Optimized Machine Learning for Cancer Classification via Three-Stage Gene Selection

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

Received: 1 November 2024 | Revised: 8 December 2024 and 20 December 2024 | Accepted: 22 December 2024

 DOI: https://doi.org/10.48084/etasr.9473

ABSTRACT

Gene selection from high-dimensional microarray data presents challenges such as overfitting, computational inefficiency, and feature redundancy. Despite significant advances, existing methods often suffer from limitations in scalability and interpretability, especially for precision oncology. This study introduces a novel Three-Stage Gene Selection (3SGS) strategy that addresses these issues through a combination of filter-based methods (signal-to-noise ratio, correlation coefficient, ReliefF) with accuracy-driven refinement and redundancy reduction. The 3SGS approach identifies minimal but highly predictive gene subsets, achieving 100% accuracy for leukemia and 98% for prostate cancer using only 3-4 genes. Compared to traditional methods, 3SGS enhances efficiency and interpretability, establishing itself as a scalable and robust solution for cancer classification.

Keywords-artificial intelligence; data mining; machine learning; pattern recognition; computer science

I. INTRODUCTION

DNA microarray technology has revolutionized biomedical research, particularly in cancer diagnosis and treatment, allowing the large-scale analysis of gene expression [1-4]. However, the high dimensionality of microarray data, where thousands of genes are profiled from only a limited number of samples, presents significant challenges for machine learning models. This often results in overfitting and poor generalizability, making accurate classification difficult [5-7]. To address these issues, effective gene selection is crucial for improving model performance and interpretability [8-10].

Gene selection methods are generally classified into filter, wrapper, and hybrid approaches. Filter methods independently rank genes based on their relevance to class separability, offering speed and simplicity but limited accuracy. Wrapper methods, in contrast, refine gene subsets iteratively based on classifier performance, improving accuracy but at a higher computational cost [11-15]. Hybrid methods seek to combine the strengths of both by enhancing accuracy while maintaining efficiency. Despite these advances, existing approaches often struggle with balancing interpretability, computational efficiency, and classification accuracy. For instance, methods combining genetic algorithms with manifold learning [16] achieve high accuracy but are computationally intensive and require complex parameter tuning. Similarly, hybrid techniques leveraging multi-objective optimization [17] improve performance but lack scalability for large datasets.

To address these gaps, this study introduces the 3SGS approach, which integrates both filter and wrapper techniques to optimize gene selection for cancer classification. By combining Signal-to-Noise Ratio (SNR), Correlation Coefficient (CC), and ReliefF with iterative refinement and redundancy reduction, the proposed method simplifies the process while achieving superior performance using minimal gene subsets. Experiments on leukemia and prostate cancer datasets demonstrate that 3SGS achieves high classification accuracy with only 3-4 genes, making it an efficient and interpretable solution for cancer classification [18-21].

II. METHOD

A. Summary of the Three-Stage Gene Selection (3SGS) Method

The 3SGS method systematically refines gene selection for cancer classification through the following stages:

1. Filter-Based Gene Selection: Identifies the most relevant genes using three filter methods: SNR, CC, and ReliefF. This stage narrows the gene pool by selecting the top-ranked candidates based on relevance metrics.
2. Accuracy-Driven Selection: Iteratively evaluates the predictive performance of combinations of the top-ranked genes using a classifier. Genes that improve or maintain accuracy are retained, creating a refined subset.
3. Redundancy Reduction: Eliminates redundant genes from the refined subset to produce a minimal, non-redundant set of genes with maximal classification power.

This structured process is designed to balance computational efficiency, predictive accuracy, and interpretability, enabling precise and effective cancer classification (Figure 1). The 3SGS model is designed to minimize computational costs while ensuring high

classification accuracy by selecting a minimal, highly predictive subset of genes.

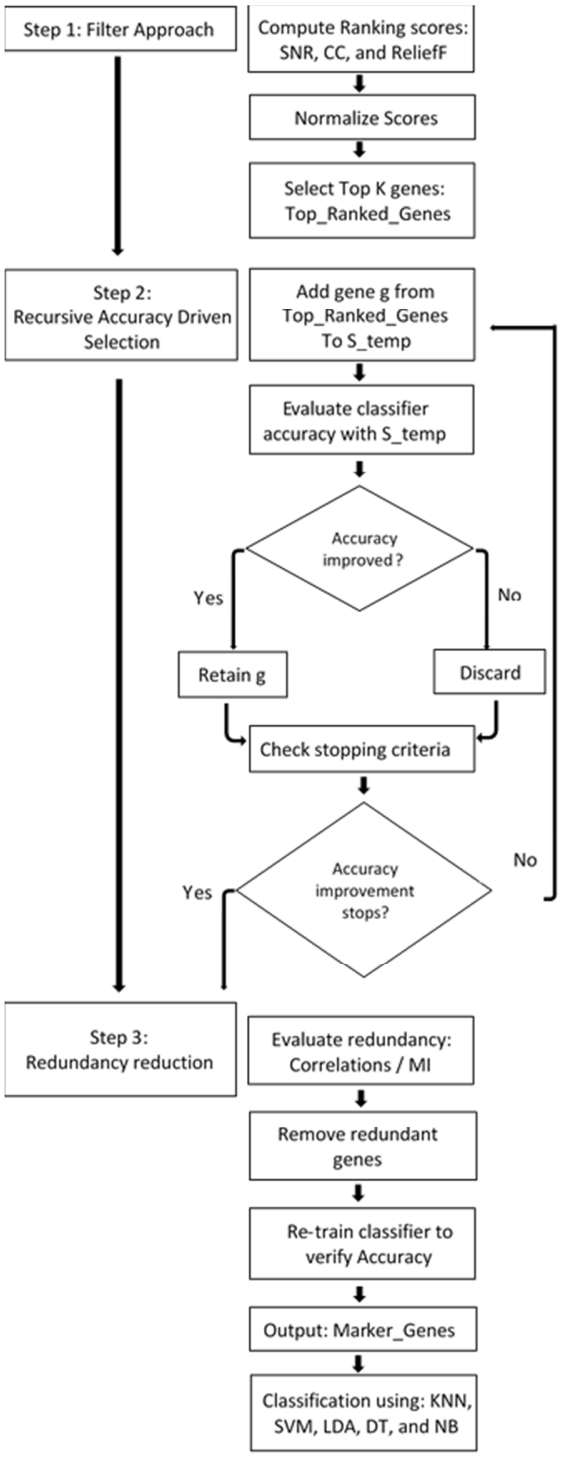

Fig. 1. Workflow of the 3SGS method.

B. Gene Selection

Gene selection is a critical component of cancer classification, especially in microarray data, where the number of genes far exceeds the number of samples, presenting challenges of high dimensionality, computational inefficiency, and the risk of overfitting. The 3SGS approach combines three filter-based techniques alongside a recursive accuracy-driven selection strategy. This approach refines the gene subset to be both minimal and maximally informative, facilitating the effective classification of cancerous tissues.

SNR ranks genes by evaluating the difference in mean expression levels between classes, normalized by within-class variability. This method is computationally efficient and performs well in datasets with low overlap between class distributions, highlighting clear separability. However, its reliance on linear assumptions limits its ability to detect non-linear patterns and gene interactions, potentially overlooking relationships critical for classification accuracy [15, 18]. The SNR for a gene j is defined as:

$$P(j) = \frac{M_{1j} - M_{2j}}{S_{1j} + S_{2j}} \quad (1)$$

where M_{kj} and S_{kj} denote the mean and the standard deviation of the gene j for samples of classes $k = 1, 2$. A higher SNR value indicates a greater distinction between classes, enabling efficient identification of candidate genes.

The CC, particularly Pearson's correlation, measures the linear relationship between gene expression levels and class labels. It is computationally efficient and effective for identifying genes with consistent expression patterns and strong class separability. However, like SNR, CC assumes linear relationships, limiting its performance with non-linear data. Additionally, CC does not account for redundancy among selected genes, which can degrade the quality of the gene subset [19]. The correlation coefficient for gene j is calculated as:

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

where r is the Pearson correlation score, X_{ij} is the i^{th} sample value for the gene j, Y_i is the corresponding class, and $\overline{X_j} = 1/n \sum_{i=1}^{n} X_{ij}$ and $\overline{Y}$ are the means for gene j and the classes, respectively.

ReliefF is a sophisticated filter method that identifies genes based on their ability to differentiate classes by considering neighboring samples. Unlike SNR and CC, ReliefF can detect non-linear interactions and is robust to noise, making it effective for complex datasets with overlapping classes. However, it is computationally more demanding, particularly for large datasets, and its performance depends on the number of nearest neighbors and iterations, which can be challenging to tune effectively [20, 21]. The weight W_A assigned to gene A is calculated as:

$$W_A = w_d - \sum_{j=1}^{k} \frac{\text{diff}(A_i, X_i, hits_j)}{m*k} + \sum_{c \neq class(Xi)} \frac{p(c)}{1-p(class(X_i))} \sum_{j=1}^{k} \frac{diff(A_i, X_i, misses_j)}{m*k} \quad (3)$$

where the distance used is defined by:

$$diff(A_i, X_1, X_2) = \frac{|value(A_i, X_1) - value(A_i, X_2)|}{\max(A) - \min(A)} \quad (4)$$

Here, X_i is an instance described by the vector A_i of n genes, m is the number of process repetitions (a user-defined parameter), k is the number of nearest misses, and $hits$ and $misses$ refer to nearest hit and miss instances, respectively.

SNR is efficient but limited to linear separability, often missing non-linear patterns in the data. CC is effective in identifying consistent gene-class relationships but may struggle with redundancy between genes, potentially reducing the overall performance. ReliefF, on the other hand, excels at handling noisy data and can detect non-linear interactions between genes, although it requires more computational resources for tuning, especially when dealing with larger datasets. Together, these three methods provide complementary strengths, enabling a more robust gene selection process in the 3SGS framework.

The filter-based selection stage of 3SGS leverages these ranking methods to independently evaluate each gene's relevance. However, while this approach is effective in identifying potentially important genes, it can also select genes that introduce noise, which could hinder classification accuracy if included without further validation. To mitigate this risk, a wrapper-based strategy is introduced in the second stage of 3SGS.

In the second stage, an iterative process is adopted where genes from the filter-selected subset are incrementally added, and the classifier's accuracy is evaluated with each addition. Only genes that show a clear improvement in classification performance are retained. This strategy allows refining the gene subset by directly optimizing classifier performance and eliminating unnecessary complexity.

By the end of the first two stages, a set of highly relevant genes is identified. The third stage performs redundancy reduction, which involves analyzing the selected genes to remove those that are highly correlated or redundant. This ensures that the final subset of genes is both minimal and highly discriminative, leading to improved model interpretability and greater computational efficiency.

C. Algorithm of the 3SGS Approach

The 3SGS algorithm is developed to optimize cancer classification by refining the gene selection process through three distinct stages. This approach combines filter and wrapper strategies to identify a minimal and highly informative subset of genes. By doing so, classification accuracy while simultaneously reducing computational complexity. The three stages of the 3SGS algorithm steps are as follows.

```
Algorithm 1: Three-Stage Gene Selection
1: Feature Selection using Filter Methods:
   Rank genes based on SNR, CC, or
   ReliefF.
   Normalize scores (if needed) for
   comparability.
   Select the top k genes as candidates.
2: Recursive Accuracy-Driven Selection:
   Initialize an empty subset S to hold
   selected genes.
   Iteratively add genes from the top-
   ranked list to S, evaluating classifier
   accuracy at each step.
   Retain genes that improve or maintain
   accuracy.
3: Redundancy Reduction:
   Evaluate redundancy in S by calculating
   correlations or mutual information.
   Remove redundant genes, keeping only
   those that provide unique information.
   Verify classifier performance using the
   reduced subset.
```

D. Parameter Tuning and Robustness Discussion

The performance of the 3SGS approach is influenced by parameters such as the number of top-ranked genes k and the classifier used. For this study, k was selected based on preliminary experiments balancing computational efficiency and classifier accuracy. Specifically:

- Filter-Based Selection: Values of k were initially set to 50, 100, and 200, and the classifier accuracy was monitored for each subset. The optimal k was chosen based on the highest cross-validation accuracy.
- ReliefF Parameters: The number of nearest neighbors m was set to 10 based on prior studies, ensuring a balance between noise resistance and computational feasibility.

To evaluate robustness, sensitivity analysis was performed by varying k and measuring its impact on classification accuracy and subset size. The results showed that small deviations in k (e.g., ±10%) did not significantly affect accuracy, highlighting the stability of the 3SGS approach. Future studies could explore a broader range of parameters or employ automated tuning methods, such as grid search or Bayesian optimization, to refine parameter selection further.

E. Classification Methods

Five widely recognized classifiers were employed to evaluate the effectiveness of the 3SGS approach, each with unique strengths in handling high-dimensional, low-sample datasets typical of microarray gene expression data. These classifiers were selected for their diverse methodological foundations, offering a robust basis for assessing the impact of gene selection on cancer classification performance.

- K-Nearest Neighbors (KNN) [22, 23] simplicity and adaptability to high-dimensional spaces make it well-suited for gene expression data, where inter-sample similarity is crucial for accurate classification.

- Support Vector Machine (SVM) [24] effectively distinguishes between gene expression patterns by maximizing the margin between classes, achieving high classification accuracy even with limited sample sizes.
- Linear Discriminant Analysis (LDA) [25] is particularly advantageous for microarray gene expression data, as it enhances interpretability by selecting a discriminative axis that captures maximum variance between cancer classes.
- Decision Tree (DT) [26, 27]: In microarray gene expression analysis, DT provides a clear framework to identify key genes that contribute to classification, making it well-suited for selecting and analyzing a compact, informative subset of genes.
- Naïve Bayes (NB) [26] classifies samples based on prior probabilities and likelihoods calculated from gene expression levels, providing a computationally efficient alternative that remains robust in the face of high-dimensional gene subsets.

Table I presents the final values for each classifier after hyperparameter tuning. It reflects the settings that were chosen to balance model performance and computational efficiency, ensuring optimal results on the cancer classification task.

TABLE I. HYPERPARAMETERS FOR KNN, SVM, LDA, NB, AND DC FOR LEUKEMIA AND PROSTATE CANCER

Classifier	Hyperparameter	Final Value
KNN	n_neighbors	5
	metric	'euclidean'
	weights	'distance'
	algorithm	'auto'
SVM	kernel	'rbf'
	C	1.0
	gamma	'scale'
LDA	solver	'lsqr'
	shrinkage	'auto'
NB	var_smoothing	1e-9 (GaussianNB)
DT	criterion	'gini'
	max_depth	5
	min_samples_split	10
	min_samples_leaf	5

Each classifier was trained and validated using the 3SGS-selected gene subsets, providing a comprehensive evaluation of classification accuracy, efficiency, and model interpretability. This multiclassifier approach ensures that the 3SGS method's impact on performance is rigorously evaluated across diverse classification paradigms, highlighting its generalizability and robustness for microarray cancer classification.

Classification accuracy [28] was used to assess classifier performance.

$$Accuracy = 100 * \frac{TP + TN}{TN + TP + FN + FP} \quad (5)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

F. Datasets

Microarray datasets are often represented as an $N \times M$ matrix, where N is the number of samples and M is the number of genes. Each value in the matrix represents the expression level of a certain gene in a given sample. This study employed two popular binary-class microarray cancer datasets (Table II).

TABLE II. MICROARRAY DATASETS

Datasets	Genes	Class	Samples data	Train data	Test data
Leukemia	7129	2	72	38	34
Prostate	12600	2	136	102	34

The leukemia dataset was taken from a collection of leukemia patient samples [29]. This dataset often serves as a benchmark for microarray analysis methods. It contains gene expressions corresponding to Acute Lymphoblast Leukemia (ALL) and Acute Myeloid Leukemia (AML) samples from bone marrow and peripheral blood. The dataset consisted of 72 samples: 49 samples of ALL and 23 samples of AML. Each sample is measured over 7,129 genes. The data are in a matrix with 7130 rows (7129 rows show the gene expressions, while the last row reports the corresponding sample's class label) and 72 columns representing the samples. The dataset can be found in [30].

The prostate cancer dataset was derived from the microarray analysis in [31]. It provides gene expression data for 12,600 genes, organized into distinct training and testing subsets. The training dataset includes 102 samples, 52 from patients diagnosed with prostate tumors and 50 from normal specimens, offering a solid basis for developing classification models. The test dataset comprises 34 samples, including 25 from prostate tumor patients and 9 from normal specimens, serving as an independent evaluation set to validate the model's predictive performance. This dataset is well-documented and publicly accessible, with detailed information about its structure, preprocessing steps, and usage guidelines available in [32].

G. Workflow: 3SGS Method for Cancer Classification

The process begins by classifying genes according to class separability using filter-based methods (SNR, CC, and ReliefF). An iterative, accuracy-driven selection process follows, retaining genes that improve classifier performance. In the final stage, redundancy is reduced by eliminating genes with minimal classification value, ensuring a minimal yet highly informative subset.

1) Dataset Preparation

- Select and preprocess two microarray datasets: leukemia and prostate cancer.
- Apply variance filtering to remove low-variance genes, reducing noise.

2) 3SGS Process

a) Stage 1: Filter-Based Selection

- Use SNR, CC, and ReliefF to rank genes by relevance.
- Select the top k genes for further analysis.

b) Stage 2: Accuracy-Driven Selection

- Evaluate classifier accuracy iteratively for gene subsets.

- Retain genes that improve or maintain accuracy to create a refined subset.

c) Stage 3 - Redundancy Reduction

- Analyze correlations or mutual information among genes.
- Remove redundant genes to finalize a minimal, predictive subset.

3) Classification

- Train classifiers on the selected gene subset.
- Evaluate performance using cross-validation and test datasets.

4) Results Analysis

- Compare classification accuracy and gene subset size with other methods.
- Discuss computational efficiency and biological relevance of selected genes.

III. RESULTS AND DISCUSSION

A. Data Preprocessing

Effective data preprocessing is essential in microarray analysis to reduce noise, ensure data quality, and manage high dimensionality. In the leukemia dataset, genes with low variance across samples were filtered out using a combination of thresholding, filtering, and logarithmic transformation techniques [33]. Thresholding kept values between 100 and 16000. Filtering tried to exclude genes whose expression is uniform and whose variation is insignificant in comparison to expression level measurements.

For each gene, the minimum value S_{min} and highest value S_{max} of the expression level were examined, and only genes with $S_{max}/S_{min} > 5$ and $S_{max} - S_{min} > 500$ were maintained. The data was then transformed logarithmically. This preprocessing resulted in a modest reduction in microarray data, from 7129 to 3051 genes.

B. Experimental Tools

MATLAB was used, which ran on a laptop with an Intel Core i5 CPU M 250 @ 2.4GHz. Each experiment took between 30 seconds and one minute to execute.

C. Classifier Performance Using the 3SGS Strategy

The 3SGS strategy, integrating filter-based and wrapper-based selection approaches, was applied to the leukemia and prostate cancer datasets. Five widely used classifiers were assessed: KNN, SVM, LDA, DT, and NB. Each classifier was evaluated based on accuracy and the number of selected genes. The feature selection methods compared include SNR, CC, and ReliefF, both independently and within the 3SGS framework.

1) Performance on the Leukemia Dataset

Table III displays the performance of various classifiers on the leukemia dataset, comparing traditional feature selection methods with the integrated 3SGS approach. The 3SGS strategy consistently achieved higher accuracy, reaching 100% with only 3-4 genes across several classifiers. The 3SGS approach outperformed traditional methods in accuracy and reduced the gene count required for high classification performance, identifying crucial markers (M27891, M23197, Y00787) for distinguishing between ALL and AML.

TABLE III. PERFORMANCE OF CLASSIFIERS WITH DIFFERENT FEATURE SELECTION METHODS FOR LEUKEMIA CANCER

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

2) Performance on the Prostate Cancer Dataset

Table IV summarizes the results on the prostate cancer dataset, where the 3SGS approach identified 3-4 genes, yielding classifier accuracies between 91% and 98%.

TABLE IV. PERFORMANCE OF VARIOUS CLASSIFIERS WITH DIFFERENT FEATURE SELECTION METHODS FOR PROSTATE CANCER

Feature selection methods	Classifiers				
	KNN	**SVM**	**LDA**	**DT**	**NB**
SNR	90% (22)	92% (8)	92% (4)	91% (19)	91% (45)
SNR_ 3SGS	98% (3)	98% (2)	97% (1)	92% (3)	92% (4)
CC	85% (6)	92% (44)	92% (6)	92% (46)	91% (65)
CC_3SGS	92% (4)	98% (3)	98% (3)	95% (4)	92% (3)
ReliefF	90% (32)	92% (34)	91% (75)	90% (36)	92% (50)
ReliefF_3SGS	95% (3)	95% (3)	91% (1)	91% (4)	95% (4)

The findings illustrate that the 3SGS method reduces the required gene subset size and consistently improves accuracy across classifiers, enhancing interpretability and computational efficiency. It achieved the highest classification accuracy using only three key discriminative genes (37720_at, 37639_at, 40435_at).

D. Discussion

Several methods have achieved notable results in leukemia and prostate cancer classification. For instance, in [16], Isomap-based dimensionality reduction was combined with genetic algorithms, achieving 100% accuracy for leukemia but requiring 43 genes, making it computationally intensive and less interpretable compared to the 3SGS approach. Similarly, in [17], perfect accuracy was achieved for leukemia using seven genes. However, this method's reliance on multi-objective

optimization increases implementation complexity. In [34], 100% accuracy was achieved for leukemia, but this method required 15 genes and leveraged fuzzy systems, which may complicate interpretability. In [35], 100% accuracy was achieved for leukemia using 10 genes, but this method lacks the accuracy-driven refinement and redundancy reduction included in 3SGS.

For prostate cancer, methods such as MC-FE combined with PCA [36] achieved high accuracy (98%) but relied on PCA for dimensionality reduction, which can reduce interpretability by transforming features into principal components. Self-regularized Lasso [37] achieved 97% accuracy but required careful tuning of regularization parameters, which may limit scalability across datasets. Ensemble-based methods focused on tree features [38] also reached 97% accuracy but were computationally demanding and involved complex tree-based ensemble strategies, potentially increasing training time and complexity.

The 3SGS approach achieved competitive performance with 100% accuracy for leukemia and 98% for prostate cancer using only 3-4 genes. Unlike the referenced methods, 3SGS balances simplicity and interpretability by combining filter-based methods (SNR, CC, ReliefF) with accuracy-driven selection and redundancy reduction. This allows the identification of minimal but highly predictive gene subsets, reducing computational demands while maintaining high performance. Integrating linear (SNR, CC) and non-linear (ReliefF) techniques ensures robustness across diverse datasets without the need for ensemble models or dimensionality reduction techniques such as PCA. These strengths place 3SGS as a scalable, efficient, and interpretable solution for precision oncology, advancing the field of cancer classification through innovative feature selection.

Despite its strengths, several limitations merit attention. First, the limited sample size raises concerns regarding overfitting. Future studies should validate 3SGS on larger, more diverse datasets to evaluate its generalizability. Second, the current application of 3SGS is limited to binary classification tasks. Expanding this framework to handle multi-class cancer classification will enhance its utility in complex biomedical contexts. Third, although the identified genes exhibit strong statistical relevance, biological validation is essential to confirm their roles as biomarkers. Therefore, collaborations with biomedical researchers can provide deeper insights into their clinical significance. Finally, although 3SGS enhances computational efficiency, applying it to larger or multi-omics datasets may present challenges. Future research should explore parallel computing or distributed frameworks to improve scalability.

IV. CONCLUSION

This study introduced the 3SGS method, a novel framework that combines filter-based techniques, accuracy-driven refinement, and redundancy reduction to address key challenges in gene selection. This approach differs from prior methods, such as those relying on genetic algorithms, ensemble models, or dimensionality reduction techniques such as PCA, by achieving similar or superior performance with far fewer genes and reduced computational demands. For example, while methods like Isomap-based genetic algorithms and XGBoost-based multi-objective optimization require 10-43 genes to achieve 100% accuracy for leukemia, 3SGS achieves the same result with only 3-4 genes. Furthermore, the interpretability and efficiency of 3SGS make it particularly advantageous compared to ensemble-based approaches, which are often computationally intensive. Thus, this novel method provides a practical, scalable, and high-performance solution for cancer classification.

Future research should focus on validating these findings across more extensive and varied datasets to assess the generalizability of 3SGS and its potential to handle multi-class classifications. Additionally, integrating deep learning techniques with 3SGS-selected features may further improve performance, especially in complex and high-dimensional settings such as multi-omics. The proposed strategy also encourages collaborations with clinical researchers to explore the biological relevance of selected genes, potentially enhancing precision oncology.

In summary, the 3SGS strategy represents a robust, scalable, and efficient solution for cancer classification, poised to advance precision medicine through improved diagnostic workflows in cancer genomics.

REFERENCES

[1] H. Z. Almarzouki, "Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile," *Journal of Healthcare Engineering*, vol. 2022, no. 1, 2022, Art. no. 4715998, https://doi.org/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, https://doi.org/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, Oct. 2023, Art. no. 102786, https://doi.org/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, Jun. 2022, Art. no. 101847, https://doi.org/10.1016/j.jksus.2022.101847.

[5] N. Behar and M. Shrivastava, "A Novel Model for Breast Cancer Detection and Classification," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9496–9502, Dec. 2022, https://doi.org/10.48084/etasr.5115.

[6] S. Larabi 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, https://doi.org/10.1016/j.jksuci.2018.06.004.

[7] A. E. Hegazy, M. A. Makhlouf, 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, https://doi.org/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, Feb. 2023, Art. no. 562, https://doi.org/10.3390/pr11020562.

[9] R. Dash, "An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 2, pp. 195–207, Feb. 2021, https://doi.org/10.1016/j.jksuci.2018.02.013.

[10] L. Moody, H. Chen, and Y.-X. Pan, "Considerations for feature selection using gene pairs and applications in large-scale dataset integration, novel oncogene discovery, and interpretable cancer screening," *BMC Medical Genomics*, vol. 13, no. 10, Oct. 2020, Art. no. 148, https://doi.org/10.1186/s12920-020-00778-x.

[11] Uzma and Z. Halim, "An ensemble filter-based heuristic approach for cancerous gene expression classification," *Knowledge-Based Systems*, vol. 234, Dec. 2021, Art. no. 107560, https://doi.org/10.1016/j.knosys.2021.107560.

[12] A. Benkessirat and N. Benblidia, "A novel feature selection approach based on constrained eigenvalues optimization," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4836–4846, Sep. 2022, https://doi.org/10.1016/j.jksuci.2021.06.017.

[13] M. K. P. Niyas and P. Thiyagarajan, "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, Sep. 2022, https://doi.org/10.1016/j.jksuci.2020.12.009.

[14] S. Bose *et al.*, "An ensemble machine learning model based on multiple filtering and supervised attribute clustering algorithm for classifying cancer samples," *PeerJ Computer Science*, vol. 7, Sep. 2021, Art. no. e671, https://doi.org/10.7717/peerj-cs.671.

[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*, vol. 8, no. 10, pp. 106–116, 2020.

[16] 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, Apr. 2023, Art. no. 139, https://doi.org/10.1186/s12859-023-05267-3.

[17] 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, https://doi.org/10.1007/s11517-021-02476-x.

[18] 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, Aug. 2004, https://doi.org/10.1109/TKDE.2004.68.

[19] J. Hou *et al.*, "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, no. 1, Feb. 2022, Art. no. 81, https://doi.org/10.1186/s12859-022-04609-x.

[20] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, Dec. 1997, https://doi.org/10.1016/S0004-3702(97)00043-X.

[21] 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, https://doi.org/10.1049/el.2015.3613.

[22] 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, https://doi.org/10.1109/ACCESS.2021.3105485.

[23] 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, https://doi.org/10.1016/j.jksuci.2021.05.004.

[24] 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, Jun. 2023, Art. no. 101565, https://doi.org/10.1016/j.jksuci.2023.101565.

[25] 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, https://doi.org/10.32604/cmc.2022.029039.

[26] 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)*, Marrakesh, Morocco, Apr. 2024, pp. 1–5, https://doi.org/10.1109/GAST60528.2024.10520752.

[27] 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." Social Science Research Network, May 06, 2024, https://doi.org/10.2139/ssrn.4811438.

[28] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazanci, and O. Oral, "Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture," *Journal of King Saud University - Science*, vol. 35, no. 6, Aug. 2023, Art. no. 102754, https://doi.org/10.1016/j.jksus.2023.102754.

[29] T. R. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999, https://doi.org/10.1126/science.286.5439.531.

[30] "Gene expression dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/crawford/gene-expression.

[31] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, Mar. 2002, https://doi.org/10.1016/S1535-6108(02)00030-2.

[32] "GNF Prostate Data." https://www.stat.cmu.edu/~jiashun/Research/software/HCClassification/Prostate/Readme.txt.

[33] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Systems with Applications*, vol. 213, Mar. 2023, Art. no. 118946, https://doi.org/10.1016/j.eswa.2022.118946.

[34] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification," *PLOS ONE*, vol. 10, no. 3, 2015, Art. no. e0120364, https://doi.org/10.1371/journal.pone.0120364.

[35] 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, Mar. 2005, Art. no. 76, https://doi.org/10.1186/1471-2105-6-76.

[36] A. Razzaque and D. A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Measurement: Sensors*, vol. 31, Feb. 2024, Art. no. 100945, https://doi.org/10.1016/j.measen.2023.100945.

[37] M. Vatankhah and M. Momenzadeh, "Self-regularized Lasso for selection of most informative features in microarray cancer classification," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5955–5970, Jan. 2024, https://doi.org/10.1007/s11042-023-15207-1.

[38] G. Dagnew and B. h. Shekar, "Ensemble learning-based classification of microarray cancer data on tree-based features," *Cognitive Computation and Systems*, vol. 3, no. 1, pp. 48–60, 2021, https://doi.org/10.1049/ccs2.12003.