gene expression data profile. Journal of Healthcare Engineering, 2022(1), 4715998.
2. Gupta, S., Gupta, M. K., Shabaz, M., & Sharma, A. (2022). Deep learning techniques for cancer classification using microarray gene expression data. Frontiers in Physiology, 13, 952709.
3. Debnath, S., Aisha, S., Malakar, A., Perveen, K., Alfagham, A. T., Khanam, M. N., ... & Mohammed, Y. A. (2023). Understanding the cross-talk of major abiotic-stress-responsive genes in rice: A computational biology approach. Journal of King Saud University-Science, 35(7), 102786.
4. Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., & Osamor, V. C. (2022). Machine learning approaches to genome-wide association studies. Journal of King Saud University-Science, 34(4), 101847.
5. Elwahsh, H., Tawfeek, M. A., Abd El-Aziz, A. A., Mahmood, M. A., Alsabaan, M., & El-shafeiy, E. (2023). A new approach for cancer prediction based on deep neural learning. Journal of King Saud University-Computer and Information Sciences, 35(6), 101565.
6. Marie-Sainte, S. L., & Alalyani, N. (2020). Firefly algorithm based feature selection for Arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(3), 320-328.
7. Hegazy, A. E., Makhlouf, M. A., & El-Tawel, G. S. (2020). Improved salp swarm algorithm for feature selection. Journal of King Saud University-Computer and Information Sciences, 32(3), 335-344.
8. Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. Processes, 11(2), 562.
9. Dash, R. (2021). An adaptive harmony search approach for gene selection and classification of high dimensional medical data. Journal of King Saud University-Computer and Information Sciences, 33(2), 195-207.
10. Hegazy, A. E., Makhlouf, M. A., & El-Tawel, G. S. (2020). Improved salp swarm algorithm for feature selection. Journal of King Saud University-Computer and Information Sciences, 32(3), 335-344.

11. Dash, R. (2021). An adaptive harmony search approach for gene selection and classification of high dimensional medical data. Journal of King Saud University-Computer and Information Sciences, 33(2), 195-207.
12. Benkessirat, A., & Benblidia, N. (2022). A novel feature selection approach based on constrained eigenvalues optimization. Journal of King Saud University-Computer and Information Sciences, 34(8), 4836-4846.
13. KP, M. N., & Thiyagarajan, P. (2022). Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification. Journal of King Saud University-Computer and Information Sciences, 34(8), 4993-5006.
14. Hegazy, A. E., Makhlouf, M. A., & El-Tawel, G. S. (2020). Improved salp swarm algorithm for feature selection. Journal of King Saud University-Computer and Information Sciences, 32(3), 335-344.
15. Uthman, K. A., Ba-Alwi, F. M., & Othman, S. M. (2020). A survey on feature selection in microarray data: Methods algorithms and challenges. International Journal of Computer Sciences and Engineering, 8(10), 106-116.
16. Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: a survey. IEEE Transactions on knowledge and data engineering, 16(11), 1370-1386.
17. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. Oct. 1999, 286: pages 531- 537.
18. Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., ... & Wei, Y. (2022). Distance correlation application to gene co-expression network analysis. BMC bioinformatics, 23(1), 81.
19. Kohavi R and John G H (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324, 1997.
20. S. Kwon, H. Lee, S. Lee (2016), "Image enhancement with Gaussian filtering in time-domain microwave imaging system for breast cancer detection", Electronics Letters, vol. 52, no. 5, pp. 342-344, 3 2016
21. Wazery, Y. M., Saber, E., Houssein, E. H., Ali, A. A., & Amer, E. (2021). An efficient slime mould algorithm combined with k-nearest neighbor for medical classification tasks. IEEE Access, 9, 113666-113682.
22. Alwohaibi, M., Alzaqebah, M., Alotaibi, N. M., Alzahrani, A. M., & Zouch, M. (2022). 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, 34(8), 5192-5203.
23. Elwahsh, H., Tawfeek, M. A., Abd El-Aziz, A. A., Mahmood, M. A., Alsabaan, M., & El-shafeiy, E. (2023). A new approach for cancer prediction based on deep neural learning. Journal of King Saud University-Computer and Information Sciences, 35(6), 101565.
24. Almalki, Y. E., Khalid, M., Alduraibi, S. K., Yousaf, Q., Zaffar, M., Almutiri, S. M., ... & Alshamrani, H. A. (2022). LBP–Bilateral Based Feature Fusion for Breast Cancer Diagnosis. Computers Materials & Continua, 73, 4103-4121.
25. Sara, H. B., & Jihad, H. B. (2024, April). Artificial Intelligence Application for the Classification of Central Nervous System Tumors Based on Blood Biomarkers. In 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST) (pp. 1-5). IEEE.
26. Abubakar, A., Jibrin, Y., Maina, M. B., & Maina, A. B. Classification of Alzheimer's Disease Using Cnn-Based Features and Vit-Global Contextual Patterns from MRI Images. Available at SSRN 4811438.
27. Çakir, M., Yilmaz, M., Oral, M. A., Kazanci, H. Ö., & Oral, O. (2023). Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. Journal of King Saud University-Science, 35(6), 102754.
28. Chanho Park, Sung Bae Cho. Evolutionary ensemble classifier for lymphoma and colon cancer classification. Conference: Evolutionary Computation, 2003, DOI: 10.1109/CEC.2003.1299385.
29. Wang, Z., Zhou, Y., Takagi, T., Song, J., Tian, Y. S., & Shibuya, T. (2023). Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data. BMC bioinformatics, 24(1), 139.
30. Deng, X., Li, M., Deng, S., & Wang, L. (2022). Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. Medical & Biological Engineering & Computing, 60(3), 663-681.
31. Liu, X., Krishnan, A., & Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. BMC bioinformatics, 6, 1-14.