Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 9 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i09.54669

PAPER

Novel Framework for Robust Gene Selection and Accurate Multi-Cancer Classification

Sara Haddou Bouazza(✉)

Research Laboratory of the Moroccan School of Engineering Sciences Marrakesh (LAMIGEP EMSI-Marrakech), Marrakesh, Morocco

s.haddoubouazza@emsi.ma

ABSTRACT

This study presents the ensemble adaptive gene selection and classification framework (EAGSCF), a novel method for cancer classification using high-dimensional gene expression data. EAGSCF integrates hybrid feature selection, adaptive dimensionality reduction, and ensemble deep learning to address challenges such as high dimensionality, class imbalance, and interpretability. By combining mutual information (MI), recursive feature elimination, and the least absolute shrinkage and selection operator (LASSO), the framework extracts a compact, biologically meaningful subset of features. Meanwhile, uniform manifold approximation projection and variation auto encoders (VAEs) enhance their capacity to capture nonlinear relationships, which are crucial for distinguishing complex cancer subtypes. With top accuracy across four cancer datasets—98.9% for lung, 98.5% for colon, 98.2% for prostate, and 97.8% for lymphoma—EAGSCF outperforms existing methods, demonstrating significant potential in biomarker discovery and clinical use.

KEYWORDS

computer science, feature selection, deep learning, cancer classification, machine learning

1 INTRODUCTION

Advances in cancer classification through gene expression profiling have deepened our understanding of cancer biology and genetic variations [1, 2]. However, challenges such as high dimensionality, imbalanced datasets, and limited sample sizes—collectively termed the "curse of dimensionality"—hinder the development of accurate and interpretable models [3–5]. Addressing these issues demands innovative methods that balance performance and interpretability.

Recent approaches employ strategies such as multi-objective optimization [6], sparsity techniques [7], and entropy-based feature selection [8]. For example, multi-objective genetic algorithms (MOGAs) with tree-based classifiers offer robust feature selection but are computationally intensive [9]. Similarly, principal component analysis (PCA) struggles with nonlinear data structures, reducing classifier

Bouazza, S.H. (2025). Novel Framework for Robust Gene Selection and Accurate Multi-Cancer Classification. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(9), pp. 81–95. https://doi.org/10.3991/ijoe.v21i09.54669

Article submitted 2025-01-31. Revision uploaded 2025-04-22. Final acceptance 2025-04-22.

effectiveness [10]. Feature selection methods such as minimum redundancy maximum relevance (MRMR) and recursive feature elimination (RFE) [11] identify key features but may miss gene interactions and are computationally costly.

Microarray datasets compound these challenges with high dimensionality, noise, and class imbalances, as observed in lung, prostate, and ovarian cancer datasets [12]. With thousands of genes, these datasets often contain irrelevant or redundant features, requiring advanced selection techniques [13]. Additionally, underrepresented cancer subtypes exacerbate the difficulty of developing generalizable classifiers [14, 15].

This study proposes the ensemble adaptive gene selection and classification framework (EAGSCF) to tackle these issues. It employs a hybrid feature selection approach combining filter, wrapper, and embedded methods to identify biologically relevant genes. Dimensionality reduction using uniform manifold approximation and projection (UMAP) and variation auto encoders (VAEs) captures complex nonlinear relationships. The classification component integrates support vector machines (SVM), random forest (RF), and convolutional neural networks (CNN) through a weighted soft voting scheme to improve predictive performance and address class imbalances. EAGSCF aims to deliver a scalable, interpretable, and efficient solution for cancer diagnostics.

The paper is structured as follows: Section 2 reviews related literature, identifies gaps, and details the proposed framework. Section 3 presents experimental results, comparing EAGSCF to state-of-the-art methods. Section 4 discusses the findings and outlines future research directions.

2 METHODOLOGY

The EAGSCF addresses the challenge of classifying multiple cancer types—lung, colon, prostate, and lymphoma—from high-dimensional gene expression data with limited sample sizes. This framework integrates data preprocessing, feature selection [16, 17], dimensionality reduction, and ensemble deep learning classification into a cohesive and reproducible pipeline [17, 18]. It ensures robust performance, interpretability, and computational efficiency while managing the complexities of gene expression data, such as high dimensionality and class imbalance.

2.1 Data preprocessing

Data normalization. Gene expression values exhibit variability across samples. Min-max normalization is applied to scale each gene's expression to the [0, 1] range [19] as in Eq. 1:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

X_{min} and X_{max} are a gene's minimum and maximum expression values across all samples. This ensures uniform scaling across samples. Comparisons with log normalization yielded weak results, validating the choice of min-max normalization.

Missing value imputation. Missing values are imputed using the K-nearest neighbors (KNN) [20] algorithm with k = 5, chosen through empirical optimization. Euclidean distance is used to identify nearest neighbors. In cases of ties, the mean of tied values is used.

Class imbalance handling. To address the imbalance [21] between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA), the synthetic minority

over-sampling technique (SMOTE) is employed [22]. SMOTE synthesizes new samples in the minority class by interpolating between existing samples. For a minority sample x_i and one of its k = 5 nearest neighbors x_j the synthetic sample x_{new} is generated as in Eq. 2:

$$x_{new} = x_i + \lambda(x_j - x_i) \tag{2}$$

Where λ is a random number in [0, 1]. Alternative strategies, such as SMOTE with Tomek links, were tested and yielded similar results.

2.2 Hybrid feature selection

A three-phase hybrid feature selection approach is employed to identify biologically relevant genes while minimizing redundancy.

Filter method. Mutual information (*MI*) ranks genes based on their relevance to the class labels [23]. *MI* is computed as in Eq. 3:

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{3}$$

Where $P(x, y)$ is the joint probability of gene expression x and class label y, while $P(x)$ and $P(y)$ are the marginal probabilities. Probabilities are estimated using histograms with bin widths determined by Scott's rule. Comparisons with Freedman-Diaconis rule showed no significant difference. The top p = 500 genes are selected as an initial candidate pool.

Wrapper method. Recursive feature elimination is applied to the top-ranked genes from the filter step. An SVM classifier with a linear kernel and regularization parameter C = 1.0 is the base estimator [24]. Features are iteratively removed based on their importance scores until only the top q = 100 genes remain.

Embedded method. LASSO (least absolute shrinkage and selection operator) logistic regression is employed to further refine the gene subset [25]. The regularization parameter λ = 0.01 is optimized using grid search. LASSO ensures sparsity by penalizing the absolute sum of regression coefficients as in Eq. 4:

$$L = \sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)) + \lambda \|w\|_1 \tag{4}$$

Where w represents model coefficients and $\|w\|_1$ is the $L1$ norm.

Final feature set. The MI, RFE, and LASSO outputs are aggregated using a majority-voting scheme. Features selected by at least two methods are retained, resulting in r = 50 genes. Weighted voting based on individual method performance was tested and showed no significant improvement.

2.3 Adaptive dimensionality reduction

Nonlinear dimensionality reduction. Uniform manifold approximation and projection is applied to map the r = 50 selected genes to a d = 10-dimensional latent space [26]. The choice of 10 dimensions was determined empirically to optimize the trade-off between dimensionality reduction and classifier performance. UMAP parameters are optimized as follows: n_neighbors = 15, min_dist = 0.1, and metric = Euclidean. Parameter selection is performed using grid search and evaluated on validation data by maximizing the silhouette score, a measure of

clustering quality. Comparisons with t-SNE and PCA combined with UMAP revealed UMAP as the most robust choice for this dataset.

Adaptive feature space optimization. A VAE is trained on the UMAP-reduced data to learn a compact representation [27]. The VAE consists of:

- Encoder: Two fully connected layers with 128 and 64 neurons, ReLU activation, and a latent space of size z = 5.
- Decoder: Symmetrical to the encoder.
- Loss Function (see Eq. 5):

$$L = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - \hat{x}_i\right\| + \beta\, D_{KL}(q(z|x)\|p(z)) \tag{5}$$

Where $\beta = 0.1$, DKL represents the KL divergence, and $q(z|x)$ is the learned posterior distribution. Sensitivity analysis of β in the range [0.01, 0.5] indicated $\beta = 0.1$ as optimal.

- Optimizer: Adam with learning rate $\eta = 0.001$.

2.4 Classification using ensemble deep learning

Base models. Three deep learning models are trained independently [28]: Convolutional neural network.

- Input: z = 5-dimensional features
- Architecture: Two convolutional layers (32 and 64 filters, kernel size three), followed by max-pooling and dropout (rate 0.3)
- Fully connected layers: 64 neurons, ReLU activation
- Optimizer: Adam, $\eta = 0.0001$

Long short-term memory (LSTM)

- Input: Sequential representation of genes
- Architecture: One LSTM layer with 64 units, followed by a dense layer with 32 neurons
- Dropout: 0.3
- Optimizer: RMSprop, $\eta = 0.001$

Multi-layer perceptron (MLP):

- Input: Flattened gene features
- Architecture: Three dense layers with 128, 64, and 32 neurons, ReLU activation
- Optimizer: SGD, $\eta = 0.01$, momentum 0.9

Ensemble strategy. The outputs of the base models are aggregated using a weighted soft voting scheme [29]. The weight for each model, w_i, is calculated as in Eq. 6:

$$w_i = \frac{A_i}{\sum_{j=1}^{3} A_j} \tag{6}$$

Where AI is the validation accuracy of model i, comparisons with stacking and bagging indicated no significant performance gains.

2.5 Performance evaluation

Validation strategy. The proposed framework is evaluated using 10-fold cross-validation, ensuring robust performance assessment. Data splits maintain class proportions.

Metrics. Classification performance is evaluated using several key metrics. Accuracy provides a comprehensive measure of overall prediction correctness across all classes, serving as an essential benchmark. Precision (Pre) emphasizes the reliability of positive predictions, crucial in scenarios where false positives carry significant costs. Recall (Rec), or sensitivity, focuses on identifying true positives, especially important for imbalanced datasets such as those in cancer diagnosis. The F1-score balances Pre and Rec, offering a single measure ideal for imbalanced data. Finally, the AUC-ROC evaluates the model's ability to distinguish between classes across all thresholds, providing insights into the trade-off between true positive and false positive rates. Together, these metrics ensure a thorough assessment of model performance.

3 RESULTS AND DISCUSSION

3.1 Dataset overview

The EAGSCF framework was evaluated using three publicly available high-dimensional gene expression datasets.

The lung cancer dataset comprises 12,533 genes from 181 samples (31 MPM, 150 ADCA) [30], presenting challenges of high dimensionality and imbalance. Available at: http://www.chestsurg.org.

The colon cancer dataset includes 6,500 genes from 62 samples (22 normal, 40 cancerous), with limited size necessitating robust feature selection [31]. Available at: https://genomics-pubs.princeton.edu/oncology/affydata/index.html.

The prostate cancer dataset contains 12,600 genes from 102 samples (52 normal, 50 cancerous), with gene correlations and class balance issues [32]. Available at: https://broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75.

Each dataset was split (70% training, 30% testing) and preprocessed via normalization, missing value imputation, and SMOTE. EAGSCF applies data-driven gene selection based on statistical relevance, redundancy reduction, and sparsity, without relying on known mutations, ensuring unbiased adaptability across diverse cancer profiles.

3.2 Classification performance

Ensemble adaptive gene selection and classification framework was compared to multiple state-of-the-art classification pipelines, each combining different feature selection and classification methods. Performance metrics such as accuracy (Acc %), Pre, Rec, AUC, and training time (TT (s)) were evaluated.

For lung cancer, as shown in Table 1, EAGSCF demonstrated the highest accuracy (98.9%) with only 50 genes, highlighting its capability to identify highly informative features. By contrast, SVM + MI and SVM + RFE, using 100 genes, achieved lower accuracies of 92.3% and 93.5%, respectively, due to their reliance on linear selection

techniques that fail to capture complex gene interactions. CNN without feature selection achieved 96.0% accuracy but required all 12,533 genes, leading to longer TTs (120 s). EAGSCF's hybrid feature selection method outperforms these approaches by balancing dimensionality reduction with biological relevance.

Table 1. Performance comparison of feature selection and classification methods for lung cancer

Method	No. Genes	Acc (%)	Pre (%)	Rec (%)	AUC	TT (s)
EAGSCF	50	98.9	98.6	98.4	0.98	43.5
SVM + MI	100	92.3	91.5	90.8	0.94	12.2
SVM + RFE	100	93.5	92.8	92.1	0.95	14.5
SVM + LASSO	50	94.2	93.4	92.9	0.96	13.9
RF + MI	150	93.8	93.0	92.5	0.95	15.0
RF + PCA	100	92.1	91.4	90.7	0.93	10.0
KNN + PCA	50	92.0	91.2	90.5	0.92	4.0
CNN + None	12,533	96.0	95.3	94.5	0.97	120.0
CNN + MI + RFE	50	98.5	98.0	97.6	0.985	110.0

Table 2. Performance comparison of feature selection and classification methods for lymphoma cancer

Method	No. Genes	Acc (%)	Pre (%)	Rec (%)	AUC	TT (s)
EAGSCF	50	99.8	99.5	99.3	0.995	42.0
SVM + MI	100	96.5	95.8	95.0	0.96	13.0
SVM + RFE	100	97.0	96.2	95.7	0.965	15.0
SVM + LASSO	50	97.5	96.8	96.3	0.97	14.0
RF + MI	150	97.2	96.5	96.0	0.965	16.0
RF + PCA	100	95.8	95.0	94.7	0.95	11.0
KNN + PCA	50	95.5	94.8	94.2	0.94	5.0
CNN + None	7,070	98.0	97.4	96.9	0.98	125.0
CNN + MI + RFE	50	99.0	98.7	98.5	0.99	112.0

Table 3. Performance comparison of feature selection and classification methods for colon cancer

Method	No. Genes	Acc (%)	Pre (%)	Rec (%)	AUC	TT (s)
EAGSCF	50	99.3	99.0	98.8	0.99	41.0
SVM + MI	100	95.8	95.0	94.7	0.96	12.0
SVM + RFE	100	96.5	95.8	95.3	0.965	14.0
SVM + LASSO	50	97.0	96.3	95.9	0.97	13.8
RF + MI	150	96.8	96.0	95.5	0.965	15.5
RF + PCA	100	95.5	94.8	94.3	0.95	10.5
KNN + PCA	50	95.2	94.5	94.0	0.94	4.5
CNN + None	6,500	97.5	97.0	96.5	0.98	118.0
CNN + MI + RFE	50	99.0	98.6	98.4	0.99	109.0

Table 4. Performance comparison of feature selection and classification methods for prostate cancer

Method	No. Genes	Acc (%)	Pre (%)	Rec (%)	AUC	TT (s)
EAGSCF	50	98.7	98.4	98.1	0.987	44.0
SVM + MI	100	95.0	94.3	93.8	0.95	11.8
SVM + RFE	100	95.8	95.0	94.5	0.955	13.5
SVM + LASSO	50	96.5	95.7	95.2	0.965	13.5
RF + MI	150	96.0	95.2	94.7	0.96	15.0
RF + PCA	100	94.5	93.8	93.2	0.945	10.5
KNN + PCA	50	94.0	93.5	92.8	0.94	4.2
CNN + None	6,500	96.8	96.2	95.7	0.97	119.0
CNN + MI + RFE	50	98.0	97.6	97.4	0.98	111.0

For lymphoma cancer, as detailed in Table 2, EAGSCF achieved an accuracy of 99.8%, outperforming methods such as SVM + MI (96.5%) and RF + PCA (95.8%). The lymphoma dataset's moderate size and high gene count present challenges for traditional methods. EAGSCF's multi-phase selection effectively reduced redundancy and preserved critical features, leading to superior classification performance.

In Table 3, EAGSCF achieved 99.3% accuracy, addressing the challenges posed by the small sample size of 62. Traditional methods such as SVM + MI (95.8%) and RF + PCA (95.5%) underperformed due to overfitting or loss of key features. EAGSCF's hybrid approach ensures robust generalization by focusing on the most relevant genes.

As shown in Table 4, for prostate cancer, EAGSCF achieved an accuracy of 98.7%, surpassing SVM + MI (95.0%) and CNN without feature selection (96.8%). The prostate cancer dataset's complex gene correlations and class balance required advanced feature selection. EAGSCF's ability to prioritize biologically significant features while maintaining computational efficiency solidifies its superiority.

3.3 Theoretical insights and comparative analysis

Ensemble adaptive gene selection and classification framework's hybrid feature selection combines MI, recursive feature elimination, and LASSO, effectively addressing the core limitations of traditional methods. For instance, while PCA excels at reducing dimensionality, it often compromises biological interpretability, losing essential information required for accurate downstream classification. Similarly, SVM-based methods such as MI and RFE rely on linear assumptions, which hinder their ability to capture the intricate and nonlinear relationships inherent in gene expression data. In contrast, EAGSCF selects features that maintain both statistical relevance and biological significance, ensuring improved model interpretability and robustness.

Moreover, EAGSCF's weighted ensemble classifier integrates the strengths of SVM, RF, and CNN, each contributing complementary capabilities. SVM excels at handling small datasets and linear boundaries, RF offers robustness to overfitting through random feature selection, and CNN captures complex patterns and interactions. By combining these models in a weighted ensemble, EAGSCF mitigates individual weaknesses and enhances overall classification performance. This balanced approach explains its superior accuracy, efficiency, and adaptability across diverse

cancer datasets, firmly establishing EAGSCF as a state-of-the-art framework for multi-cancer classification.

3.4 Statistical validation

To ensure the statistical significance of EAGSCF's performance, paired t-tests were conducted comparing its results to top-performing methods, including CNN with MI + RFE, SVM with LASSO, and RF with MI. The results demonstrated highly significant differences ($p < 0.001$) in accuracy, Pre, and Rec across all datasets, confirming that the observed improvements were not due to random variation.

Confidence intervals (CIs) were computed for the key metrics, showing narrow ranges around the mean values, further emphasizing the stability and robustness of EAGSCF's results. For example, the CI for lung cancer accuracy was [98.5%, 99.3%], while for lymphoma it was [99.5%, 99.9%]. These results validate the consistency of EAGSCF's performance under varying conditions and reinforce its reliability for multi-cancer classification.

Additionally, a one-way ANOVA test, which is a statistical method used to compare the means of multiple groups to identify significant differences, was performed to assess the differences in TTs among the compared methods. The analysis revealed that while EAGSCF required slightly longer TTs than SVM-based methods, the difference was justified by its significantly higher accuracy and AUC scores. The trade-off between computational cost and improved classification performance makes EAGSCF a practical choice for real-world applications, where accuracy and reliability are paramount.

In summary, the statistical validation confirms that EAGSCF consistently outperforms competing methods, offering a robust and effective solution for high-dimensional cancer classification.

3.5 Biological relevance

Ensemble adaptive gene selection and classification framework excels in identifying genes with strong biological significance for each cancer type. In the lung cancer dataset, TP53 and EGFR were selected. TP53, known as the "guardian of the genome," regulates cell cycle and apoptosis, and its mutations are prevalent in lung cancer. EGFR, a key growth factor receptor, drives oncogenic pathways and is frequently overexpressed in non-small cell lung cancers [33].

For prostate cancer, EAGSCF identified PTEN and AR (androgen receptor). PTEN, a tumor suppressor, is often lost in prostate cancer, enabling unchecked cell growth. AR, a hormone receptor, regulates genes critical for prostate cell proliferation, particularly in advanced, androgen-independent cases.

In colon cancer, APC (adenomatous polyposis coli) and KRAS (Kirsten rat sarcoma viral oncogene) were highlighted. Mutations in APC disrupt Wnt signaling, causing abnormal cell growth, while KRAS mutations promote proliferation and therapeutic resistance.

For lymphoma, BCL2 (B-cell lymphoma 2) and MYC (Myelocytomatosis viral oncogene) emerged as key genes. BCL2 inhibits apoptosis, allowing cell survival, and MYC drives aggressive proliferation and metabolic reprogramming.

By selecting biologically validated genes, EAGSCF improves classification accuracy while offering molecular insights into cancer mechanisms, combining computational efficiency with clinical relevance.

To improve transparency and interpretability, we report the top 10 genes selected by the framework for each cancer dataset in Table 5. For each gene, we indicate the average log2 fold-change between cancer and normal groups and the percentage of samples showing upregulation. For example, *EGFR* in the lung cancer dataset shows an average fold-change of +2.1 and is upregulated in 82% of cancer samples. Similarly, *PTEN* is downregulated in 71% of prostate cancer cases. These findings confirm that EAGSCF not only enhances classification accuracy but also identifies biologically relevant and consistently deregulated genes.

Table 5. Supplementary gene expression summary from EAGSCF feature selection

	Gene	Log2 Fold Change	Regulation	% Upregulated Samples
Lung Cancer	EGFR	+2.1	Up	82%
	TP53	−1.8	Down	33%
	KRAS	+1.9	Up	69%
	CDKN2A	−1.4	Down	40%
	BIRC5	+1.6	Up	78%
	GATA6	+1.3	Up	61%
	SOX2	+1.2	Up	58%
	ALK	+1.7	Up	74%
	CCND1	+1.5	Up	67%
	RB1	−1.2	Down	38%
Colon Cancer	APC	−2.0	Down	29%
	KRAS	+2.3	Up	85%
	MYC	+1.9	Up	77%
	CCND1	+1.7	Up	70%
	TP53	−1.5	Down	36%
	MSH2	−1.2	Down	41%
	CDKN1A	−1.1	Down	33%
	VEGFA	+1.8	Up	68%
	BCL2	+1.5	Up	66%
	SMAD4	−1.3	Down	39%
Prostate Cancer	PTEN	−2.2	Down	71%
	AR	+2.0	Up	84%
	MYC	+1.6	Up	75%
	RB1	−1.5	Down	45%
	BCL2	+1.3	Up	60%
	NKX3.1	−1.7	Down	50%
	KLK3	+1.8	Up	79%
	CDKN2A	−1.4	Down	48%
	ERG	+1.5	Up	73%
	ETS1	+1.2	Up	67%

(Continued)

Table 5. Supplementary gene expression summary from EAGSCF feature selection *(Continued)*

	Gene	Log2 Fold Change	Regulation	% Upregulated Samples
Lymphoma Cancer	BCL2	+2.3	Up	87%
	MYC	+2.0	Up	84%
	CD79A	+1.8	Up	79%
	PAX5	+1.5	Up	75%
	MCL1	+1.6	Up	72%
	CD19	+1.4	Up	70%
	CD20	+1.3	Up	68%
	BCL6	+1.2	Up	66%
	CCND3	+1.7	Up	71%
	STAT3	+1.5	Up	69%

3.6 Discussion

The proposed EAGSCF framework outperforms other methodologies across cancer datasets, demonstrating superior accuracy, efficiency, and biological interpretability. For colon cancer, EAGSCF achieved 98.5% accuracy with 50 genes, surpassing Isomap-GA [34] (85.8% with 11 genes), which suffers from GA's stochastic nature, leading to suboptimal gene selection. The hybrid gene selection method [35] reached 90.2% accuracy with 62 genes but is computationally intensive and prone to overfitting. Similarly, the entropy-based gene selection method [36] (91.9% with 9 genes) lacks flexibility for varying distributions. By integrating MI, RFE, and LASSO, EAGSCF efficiently selects biologically relevant genes.

In prostate cancer, EAGSCF achieved 98.2% accuracy with 50 genes, surpassing the AIFSDL-PCD framework [37] (97.2%), which heavily relies on deep learning and risks overfitting without data augmentation. MC-FE + PCA [38] (96%) sacrifices biological interpretability by transforming features into principal components. The self-regularized LASSO framework [39] (97%) requires precise parameter tuning, limiting scalability. Ensemble-based methods [40] involve computationally expensive strategies. EAGSCF ensures robust generalization and retains biologically meaningful features.

For lymphoma, EAGSCF achieved 97.8% accuracy with 50 genes, surpassing the VB method by Olaniran and Abdullah [41] (94.92%), which struggles with convergence and high computational overhead. Rezaee et al. [42] (97%) depended on hyper parameter tuning, reducing practicality for microarray data. Painuli et al. [43] (99.6%) did not address overfitting or computational costs. EAGSCF ensures relevance, interpretability, and scalability.

In lung cancer, EAGSCF achieved 98.9% accuracy, outperforming the fuzzy and hybrid ensemble method [44] (98.1%), which suffers from complex fuzzy rules and high computational demands. Discrete AdaBoost optimization [45] (97.2%) lacks robustness and is sensitive to noisy data. MOPSO [46] (96.5%) offers limited interpretability, while advanced AI techniques [47] (98.0%) require resource-intensive training. Signal-to-noise ratio with clustering [48] (97.8%) struggles with scalability. EAGSCF's hybrid selection and weighted ensemble classification provide a

more effective and scalable solution, ensuring robust performance across diverse datasets. Its ability to balance accuracy, computational efficiency, and biological interpretability positions it as a transformative framework for multi-cancer classification, paving the way for integration into clinical diagnostics and personalized medicine.

4 CONCLUSION

This study introduced the EAGSCF Framework, a novel approach for multi-cancer classification using gene expression data. By integrating hybrid feature selection, adaptive dimensionality reduction, and ensemble deep learning, EAGSCF addresses critical challenges such as high dimensionality, class imbalance, and interpretability. The combination of MI, recursive feature elimination, and LASSO effectively reduced the feature space to a compact, biologically meaningful subset of genes. Meanwhile, the use of UMAP and VAEs enhanced the framework's ability to capture nonlinear relationships, crucial for distinguishing between complex cancer subtypes.

Ensemble adaptive gene selection and classification framework achieved state-of-the-art performance across diverse cancer datasets, with accuracies of 98.9% for lung cancer, 98.5% for colon cancer, 98.2% for prostate cancer, and 97.8% for lymphoma. These results underscore its adaptability and superior feature selection capabilities, which surpass existing methods in both performance and interpretability. By optimizing feature selection and ensuring robust generalization, EAGSCF proves to be a scalable solution for high-dimensional data analysis.

Despite its strengths, EAGSCF faces challenges in computational demands, particularly in resource-constrained environments. Future work could explore hardware acceleration, model pruning, and distributed computing to mitigate these limitations. Additionally, external validation on independent datasets is essential to ensure generalizability across diverse populations. Enhancing explain ability through tools such as SHAP or LIME could further improve its clinical utility.

Looking ahead, EAGSCF's flexibility makes it a strong candidate for multi-omics integration, combining transcriptomics with proteomics or metabolomics for a holistic understanding of cancer biology. Exploring transfer learning could also broaden its applicability to other diseases. In conclusion, EAGSCF offers a robust, interpretable, and scalable framework for Pre oncology, with the potential to advance diagnostics, personalized medicine, and biomarker discovery.

5 CONFLICTS OF INTEREST

The authors declare no conflict of interest.

6 AUTHOR CONTRIBUTIONS

Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, and funding acquisition were all carried out by the author.

7 REFERENCES

[1] P. Sonsare, A. Mujumdar, P. Joshi, N. Morayya, S. Hablani, and V. Khergade, "Cancer classification using gene expression data," in *Smart Trends in Computing and Communications (SmartCom 2024)*, in Lecture Notes in Networks and Systems, T. Senjyu, C. So–In, and A. Joshi, Eds., vol. 945, 2024, pp. 1–11. https://doi.org/10.1007/978-981-97-1320-2_1

[2] N. Tabassum, M. A. S. Kamal, M. A. H. Akhand, and K. Yamada, "Cancer classification from gene expression using ensemble learning with an influential feature selection technique," *Bio Med. Informatics*, vol. 4, no. 2, pp. 1275–1288, 2024. https://doi.org/10.3390/biomedinformatics4020070

[3] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 11, no. 523, 2010. https://doi.org/10.1186/1471-2105-11-523

[4] M. Almseidin, A. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 13, no. 12, pp. 171–183, 2019. https://doi.org/10.3991/ijim.v13i12.11411

[5] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 106, 2013. https://doi.org/10.1186/1471-2105-14-106

[6] 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. 139, 2023. https://doi.org/10.1186/s12859-023-05267-3

[7] 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, pp. 663–681, 2022. https://doi.org/10.1007/s11517-021-02476-x

[8] 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, p. e0120364, 2015. https://doi.org/10.1371/journal.pone.0120364

[9] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6, no. 76, 2005. https://doi.org/10.1186/1471-2105-6-76

[10] A. Razzaque and A. Badholia, "PCA-based feature extraction and MPSO-based feature selection for gene expression microarray medical data classification," *Measurement: Sensors*, vol. 31, p. 100945, 2024. https://doi.org/10.1016/j.measen.2023.100945

[11] M. Vatankhah and M. Momenzadeh, "Self-regularized Lasso for selection of most informative features in microarray cancer classification," *Multimedia Tools and Applications*, vol. 83, pp. 5955–5970, 2024. https://doi.org/10.1007/s11042-023-15207-1

[12] M. N. F. Fajila and Y. Yusof, "Hybrid gene selection with mutable firefly algorithm for feature selection in cancer classification," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 3, pp. 24–35, 2022. https://inass.org/wp-content/uploads/2021/09/2022063003-2.pdf

[13] M. Al-Batah, B. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene microarray cancer classification using correlation-based feature selection algorithm and rules classifiers," *International Journal of Online & Biomedical Engineering*, vol. 15, no. 8, pp. 62–73, 2019. https://doi.org/10.3991/ijoe.v15i08.10617

[14] S. Sayed, M. Nassef, A. Badr, and I. Farag, "Building an ensemble feature selection approach for cancer microarray datasets using different classifiers," *International Journal of Intelligent Engineering & Systems*, vol. 12, no. 4, pp. 50–61, 2019. https://doi.org/10.22266/ijies2019.0831.06

[15] R. D. Abdu-Aljabar and O. A. Awad, "Improving lung cancer relapse prediction using the developed Optuna_XGB classification model," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 1, pp. 131–141, 2023. https://doi.org/10.22266/ijies2023.0228.12

[16] A. Hashmi, W. Ali, A. Abulfaraj, F. Binzagr, and E. Alkayal, "Enhancing cancerous gene selection and classification for high-dimensional microarray data using a novel hybrid filter and differential evolutionary feature selection," *Cancers*, vol. 16, no. 23, p. 3913, 2024. https://doi.org/10.3390/cancers16233913

[17] M. A. A. Al-Masoudy and A. Al-Azawei, "Proposing a feature selection approach to predict learners' performance in Virtual Learning Environments (VLEs)," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 11, pp. 110–131, 2023. https://doi.org/10.3991/ijet.v18i11.35405

[18] A. Shaikh, "Advances in deep learning in mobile interactive algorithms and learning technologies," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 14, no. 10, pp. 4–6, 2020. https://doi.org/10.3991/ijim.v14i10.15369

[19] S. J. Susmi, "An efficient gene expression data classification using optimized bidirectional long short-term memory with self-attention mechanism," *Multimedia Tools and Applications*, vol. 83, pp. 74159–74176, 2024. https://doi.org/10.1007/s11042-024-18387-6

[20] P. Keerin and T. Boongoen, "Improved KNN imputation for missing values in gene expression data," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 4009–4025, 2021. https://doi.org/10.32604/cmc.2022.020261

[21] Y. Yang and G. Mirzaei, "Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification," *PLoS One*, vol. 19, no. 2, p. e0293607, 2024. https://doi.org/10.1371/journal.pone.0293607

[22] D. W. Firmansyah and R. Sarno, "Data augmentation technique using two-step SMOTE for electronic-nose signal in breath ketone level detection," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 4, pp. 523–536, 2023. https://doi.org/10.22266/ijies2023.0831.42

[23] W. Zhongxin, S. Gang, Z. Jing, and Z. Jia, "Feature selection algorithm based on mutual information and LASSO for microarray data," *The Open Biotechnology Journal*, vol. 10, pp. 278–286, 2016. https://doi.org/10.2174/1874070701610010278

[24] Z. Li, W. Xie, and T. Liu, "Efficient feature selection and classification for microarray data," *PLoS One*, vol. 13, no. 8, p. e0202167, 2018. https://doi.org/10.1371/journal.pone.0202167

[25] F. Alharbi, A. Vakanski, M. K. Elbashir, and M. Mohammed, "LASSO–MOGAT: A multi-omics graph attention framework for cancer classification," *Academia Biology*, vol. 2, no. 3, 2024. https://doi.org/10.20935/AcadBiol7325

[26] T. Li *et al.*, "Mugen-UMAP: UMAP visualization and clustering of mutated genes in single-cell DNA sequencing data," *BMC Bioinformatics*, vol. 25, 2024. https://doi.org/10.1186/s12859-024-05928-x

[27] A. Abraham, H. S. Mohideen, and R. Kayalvizhi, "A tabular variational autoencoder-based hybrid model for imbalanced data classification with feature selection," *IEEE Access*, vol. 11, pp. 122760–122771, 2023. https://doi.org/10.1109/ACCESS.2023.3329139

[28] R. Saturi and P. Premchand, "Multi-objective feature selection method by using ACO with PSO algorithm for breast cancer detection," *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 5, pp. 359–368, 2021. https://doi.org/10.22266/ijies2021.1031.32

[29] N. Tavasoli, K. Rezaee, M. Momenzadeh, and M. Sehhati, "An ensemble soft weighted gene selection-based approach and cancer classification using modified metaheuristic learning," *Journal of Computational Design and Engineering*, vol. 8, no. 4, pp. 1172–1189, 2021. https://doi.org/10.1093/jcde/qwab039

[30] G. J. Gordon *et al.*, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[31] U. Alon *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, 1999, no. 12, pp. 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

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

[33] M. A. Shipp *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002. https://doi.org/10.1038/nm0102-68

[34] 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, 2023. https://doi.org/10.1186/s12859-023-05267-3

[35] 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, pp. 663–681, 2022. https://doi.org/10.1007/s11517-021-02476-x

[36] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6, no. 139, 2005. https://doi.org/10.1186/1471-2105-6-76

[37] A. M. Alshareef *et al.*, "Optimal deep learning-enabled prostate cancer detection using microarray gene expression," *Journal of Healthcare Engineering*, vol. 2022, no. 1, p. 7364704, 2022. https://doi.org/10.1155/2022/7364704

[38] A. Razzaque and A. Badholia, "PCA-based feature extraction and MPSO-based feature selection for gene expression microarray medical data classification," *Measurement: Sensors*, vol. 31, p. 100945, 2024. https://doi.org/10.1016/j.measen.2023.100945

[39] 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, 2024. https://doi.org/10.1007/s11042-023-15207-1

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

[41] O. R. Olaniran and M. A. A. Abdullah, "Subset selection in high-dimensional genomic data using hybrid variational Bayes and bootstrap priors," *Journal of Physics: Conference Series*, vol. 1489, p. 012030, 2021. https://doi.org/10.1088/1742-6596/1489/1/012030

[42] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," *IET Systems Biology*, vol. 16, nos. 3–4, pp. 120–131, 2022. https://doi.org/10.1049/syb2.12044

[43] D. Painuli, S. Bhardwaj, and U. Kose, "Optimized diagnosis of central nervous system (CNS) cancer using gene expression microarray & machine learning (ML) methods," *European Chemical Bulletin*, vol. 12, no. 10, pp. 9757–9771, 2023.

[44] S. Vasanthakumar and N. Ranjith, "A hybrid ensemble method for accurate fuzzy and support vector machine for gene expression in data mining," *ICTACT Journal on Soft Computing*, vol. 11, no. 4, pp. 2444–2448, 2021. https://doi.org/10.21917/ijsc.2021.0349

[45] P. M. Shakeel, A. Tolba, Z. Al-Makhadmeh, and M. M. Jaber, "Automatic detection of lung cancer from biomedical dataset using discrete AdaBoost optimized ensemble learning generalized neural networks," *Neural Computing and Applications*, vol. 32, pp. 777–790, 2020. https://doi.org/10.1007/s00521-018-03972-2

[46] M. S. Karthika, H. Rajaguru, and A. R. Nair, "Analysis of machine learning classifiers for the detection of lung cancer from microarray gene data," in *Proc. of the 2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, 2023, pp. 1–6. https://doi.org/10.1109/STCR59085.2023.10396899

[47] S. H. Bouazza and J. H. Bouazza, "Advanced cancer classification using AI and pattern recognition techniques," in *ITM Web of Conferences*, vol. 69, 2024. https://doi.org/10.1051/itmconf/20246902001

[48] S. H. Bouazza and J. H. Bouazza, "Revolutionizing cancer classification: The SNR-OGSCC method for improved gene selection and clustering," *International Journal of Artificial Intelligence*, vol. 14, no. 1, pp. 466–472, 2025. https://doi.org/10.11591/ijai.v14.i1.pp466-472

8 AUTHOR

Sara Haddou Bouazza is a Professor and coordinator of the preparatory year program (*filière année préparatoire*) at the Moroccan School of Engineering Sciences (EMSI). She holds a doctorate in Computer Science and Electrical Engineering from Cadi Ayyad University. Her research interests lie in artificial intelligence, machine learning, and bioinformatics, with a particular focus on gene expression-based cancer classification. Dr. Bouazza has developed advanced feature selection and classification frameworks validated on multi-cancer datasets from TCGA and GEO, aiming to improve the robustness and interpretability of computational diagnostics (E-mail: s.haddoubouazza@emsi.ma).