

Optimized colon cancer classification via feature selection and machine learning

Sara Haddou Bouazza, Jihad Haddou Bouazza

Research Laboratory of the Moroccan School of Engineering Sciences, LAMIGEP, EMSI, Marrakech, Morocco

Article Info

Article history:

Received Sep 10, 2024

Revised Oct 25, 2024

Accepted Nov 19, 2024

Keywords:

Artificial intelligence

Cancer classification

Computer science

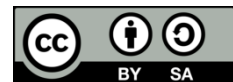
Feature selection

Machine learning

ABSTRACT

The increasing dimensionality of gene expression data poses significant challenges in cancer classification, particularly in colon cancer. This study presents a novel filtering approach (FA) and a gene classifier (GC) to enhance gene selection and classification accuracy. Utilizing a dataset of 62 samples, our methods integrate statistical measures and machine learning classifiers, achieving classification accuracies of 96% and 97%, respectively. The FA effectively filters out noise and redundancy, allowing for accurate predictions with a minimal subset of genes, while the GC leverages multiple classifiers for optimal performance. These findings underscore the importance of robust feature selection in improving cancer diagnostics and suggest potential applications in personalized medicine. By addressing the limitations of existing methodologies, our work lays the groundwork for future research in cancer genomics, emphasizing the need for adaptive strategies to handle complex datasets.

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



Corresponding Author:

Sara Haddou Bouazza

Research Laboratory of the Moroccan School of Engineering Sciences, LAMIGEP, EMSI

Marrakech, Morocco

Email: sara.hb.sara@gmail.com

1. INTRODUCTION

Deoxyribonucleic acid (DNA) microarray technology has transformed cancer research, enabling simultaneous analysis of thousands of genes, and offering valuable insights into gene interactions crucial for early detection, diagnosis, and prognosis [1], [2]. Despite this, significant challenges remain due to the imbalance between the large number of genes and the limited sample size in such datasets. Many genes are irrelevant to cancer progression or highly interdependent, complicating analyses and potentially leading to inaccurate predictions if the entire gene set is used indiscriminately [3], [4].

Feature selection is pivotal in overcoming these challenges by reducing dimensionality and excluding irrelevant or noisy genes, enhancing classification accuracy and model interpretability [5], [6]. Recent advancements in feature selection methods have been made. For instance, Hegazy *et al.* [7] demonstrated the efficacy of differential evolution techniques for colon cancer gene selection, while Ali and Saeed [8] developed a hybrid filter-genetic algorithm (GA) that improved classification across various cancer types.

Other studies, including those by Kourou *et al.* [9] and Hambali *et al.* [10], highlighted the importance of feature selection in cancer classification, noting the need for more refined techniques to address the complexity of microarray data. Chowdhary *et al.* [11] further identified ongoing issues in feature selection for high-dimensional datasets. Hybrid methods, such as those combining filter methods with algorithms like C5.0, have shown promise, as illustrated by Hamim *et al.* [12], who achieved significant improvements in breast cancer classification. Despite these advancements, current feature selection methods

require further refinement to balance classification accuracy, computational efficiency, and gene interpretability, especially in colon cancer classification. This study addresses these gaps by proposing an innovative approach integrating filter and wrapper methods to identify the most relevant genes, aiming to enhance both accuracy and efficiency in cancer diagnostics.

The structure of this paper is organized as follows: section 2 discusses the materials and methods used, detailing the dataset and feature selection strategies. Section 3 presents classification results, focusing on accuracy metrics and comparisons with state-of-the-art methods. Section 4 offers a discussion of the findings, limitations of the study, and directions for future research. Finally, section 5 concludes the paper by summarizing the main contributions and highlighting the significance of the proposed approach.

2. METHOD

Our study employs a streamlined, three-step gene selection approach integrated with classification techniques to enhance tumor classification from microarray data. First, a filter approach (FA) reduces the feature space using statistical methods like signal-to-noise ratio (SNR) and Pearson correlation coefficient (CC), identifying the most relevant genes while eliminating redundancy and noise [13], [14]. Next, a wrapper approach refines this selection by iteratively incorporating genes that improve classification performance, ensuring the model is optimized for accuracy [14]. Finally, a minimal subset selection ensures the smallest gene set is retained, balancing high classification accuracy with computational efficiency [13], [14].

2.1. Gene selection procedure

The three-step gene selection process starts with a FA, followed by a wrapper approach, and ends with selecting a minimal subset of genes. Below are the algorithms for each step along with the methods utilized.

2.1.1. Step 1: filter approach

The FA employs statistical methods to assess gene relevance before further refinement. We utilize three primary techniques: SNR, Pearson CC, and ReliefF. Algorithm 1 uses statistical measures, including SNR, Pearson CC, and ReliefF, to rank genes based on their relevance, selecting the top k genes for further analysis. The SNR differentiates gene expression patterns by measuring the maximal average expression difference between groups relative to minimal within-group variability [15], [16].

Algorithm 1. Filter approach for initial gene selection

Input: Microarray dataset D with m genes and n samples.

Output: Ranked gene subset G' based on feature relevance.

Algorithm Filter_Approach(D)

1. Initialize an empty list G' (selected genes)
2. For each gene g_i in D :
 - a. Compute Signal-to-Noise Ratio (SNR)
 - b. Compute Pearson Correlation Coefficient (CC)
 - c. Compute ReliefF score
3. Rank genes based on the combined score of SNR, CC, and ReliefF.
4. Select top k genes (k can be predefined or based on an accuracy threshold).
5. Return G' (top k genes)

End Algorithm

The CC quantifies the strength of the linear relationship between genes. The coefficient ranges from -1 to 1, with -1 and 1 indicating a perfect negative or positive linear relationship, respectively, while 0 represents no linear correlation. A positive correlation means that both genes increase together, whereas a negative correlation indicates that one increases while the other decreases [17], [18]. ReliefF, a supervised feature-weighting technique, identifies gene relevance by assessing the quality of features, initially developed by [19] and later enhanced by [20].

2.1.2. Step 2: wrapper approach

In the wrapper approach, we utilize the performance of various classifiers to further refine the gene selection process. The method iteratively adds genes and evaluates classification accuracy, retaining only those that enhance model performance.

Algorithm 2. Wrapper-based gene selection

Input: Ranked gene subset G' from Step 1, Classification algorithm C , Accuracy threshold T .

Output: Gene subset G'' with the highest classification accuracy.

Algorithm Wrapper_Selection(G' , C , T)

1. Initialize an empty set G'' (final gene subset)

```

2. Initialize best_accuracy = 0
3. For each gene g_i in G':
    a. Add g_i to G''
    b. Train classifier C using G'' and compute accuracy A
    c. If A > best_accuracy:
        i. best_accuracy = A
        ii. Keep g_i in G''
    d. Else, remove g_i from G''
    e. If best_accuracy > T, break
4. Return G'' (genes that maximize accuracy)
End Algorithm

```

The wrapper approach adds one gene at a time to the model, evaluates classification accuracy, and only keeps genes that improve accuracy. This iterative process continues until accuracy stabilizes or reaches a predefined threshold T.

2.1.3. Step 3: minimal subset selection

In this final step, we aim to select the smallest possible gene subset that maintains classification accuracy. This method focuses on optimizing both performance and computational efficiency. Algorithm 3 ensures that the final gene subset is minimal, retaining only those genes necessary for maintaining classification performance.

Algorithm 3. Minimal subset selection

Input: Gene subset G' from Step 2, Classifier C.

Output: Minimal subset G_{min}.

```

1. Initialize G_{min} = G'
2. For each gene g_i in G':
    a. Temporarily remove g_i from G_{min}
    b. Train classifier C on G_{min} and compute accuracy A
    c. If A decreases, add g_i back to G_{min}
3. Return G_{min}
End Algorithm

```

2.2. Classification methods

We implemented five classification algorithms: K-nearest neighbor (KNN), support vector machine (SVM), linear discriminant analysis (LDA), decision tree (DT), and naive Bayes (NB). Below are the pseudocode descriptions of these classifiers.

2.2.1. K-nearest neighbor

KNN is a simple, instance-based learning algorithm that classifies a test sample based on the classes of its nearest neighbors in the feature space. The majority voting mechanism among the KNN determines the predicted class. KNN is effective for small datasets and is particularly useful when the decision boundary is irregular [21], [22].

Algorithm 4. KNN classification

```

1. For each sample s in D_{train}:
    a. Compute Euclidean distance d between x and s
2. Sort D_{train} by distance d
3. Select top K nearest neighbors
4. Return the majority class of the K neighbors as y
End Algorithm

```

2.2.2. Classification algorithm: support vector machine

SVM is a supervised learning model that aims to find the optimal hyperplane that maximizes the margin between different classes in a high-dimensional space. By using kernel functions, SVM can effectively handle non-linear decision boundaries. It is known for its robustness and effectiveness in high-dimensional datasets, making it suitable for gene expression data [23].

Algorithm 5. SVM classification

Input: Training set D_{train}, Test sample x, Kernel function K.

Output: Predicted class y.

Algorithm SVM(D_{train}, x, K)

```

1. Train the SVM model with D_{train} and kernel function K
2. Compute the optimal hyperplane H
3. Project x onto H
4. Return the class of x based on its position relative to H
End Algorithm

```

2.2.3. Classification algorithm: linear discriminant analysis

LDA is a linear classifier that finds the linear combination of features that best separates two or more classes. It minimizes within-class variance while maximizing between-class variance, making it particularly effective for classifying linearly separable data. LDA is widely used in situations where the assumption of normality holds [24].

Algorithm 6. LDA classification

Input: Training set D_{train} , Test sample x .

Output: Predicted class y .

Algorithm $\text{LDA}(D_{\text{train}}, x)$

1. Compute within-class scatter matrix and between-class scatter matrix
 2. Calculate the projection vector W that maximizes the ratio of between-class to within-class variance
 3. Project x onto W
 4. Return the class based on the projection
- End Algorithm

2.2.4. Classification algorithm: decision tree

DT are tree-like structures that recursively partition the feature space based on attribute values to make predictions. They are intuitive, easy to interpret, and can handle both categorical and continuous data. However, they may suffer from overfitting if not properly pruned [25], [26].

Algorithm 7. Decision tree classification

Input: Training set D_{train} , Test sample x .

Output: Predicted class y .

Algorithm $\text{Decision_Tree}(D_{\text{train}}, x)$

1. Train a decision tree using D_{train}
 2. Traverse the tree from the root node based on feature values of x
 3. Arrive at a leaf node representing the predicted class
 4. Return the class at the leaf node as y
- End Algorithm

2.2.5. Classification algorithm: Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features given the class label [25].

Algorithm 8. Naive Bayes classification

Input: Training set D_{train} , Test sample x .

Output: Predicted class y .

Algorithm $\text{Naive_Bayes}(D_{\text{train}}, x)$

1. For each class c in D_{train} :
 - a. Compute prior probability $P(c)$
 - b. For each feature f in x , compute likelihood $P(f|c)$
 2. Compute posterior probability $P(c|x)$ for each class
 3. Return class with highest posterior probability as y
- End Algorithm

Classification accuracy was considered the key evaluation metric for the classifiers [27]:

$$\text{Accuracy} = 100 \frac{TP + TN}{TN + TP + FN + FP} \quad (1)$$

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

2.3. Our proposition for gene classification for binary class problems

In binary class problems, we introduce ensemble-based voting using classifiers such as random forest (RF) and extreme gradient boosting (XGBoost), along with other techniques to improve performance.

2.3.1. Step 1: statistical measures of selected genes

We employ the enhanced statistical techniques described earlier to extract meaningful patterns from the gene expression data. Each gene's statistical measures are calculated, and classification intervals are defined using robust methods such as IQR.

2.3.2. Step 2: ensemble-based voting with random forest and extreme gradient boosting

We extend our ensemble approach by introducing stacking and weighted voting. Our ensemble consists of RF, XGBoost, and additional classifiers like light gradient boosting machine (LightGBM), adaptive boosting (AdaBoost), and categorical boosting (CatBoost) to increase the robustness and accuracy

of classification. A meta-classifier (e.g., logistic regression) is trained to combine the outputs of these base classifiers. Additionally, we apply weighted voting, where classifiers with higher individual performance are assigned more importance in the final vote. If the confidence of the majority vote falls below a threshold, further analysis (e.g., fallback KNN) is triggered.

Algorithm 9. Enhanced_Ensemble_Gene_Voting (G_{opt} , x , Classifiers)

```

1. Input: Optimized gene subset  $G_{opt}$ , Test sample  $x$ , List of Classifiers
2. Output: Predicted class  $y$ 
3. Initialize  $vote\_class1 = 0$ ,  $vote\_class2 = 0$ 
4. For each gene  $g_i$  in  $G_{opt}$ :
    a. Compute the expression interval [ $Mean_i - Std_i$ ,  $Mean_i + Std_i$ ]
    b. If  $x_i$  lies outside the interval for class 1, vote for class 2
    c. If  $x_i$  lies outside the interval for class 2, vote for class 1
5. Apply Random Forest, XGBoost, LightGBM, AdaBoost, and CatBoost classifiers to  $x$ 
6. Perform stacking to combine predictions using a meta-classifier (e.g., logistic regression)
7. Apply weighted voting scheme based on classifier performance
8. If majority vote confidence < threshold (e.g., 70%), apply fallback KNN or Isolation Forest
9. Tally the votes from individual gene decisions and ensemble methods
10. Return the class with the majority or weighted vote as  $y$ 
End Algorithm

```

By applying advanced statistical measures, robust ensemble methods, and techniques like stacking, weighted voting, and outlier detection, our proposed approach offers a powerful and flexible system for selecting the minimal subset of genes and classifying tumors based on microarray data. The combination of RF, XGBoost, and other classifiers ensures that our system handles complex datasets with high accuracy and robustness.

2.4. Data analysis of microarray gene expression in colon cancer

The microarray dataset utilized in our study is structured as an $N \times M$ matrix, where N corresponds to the number of biological samples (40 cancerous and 22 normal colon tissue samples) and M refers to the total number of genes (over 6,500 human genes). Each entry in this matrix represents the expression level of a specific gene in a particular sample.

2.4.1. Context and significance

Colon cancer, also referred to as colorectal cancer, arises from malignancies in the lining of the colon and rectum, resulting from uncontrolled cell proliferation. This proliferation can lead to the invasion of surrounding tissues and potential metastasis to distant organs, making early detection and accurate classification critical for effective treatment. The gene expression profiling of colon tissues can provide insights into the underlying molecular mechanisms of cancer progression and help identify potential biomarkers for diagnosis and therapy.

2.4.2. Dataset description and preprocessing

The dataset includes gene expression measurements from 40 cancerous and 22 normal samples, obtained via the Affymetrix oligonucleotide array platform, which analyzes over 6,500 genes with high accuracy. To ensure reliability, 2,000 genes were selected based on high signal intensity, variability between cancerous and normal samples, and excluding those with more than 10% missing data. Normalization techniques, like quantile normalization and robust multi-array average (RMA), were applied to adjust for technical variability, while outlier detection helped exclude anomalous samples. The final dataset, with 2,000 genes, was used for gene selection, classification, and validation. Full dataset details are available at [28]. A complete description of the dataset can be found at [28] and the data is downloadable from genomics-pubs.princeton.edu/oncology/affydata/index.html.

3. RESULTS AND DISCUSSION

The experiments were conducted on a laptop with the following hardware configuration: an Intel® Core™ i5 CPU M 250 @ 2.4 GHz processor, dual-core, with 4 GB of RAM. The system operated on Windows 10 (64-bit), providing a standard computational environment. The MATLAB software package (version R2023a) was used to implement the algorithms, execute the data analysis, and perform classification tasks. Although this setup is not highly specialized for large-scale computations, it was adequate for handling the size of the colon cancer dataset, consisting of 2,000 genes across 62 samples.

3.1. Results corresponding to the new filtering approach

In this study, we assessed the performance of various classifiers combined with different feature selection methods on colon cancer data, aiming to achieve optimal classification accuracy while minimizing the number of selected genes. The results demonstrate the significant impact of the FA in enhancing classification performance. As shown in Table 1, the classification accuracies varied across classifiers depending on the feature selection technique used. For the SNR method, classification accuracy ranged from 85.7% to 92.8% when using 2 to 29 genes. However, when SNR was combined with the FA method, the accuracy improved to 96%, with only 4 to 6 genes selected. Similarly, the CC method yielded accuracies between 85.7% and 92.8% with 2 to 27 genes, while the FA integration raised the accuracy to 96%, requiring just 3 to 5 genes. The ReliefF method, which initially produced accuracies ranging from 78.5% to 90% using 11 to 78 genes, achieved 94% accuracy when combined with FA, using only 4 to 5 genes.

Table 1. Performance of various classifiers with different feature selection methods for colon cancer

Feature selection methods	Classifiers									
	KNN		SVM		LDA		DT		NB	
SNR	92.80%	(5)	85.70%	(29)	92.80%	(2)	91%	(21)	85.70%	(22)
SNR_FA	96%	(4)	92.80%	(9)	94%	(5)	92.80%	(6)	91%	(6)
CC	92.80%	(7)	85.70%	(2)	92.80%	(27)	92.80%	(21)	85.70%	(5)
CC_FA	96%	(5)	95%	(4)	95%	(4)	95%	(5)	91%	(4)
ReliefF	85.70%	(40)	85.70%	(11)	78.50%	(78)	90%	(26)	85.70%	(64)
ReliefF_FA	91%	(5)	92.80%	(4)	91%	(4)	94%	(5)	92.80%	(5)

These results indicate that the proposed FA strategy consistently improved classification accuracy while simultaneously reducing the number of selected genes. This dimensionality reduction is critical for enhancing efficiency and minimizing overfitting in cancer diagnostics. Table 2 shows a summary of the selected genes and the corresponding classification accuracies, where the classification was re-predicted by each classifier using the FA method to better assess their importance. In the following table, one can see which genes are chosen by each specific method and their efficacy in achieving high classification accuracy.

Table 2. Colon classification accuracy of different classifiers using genes selected by various feature selection methods

Classifier	Selection method	Selected genes	Classification accuracy (%)
kNN	SNR-FA	M22382, J02854, T57619, T92451	96
	CC-FA	M76378, T71025, H64489, Z24727, T57619,	96
	ReliefF-FA	T92451, L09209, M63391, Z50753, H64489	91
SVM	SNR-FA	M63391, R87126, M22382, T92451, U09564, T57619,	92.8
		M26697, R08183, H64489	
	CC-FA	M76378, H64489, T95018, R87126	95
LDA	ReliefF-FA	M63391, H43887, T60155, T62947	92.8
	SNR-FA	H64489, M63391, T92451, T57619, M22382	94
	CC-FA	H64489, M63391, M76378, R08183	95
DT	ReliefF-FA	M63391, T92451, U09564, Z24727	91
	SNR-FA	H64489, M63391, T92451, T57619, M22382, J02854	92.8
	CC-FA	H64489, M63391, R87126, R08183, T95018	95
NB	ReliefF-FA	H64489, M63391, U09564, M26697, H43887	94
	SNR-FA	H64489, M63391, T92451, T57619, M76378, J02854	91
	CC-FA	M63391, M76378, R87126, M26697	91
	ReliefF-FA	H64489, M63391, T92451, T57619, T71025	92.8

This table reveals that the chosen genes are important in achieving good classification accuracy across various classifiers. The repeated selection of some genes, such as H64489 and M63391, demonstrates their critical role in distinguishing between tumor and non-tumor samples. After identifying the major genes for colon cancer, we used all categorization techniques to assess how often each gene is selected. The frequency of picking each gene, as well as the most useful and important genes in the classification of colon cancer, are depicted in Figure 1.

Following the identification of the major genes, their expression levels were studied across 62 samples. Figures 2(a)-(d) details the gene expression of the four major genes—M63391, H64489, T92451, and T57619—in various samples, categorized into tumor and normal classes. The x-axis represents the samples, numbered from 1 to 62, while the y-axis represents the gene expression levels. This figure clearly

captures the variation in expression levels for these genes between the colon cancer groups, facilitating more accurate sample classification.

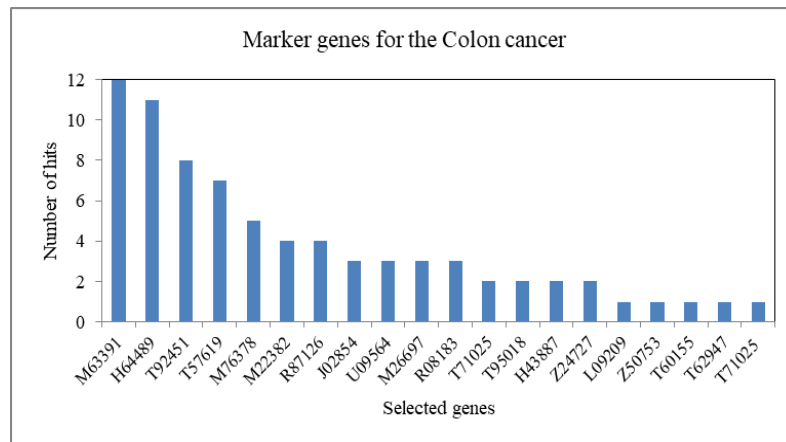


Figure 1. Number of hits for each selected gene for colon cancer

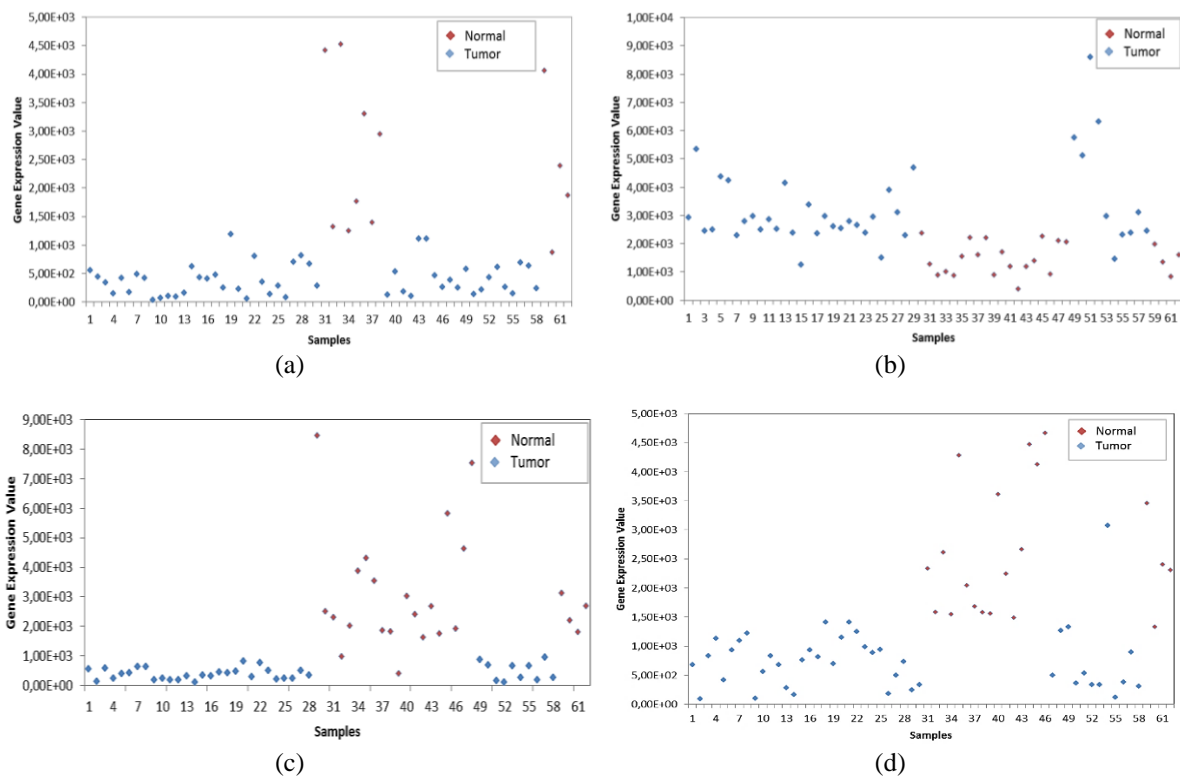


Figure 2. Gene expression values for: (a) gene T92451, (b) gene T57619, (c) gene M63391, and (d) gene H64489 in normal and tumor samples

3.2. Results corresponding to the new gene classifier

This section compares the performance of the proposed gene classifier (GC) with KNN, SVM, LDA, DT, and NB. The GC's effectiveness is evaluated using gene subsets selected through the SNR filter within the FA. For colon cancer classification, the SNR-FA method identified four key genes (Table 3), enhancing both accuracy and efficiency across all models, with GC showing superior performance. The results demonstrate that our proposed GC achieved the best classification accuracy of 97%, surpassing all

other classifiers, including KNN (96%), LDA (94%), DT (94%), and NB (94%). In addition, the GC significantly reduced computation time, requiring only 19 seconds to perform the classification, which is a notable improvement compared to the other classifiers, particularly DT (47 s) and SVM (40 s).

Table 3. Colon cancer classification accuracy and processing time of various classifiers using SNR_FA

Selection method	Classifiers					
	KNN	SVM	LDA	DT	NB	GC
SNR_FA	96%	92.8%	94%	94%	94%	97%
cost time	37 s	40 s	33 s	47 s	41 s	19 s

3.3. Interpretation of results

The results from both the FA and the GC reveal substantial improvements in classification accuracy and computational efficiency for colon cancer classification. By implementing the FA, we significantly reduced the number of selected genes while maintaining or enhancing performance across classifiers. This dimensionality reduction is particularly valuable for high-dimensional datasets like microarrays, where many irrelevant or redundant genes can negatively impact classifier performance.

The SNR filter, when combined with FA, consistently boosted classification accuracy across multiple classifiers. Specifically, it achieved 96% accuracy with KNN and SVM, and 94% with DT and NB, using only 4-6 genes. This demonstrates that the FA method effectively identifies a compact yet informative set of genes, thereby improving both efficiency and accuracy in classification.

Furthermore, the GC further refined classification performance, achieving the highest accuracy of 97% with a processing time of just 19 seconds. This highlights the potential of a custom classifier optimized for selected gene subsets, surpassing traditional classifiers like KNN, SVM, and LDA in both speed and accuracy. These findings underscore the importance of integrating advanced feature selection techniques with tailored classifiers to achieve robust cancer classification from gene expression data.

3.4. Limitations

Despite the promising results, several limitations must be addressed. Firstly, the small sample size (62 samples) used in this study may limit the generalizability of the findings. Larger datasets are needed to validate the robustness and scalability of both the FA and GC methods across diverse clinical contexts. Secondly, the dataset's class imbalance, with more tumor samples than normal ones, could potentially bias the classifiers, leading to an overestimation of accuracy for the majority class. Addressing this issue through techniques like resampling or stratification could ensure more reliable evaluations.

Moreover, the proposed methods were tested exclusively on colon cancer data. Future research should investigate the applicability of FA and GC across other cancer types and broader biological datasets. Lastly, while the computational efficiency of GC was demonstrated, further testing on larger datasets is necessary to evaluate its scalability in more complex scenarios, especially when handling hundreds or thousands of samples.

4. DISCUSSION

In this section, we critically evaluate the proposed FA and GC against prominent methodologies in cancer classification using microarray data. The Isomap-GA method [29], which integrates Isomap for nonlinear dimensionality reduction with a GA for gene selection, achieved 85.8% accuracy using 11 genes. In contrast, our FA method excels by removing noise and redundancy, yielding a higher accuracy of 96% with fewer genes, thereby offering more robust classification while enhancing computational efficiency.

Similarly, the hybrid gene selection method employing XGBoost and a multi-objective genetic algorithm (MOGA) [30] achieved 90.2% accuracy with 62 genes, but at the cost of computational intensity. Our FA, focusing on key statistical measures, simplifies gene selection, achieving higher efficiency. Additionally, the GC outperformed this method by reaching 97% accuracy in just 19 seconds, a notable improvement in both accuracy and speed.

Compared to the entropy-based gene selection method [31], which achieved 91.9% accuracy with 9 genes, our FA addresses a critical limitation by incorporating noise filtering, ensuring that only relevant genes are selected. When combined with the GC's majority voting across classifiers such as KNN, SVM, and LDA, our approach achieved 97% accuracy with minimal computational overhead. In summary, the FA and GC methods surpass existing techniques in accuracy, computational efficiency, and robustness in gene selection. By effectively eliminating noise and redundancy while reducing the number of selected genes, our

approach significantly improves colon cancer diagnostics and paves the way for further advancements in cancer genomics.

5. CONCLUSION

In this study, we have introduced a novel FA for gene selection and a GC for colon cancer classification, significantly enhancing classification accuracy while reducing the number of genes required. Our results demonstrate that by effectively filtering out noise and redundancy, we can improve diagnostic accuracy to 96% and 97% for our proposed methods, outperforming existing techniques in the field. These findings not only contribute to a deeper understanding of gene interactions in colon cancer but also offer a practical framework for the application of microarray data in clinical settings.

The implications of our research extend beyond the immediate findings, highlighting the importance of robust gene selection methods in cancer diagnostics. By focusing on minimizing dimensionality while maximizing classification performance, our approach provides a valuable tool for researchers and clinicians aiming to utilize gene expression data for early detection and personalized treatment strategies. Furthermore, the integration of diverse classifiers in our GC emphasizes the need for adaptive methodologies in the evolving landscape of cancer genomics.

Looking forward, future research could explore the application of our methods in other cancer types, as well as their integration with emerging technologies such as artificial intelligence and machine learning. This could facilitate more comprehensive analyses of large-scale genomic datasets and improve the accuracy of cancer classifications. Moreover, investigating the biological relevance of the selected genes could provide insights into the underlying mechanisms of tumorigenesis and inform the development of targeted therapies. In conclusion, our study not only addresses current challenges in gene selection and cancer classification but also sets the stage for future advancements in the field. By fostering a collaborative approach between computational methods and biological research, we can enhance our understanding of cancer and ultimately improve patient outcomes.




REFERENCES

- [1] H. Z. Almarzouki, "Deep-learning-based cancer profiles classification using gene expression data profile," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–13, Jan. 2022, doi: 10.1155/2022/4715998.
- [2] S. Gupta, M. K. Gupta, M. Shabaz, and A. Sharma, "Deep learning techniques for cancer classification using microarray gene expression data," *Frontiers in Physiology*, vol. 13, Sep. 2022, doi: 10.3389/fphys.2022.952709.
- [3] S. Debnath *et al.*, "Understanding the cross-talk of major abiotic-stress-responsive genes in rice: a computational biology approach," *Journal of King Saud University - Science*, vol. 35, no. 7, p. 102786, Oct. 2023, doi: 10.1016/j.jksus.2023.102786.
- [4] D. O. Enoma, J. Bishung, T. Abiodun, O. Ogunlana, and V. C. Osamor, "Machine learning approaches to genome-wide association studies," *Journal of King Saud University - Science*, vol. 34, no. 4, p. 101847, Jun. 2022, doi: 10.1016/j.jksus.2022.101847.
- [5] H. Elwahsh, M. A. Tawfeek, A. A. Abd El-Aziz, M. A. Mahmood, M. Alsabaan, and E. El-shafeiy, "A new approach for cancer prediction based on deep neural learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 6, p. 101565, Jun. 2023, doi: 10.1016/j.jksuci.2023.101565.
- [6] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 3, pp. 320–328, Mar. 2020, doi: 10.1016/j.jksuci.2018.06.004.
- [7] A. E. Hegazy, M. A. Makhoul, 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, Mar. 2020, doi: 10.1016/j.jksuci.2018.06.003.
- [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, p. 562, Feb. 2023, doi: 10.3390/pr11020562.
- [9] K. Kourou *et al.*, "A machine learning-based pipeline for modeling medical, socio-demographic, lifestyle and self-reported psychological traits as predictors of mental health outcomes after breast cancer diagnosis: an initial effort to define resilience effects," *Computers in Biology and Medicine*, vol. 131, p. 104266, Apr. 2021, doi: 10.1016/j.combiomed.2021.104266.
- [10] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: review, challenges and research directions," *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 78–97, Jun. 2020, doi: 10.1016/j.ijcce.2020.11.001.
- [11] C. L. Chowdhary, N. Khare, H. Patel, S. Koppu, R. Kaluri, and D. S. Rajput, "Past, present and future of gene feature selection for breast cancer classification—a survey," *International Journal of Engineering Systems Modelling and Simulation*, vol. 13, no. 2, pp. 140–153, 2022, doi: 10.1504/IJESMS.2022.123345.
- [12] M. Hamim, I. El Moudden, H. Moutachaouik, and M. Hain, "Decision tree model based gene selection and classification for breast cancer risk prediction," in *Communications in Computer and Information Science*, vol. 1207 CCIS, 2020, pp. 165–177, doi: 10.1007/978-3-030-45183-7_12.
- [13] M. N. Muhammed 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.2020.12.009.
- [14] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: a review," *Bioengineering*, vol. 10, no. 2, p. 173, Jan. 2023, doi: 10.3390/bioengineering10020173.
- [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 Sciences and Engineering*, no. November, 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, Nov. 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, no. October, pp. 531–537, 1999.
- [18] J. Hou *et al.*, "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, no. 1, p. 81, Dec. 2022, doi: 10.1186/s12859-022-04609-x.
- [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.2015.3613.
- [21] Y. M. Wazery, E. Saber, E. H. Houssein, A. A. Ali, and E. Amer, "An efficient slime mould algorithm combined with k-nearest neighbor for medical classification tasks," *IEEE Access*, vol. 9, pp. 113666–113682, 2021, doi: 10.1109/ACCESS.2021.3105485.
- [22] M. Alwohaibi, M. Alzaqebah, N. M. Alotaibi, A. M. Alzahrani, and M. Zouch, "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, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.004.
- [23] S. H. Bouazza, L. Wakrim, and L. G. Benedress, "Classifying leukemia through dna expression data mining techniques," in *2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, IEEE, May 2023, pp. 387–391, doi: 10.1109/MI-STA57575.2023.10169443.
- [24] Y. E. Almalki *et al.*, "LBP-bilateral based feature fusion for breast cancer diagnosis," *Computers, Materials & Continua*, vol. 73, no. 2, pp. 4103–4121, 2022, doi: 10.32604/cmc.2022.029039.
- [25] H. B. Sara and H. B. Jihad, "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)*, IEEE, Apr. 2024, pp. 1–5, doi: 10.1109/GAST60528.2024.10520752.
- [26] A. Abubakar, Y. Jibrin, M. B. Maina, and A. B. Maina, "Classification of alzheimer's disease using cnn-based features and vit-global contextual patterns from mri images," *SSRN*, pp. 1–22, 2024, doi: 10.2139/ssrn.4811438.
- [27] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazanci, and O. Oral, "Accuracy assessment of rfems, nb, svm, and knn machine learning classifiers in aquaculture," *Journal of King Saud University - Science*, vol. 35, no. 6, p. 102754, Aug. 2023, doi: 10.1016/j.jksus.2023.102754.
- [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.*, IEEE, 2003, pp. 2378–2385, doi: 10.1109/CEC.2003.1299385.
- [29] Z. Wang, Y. Zhou, T. Takagi, J. Song, Y.-S. Tian, and T. Shibuya, "Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data," *BMC Bioinformatics*, vol. 24, no. 1, p. 139, Apr. 2023, doi: 10.1186/s12859-023-05267-3.
- [30] 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, Mar. 2022, doi: 10.1007/s11517-021-02476-x.
- [31] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6, no. 1, p. 76, Mar. 2005, doi: 10.1186/1471-2105-6-76.

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 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.