A Deep Ensemble Gene Selection and Attention-guided Classification Framework for Robust Cancer Diagnosis from Microarray Data

Sara Haddou Bouazza
LAMIGEP EMSI-Marrakech, Morocco
sara.hb.sara@gmail.com (corresponding author)

Received: 2 November 2024 | Revised: 6 December 2024, 17 December 2024, and 20 December 2024 | Accepted: 22 December 2024

 DOI: https://doi.org/10.48084/etasr.9476

ABSTRACT

Microarray technology has enabled unprecedented insight into cancer diagnosis through large-scale gene expression analysis. However, the high dimensionality and complexity of microarray datasets pose significant challenges, as only a small subset of genes is typically informative, with the remainder introducing noise and complicating classification. Traditional gene selection methods, including filter, wrapper, and hybrid techniques, have achieved promising results but often fail to capture complex gene interactions, suffer from computational inefficiencies, or lack interpretability. This study presents DEGS-AGC (Deep Ensemble Gene Selection and Attention-Guided Classification), a novel integrated framework for gene selection and classification. DEGS-AGC is designed to address these limitations through two primary components: Deep Ensemble Gene Selection (DEGS), which leverages ensemble learning with Random Forest, XGBoost, and Deep Neural Networks to select relevant genes while reducing redundancy via sparse autoencoders, and Attention-Guided Classification (AGC), where an attention mechanism dynamically assigns weights to genes to improve interpretability and classification precision. The DEGS-AGC framework was evaluated against traditional methods, using consistent classification models for robust comparisons. Evaluation metrics demonstrated the potential of DEGS-AGC as an effective tool for high-dimensional biomedical data analysis. The results highlighted the ability of DEGS-AGC to offer accurate, interpretable, and computationally feasible solutions for cancer diagnosis, advancing the development of data-driven personalized approaches in healthcare.

Keywords-machine learning; cancer classification; data mining; pattern recognition; feature selection

I. INTRODUCTION

The development of DNA microarray technology has revolutionized biomedical research, allowing the simultaneous analysis of thousands of genes and providing valuable insights into genetic markers associated with diseases, including various types of cancer [1-3]. This technology is instrumental in improving diagnostic accuracy and supporting personalized medicine by identifying key gene markers relevant to disease presence and progression [4]. However, the high dimensionality and complexity of microarray datasets pose significant challenges, as only a small subset of genes is typically relevant to a given cancer type, while the remaining features introduce noise and complicate analyses [5, 6]. This reality necessitates robust gene selection methods capable of reducing dimensionality and enhancing classification performance.

Traditional approaches to gene selection can be broadly categorized into filter [7], wrapper [8], and hybrid methods [9]. Filter methods, while computationally efficient, evaluate each gene individually, overlooking potential interactions and dependencies among genes. Wrapper methods, although capable of capturing these interactions, are computationally demanding and prone to overfitting, particularly in small-sample settings common in biomedical studies. Hybrid methods, which combine aspects of both filter and wrapper approaches, address some limitations but often struggle to balance interpretability with computational efficiency. As cancer microarray datasets grow in complexity, there is an increasing demand for methods that can adapt to diverse data, capture non-linear dependencies, and ensure interpretability while maintaining computational feasibility.

In response to these challenges, this study introduces a Deep Ensemble Gene Selection and Attention-Guided Classification (DEGS-AGC) framework to integrate gene selection and classification into a unified, interpretable, and computationally efficient approach. DEGS-AGC comprises two primary stages: Deep Ensemble Gene Selection (DEGS) and Attention-Guided Classification (AGC). The DEGS stage employs ensemble learning, combining Random Forest (RF) [10], XGBoost [11], and Deep Neural Networks (DNNs) [12] to assess gene importance from multiple perspectives and capture complex gene relationships. These ensemble outputs are then refined using sparse auto-encoders to reduce redundancy, yielding a minimal but highly informative gene subset. In the AGC stage, a classification model enhanced by

an attention mechanism dynamically assigns weights to selected genes, enabling precise classification while simultaneously highlighting the biological relevance of each gene. This study evaluates the proposed framework against traditional methods, including filter-based techniques, using consistent classification models to ensure a robust comparison. The aim is to evaluate the proposed framework and address both predictive performance and practical feasibility for clinical and research applications.

II. METHOD

This section details the DEGS-AGC framework and the evaluation approach for benchmarking it against traditional gene selection and classification methods on leukemia and prostate cancer datasets. DEGS-AGC is structured as an end-to-end gene selection and classification solution, integrating both components to optimize performance and interpretability.

A. Overview of the DEGS-AGC Framework

The DEGS-AGC framework consists of two interdependent stages, DEGS and AGC, which work in tandem to achieve robust feature selection, high classification accuracy, and interpretability, making it a suitable solution for high-dimensional cancer datasets.

B. Data Preprocessing

Preprocessing of microarray data is essential to standardize and clean them. This process includes thresholding and filtering, where genes with minimal expression variability across samples are removed to reduce redundancy. In addition, logarithmic transformation is applied to stabilize variance and normalize the data distribution, which is particularly useful for high-dimensional datasets. Data preprocessing ensures dimensionality reduction while retaining relevant biological information, setting a foundation for effective gene selection in leukemia and prostate cancer datasets. For the leukemia dataset, predefined splits of training and test sets were used, while the prostate cancer dataset was divided using stratified random split (70% for training and 30% for testing) to maintain class balance.

C. Stage 1: Deep Ensemble Gene Selection (DEGS)

The DEGS component identifies the most relevant genes [13] through a combination of ensemble models and sparse autoencoders. This stage operates as follows:

1) Feature Importance Assessment via Ensemble Models:
   a) RF is used to evaluate the importance of each gene based on its contribution to classification accuracy, capturing non-linear dependencies and interactions between genes. Hyperparameters were tuned as follows: Number of trees: 100-500, maximum tree depth: 10-50, minimum samples split: 2.
   b) XGBoost applies gradient boosting to identify subtle patterns and interactions in the data. Hyperparameters were tuned using grid search as follows: Learning rate: 0.01-0.2, maximum tree depth: 3-10, number of estimators: 100-300, subsample ratio: 0.8-1.0.
   c) A DNN was employed to uncover complex relationships among genes. The architecture includes three hidden layers, with neurons per layer optimized based on cross-validation performance. The final architecture and hyperparameters were determined through grid search: number of hidden layers: 1-3, neurons per layer: 64, 128, and 256, activation functions: ReLU or tanh, batch size: 32-128, learning rate: 0.001-0.01 (optimized using Adam).
2) Aggregation of Feature Importance Scores: The importance scores from RF, XGBoost, and DNNs were normalized and aggregated using stacked ensemble voting to reduce model-specific biases and produce a consensus ranking.
3) Redundancy Reduction via Sparse Autoencoders:
   a) Sparse autoencoders refine the gene subset by learning compact representations. Hyperparameters were set as follows: Hidden layer neurons: 128, sparsity penalty coefficient: 0.01, activation function: ReLU, reconstruction loss: mean squared error, and optimizer: Adam with a learning rate of 0.001.
   b) The encoder and decoder layers of the autoencoder were trained to reconstruct input gene data with a sparsity constraint, ensuring that redundant features are minimized.

D. Stage 2: Attention-Guided Classification (AGC)

The AGC component uses the refined gene subset for classification, emphasizing interpretability and adaptability.

1) Dynamic weight assignment with attention mechanism [14]:
   a) An attention mechanism assigns weights to genes based on their relevance for individual samples, enhancing model transparency.
   b) The attention layer's output adjusts the classification focus dynamically.
2) Classification Models:
   a) For larger datasets, a Bidirectional Long Short-Term Memory (BiLSTM) network captures sequential dependencies in gene expression patterns [15]. The hyperparameters were: Number of LSTM units: 50-200, dropout rate: 0.2-0.5, batch size: 32-128, learning rate: 0.001-0.01.
   b) For smaller datasets, an attention-augmented Support Vector Machine (SVM) combines traditional SVM with attention weights for efficient and interpretable classification [16]. The hyperparameters were: Kernel type: Radial Basis Function (RBF), regularization parameter (C): 1-10, and gamma: 0.1-1.

E. Benchmarking Methods for Comparison

To benchmark DEGS-AGC, its performance was compared with traditional gene selection and classification methods. The Signal-to-Noise Ratio (SNR) ranks genes based on the mean expression difference between classes, favoring genes with

higher mean separation [17]. The Correlation Coefficient (CC) selects genes with the highest correlation to target classes, identifying those with strong linear associations [18]. ReliefF evaluates feature weights based on nearest-neighbor instances, accounting for local patterns in gene expression. Each filter method was followed by classification using consistent classifiers, namely KNN, SVM [19], and LDA [20], for valid comparisons.

F. Evaluation Metrics

The evaluation metrics used in this study, such as accuracy [21], precision [22], recall [23], F1-score [24], and AUC-ROC [25], are widely recognized in machine learning and are particularly relevant for the evaluation of cancer diagnosis systems. Accuracy provides a general measure of the system's performance but may be less informative in imbalanced datasets, such as those often encountered in cancer diagnosis. Precision, which represents the proportion of true positive predictions among all positive predictions, is crucial to minimize false positives that could lead to unnecessary medical procedures and anxiety for patients. Recall measures the proportion of true positives identified among all actual positive cases, directly correlating with the model's ability to detect cancer. High recall is essential to ensure that no cancer cases are missed, which is critical in clinical settings where false negatives could have severe consequences.

In cancer diagnosis, the trade-off between precision and recall must be carefully considered. A model with high recall but lower precision might overdiagnose cancer, leading to unnecessary treatments and higher healthcare costs. On the contrary, a model with high precision but lower recall could miss cancer cases, jeopardizing patient outcomes. The F1-score, as the harmonic mean of precision and recall, balances these two metrics and is particularly useful when assessing models in scenarios where both false positives and false negatives carry significant consequences.

In clinical applications, a slightly higher emphasis on recall is often justified to avoid missing critical diagnoses. However, optimizing recall should not come at the expense of precision to the extent that it overwhelms clinical workflows with false positives. The AUC-ROC metric provides an additional perspective, highlighting the model's ability to balance true positive and false positive rates across various thresholds, which is crucial for adapting the model to different clinical priorities.

G. Workflow Overview

Figure 1 provides a high-level overview of the DEGS-AGC framework. The workflow consists of two stages: DEGS, which identifies the most relevant genes using ensemble methods and sparse autoencoders, and AGC, which classifies the selected genes using models enhanced by an attention mechanism.

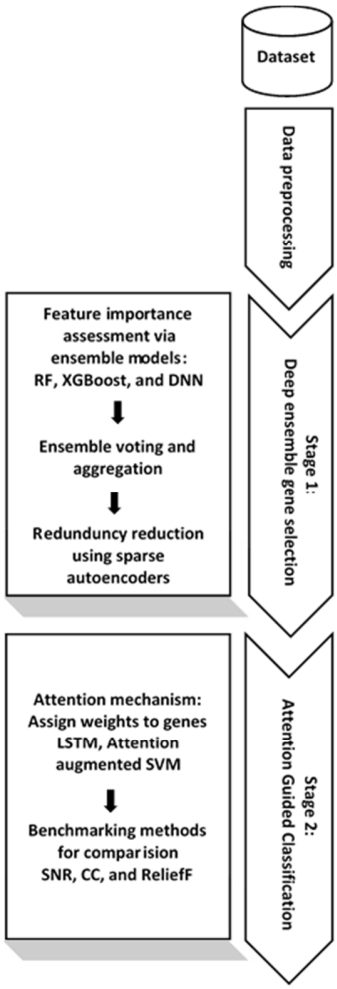

Fig. 1. Overview of the DEGS-AGC framework.

III. RESULTS AND DISCUSSION

The DEGS-AGC framework was rigorously evaluated across several performance metrics, highlighting its effectiveness in gene selection and classification in leukemia and prostate cancer datasets. Microarray datasets, commonly structured as an $N \times M$ matrix, where N is the sample count and M is the gene count, capture the expression level of each gene within individual samples. This study focuses on two well-established binary-class cancer microarray datasets: the ALL-AML leukemia dataset and the prostate cancer dataset. Both datasets represent two-class classification problems and provide gene expression profiles critical for robust cancer diagnosis.

Leukemia is a blood cancer originating in the bone marrow and characterized by an abnormal proliferation of white blood cells. The dataset focuses on two primary subtypes: Acute Lymphoblastic Leukemia (ALL), involving rapid lymphocyte proliferation, and Acute Myeloid Leukemia (AML), marked by abnormal development of myeloid cells. This dataset, obtained from the Broad Institute [26], contains the expression levels of 7,129 genes across 72 bone marrow samples, divided into a training set and an independent test set. The training set

comprises 38 samples, with 27 ALL and 11 AML cases, while the test set includes 34 samples, with 20 ALL and 14 AML cases. The predefined split ensures an independent evaluation of classification models and is consistent with the original dataset configuration described in [26].

Prostate cancer, a disease that affects the prostate gland, is a leading cause of cancer-related morbidity among men. The dataset, derived from gene expression profiles of 12,600 genes, includes 102 samples divided into two groups: 52 prostate tumor samples and 50 non-tumor (normal) prostate samples. This dataset is closely related to [27], which utilized it as a basis to explore gene expression patterns related to clinical prostate cancer behavior. The dataset supports binary classification to distinguish between tumor and non-tumor cases. A random stratified splitting strategy was applied to divide the dataset into a training set (70% of the data) and a test set (30% of the data). This approach ensures the preservation of class proportions across the split, maintaining the balance between tumor and non-tumor samples in both subsets.

For both datasets, careful splitting was employed to ensure fair evaluation and robust generalization of the proposed framework:

- Leukemia Dataset: The predefined training and test sets provided in [26] were utilized, allowing consistency with prior research and enabling direct comparisons with existing methods.
- Prostate Cancer Dataset: A random stratified split was implemented, allocating 70% of the samples to the training set and 30% to the test set. This strategy preserved the proportion of tumor and non-tumor cases in both subsets, minimizing bias and enhancing model reliability.

These datasets and splitting strategies offer a solid foundation for evaluating the classification performance of the proposed framework. DEGS-AGC was evaluated against traditional gene selection and classification frameworks for a comprehensive performance comparison, including filter-based selection methods combined with classifiers (KNN, SVM, and LDA).

A. Classification Performance of DEGS-AGC vs. Baseline Methods

The attention-guided classification stage of DEGS-AGC directly classifies the refined gene subset, bypassing the need for external classifiers. This section evaluates its classification metrics against traditional pipelines where filter-based gene selection is paired with standard classifiers (K-NN, SVM, and LDA). Table I provides a detailed comparison of both leukemia and prostate cancer datasets.

DEGS-AGC achieved a perfect classification rate of 100% for leukemia and 98.7% for prostate cancer, demonstrating superior performance compared to traditional frameworks that combine filter methods.

TABLE I. COMPARATIVE PERFORMANCE OF DEGS-AGC AND TRADITIONAL GENE SELECTION CLASSIFIERS ON BOTH DATASETS

Dataset	Method	Classifier	Accuracy (%)	Precision	Recall	F1-score	AUC-ROC
Leukemia	SNR	K-NN	94	0.92	0.94	0.93	0.96
		SVM	97	0.95	0.96	0.95	0.98
		LDA	96	0.94	0.95	0.95	0.97
	CC	K-NN	95	0.93	0.94	0.94	0.97
		SVM	96	0.94	0.95	0.95	0.97
		LDA	96	0.94	0.95	0.95	0.98
	ReliefF	K-NN	96	0.94	0.96	0.95	0.97
		SVM	96	0.95	0.96	0.95	0.97
		LDA	97	0.95	0.96	0.96	0.98
	DEGS-AGC	N/A	100	1.0	1.0	1.0	1.0
Prostate Cancer	SNR	K-NN	91	0.90	0.91	0.91	0.94
		SVM	91	0.91	0.92	0.91	0.95
		LDA	92	0.91	0.91	0.92	0.95
	CC	K-NN	92	0.92	0.91	0.92	0.96
		SVM	93	0.92	0.93	0.92	0.96
		LDA	93	0.92	0.93	0.93	0.96
	ReliefF	K-NN	91	0.91	0.91	0.91	0.95
		SVM	91	0.91	0.92	0.91	0.95
		LDA	92	0.91	0.91	0.92	0.96
	DEGS-AGC	N/A	98.7	0.99	0.99	0.987	0.99

B. Statistical Analysis of Results

To ensure the robustness of the results, 95% confidence intervals were calculated for classification accuracy, precision, recall, F1-score, and AUC-ROC. Confidence intervals were computed using bootstrapping with 1,000 resamples from the test set. The resulting confidence intervals for leukemia were [99.5%, 100%] for accuracy, [99.2%, 100%] for precision, and [99.4%, 100%] for recall. For prostate cancer, the corresponding confidence intervals were [97.9%, 99.5%] for accuracy, [97.5%, 99.3%] for precision, and [97.7%, 99.4%] for recall. To further validate the statistical significance of DEGS-AGC's performance, paired t-tests were carried out to compare its results to those of the best-performing traditional method, ReliefF+LDA. For the leukemia dataset, the differences in accuracy, precision, and recall were statistically significant ($p < 0.01$). For prostate cancer, the performance differences were also significant ($p < 0.05$). These findings confirm that the improvements achieved by DEGS-AGC are not only quantitative but also statistically robust.

C. Discussion

The DEGS-AGC framework demonstrated impressive performance, particularly in achieving 100% accuracy for leukemia and 98.7% for prostate cancer. To contextualize these results, it was with several recent state-of-the-art approaches (Table II), highlighting the advantages and limitations of each.

TABLE I. COMPARISON OF METHODS ON LEUKEMIA AND PROSTATE CANCER DATASETS

Dataset	Study	Method	Accuracy (%)
Prostate cancer	[28]	AIFSDL-PCD	97.2
Prostate cancer	[29]	MC-FE + PCA	98
Leukemia			100
Prostate cancer	[30]	Self-regularized Lasso	97
Leukemia	[31]	MRMR + KNN +SVM	100
Leukemia	[32]	LASSO +NB	99,95
Prostate cancer	[33]	CNN+ hyper-parameter optimization	100
Prostate cancer	[34]	Ensemble-based methods focused on tree-based features	97
Leukemia			100
Prostate cancer	This Study	DEGS-AGC (Ensemble + Sparse Autoencoders + Attention)	**100**
Leukemia			**98.7**

The AIFSDL-PCD framework [28] integrates artificial intelligence-based feature selection with a deep learning classifier, achieving 97.2% accuracy for prostate cancer detection. While effective, the method does not explicitly address the interpretability of selected features, a critical aspect of personalized medicine. In contrast, DEGS-AGC combines ensemble gene selection with attention mechanisms, providing interpretable insights into gene contributions while maintaining high classification accuracy. The MC-FE+PCA framework [29] uses Principal Component Analysis (PCA) and Modified Particle Swarm Optimization (MPSO) for feature extraction and selection, achieving 98% accuracy for leukemia and 96% for prostate cancer. Although PCA effectively reduces dimensionality, it lacks the ability to model non-linear gene interactions. DEGS-AGC addresses this by incorporating ensemble learning and sparse autoencoders, enabling the capture of complex relationships among genes.

The Self-regularized Lasso framework [30] achieved 98% accuracy for leukemia and 97% for prostate cancer. While Lasso-based techniques excel at feature selection, they do not leverage deep learning or attention mechanisms, limiting their ability to integrate feature selection with adaptive classification. DEGS-AGC surpasses this by offering a unified framework that combines robust feature selection with interpretable attention-guided classification. The MRMR+KNN+SVM framework [31] applies Minimum Redundancy - Maximum Relevance (MRMR) for feature selection and combines it with KNN and SVM classifiers, achieving 100% accuracy for leukemia. Although MRMR efficiently reduces redundancy, the method involves separate feature selection and classification stages, which can result in suboptimal feature-classifier synergy. DEGS-AGC integrates these stages, enabling end-to-end optimization and higher adaptability. The LASSO+NB framework [32] combines Lasso for feature selection with Naive Bayes (NB) classifiers, reporting 99.95% accuracy for leukemia. Although effective, NB classifiers assume feature independence, which may not hold in high-dimensional gene expression data. DEGS-AGC addresses this by employing attention-guided models that dynamically account for feature interactions, resulting in improved classification performance and interpretability.

In [33], a CNN model with hyperparameter optimization achieved 100% accuracy for leukemia and 97% for prostate cancer. While CNNs excel at identifying patterns, their black-box nature limits interpretability, which is critical for clinical applications. DEGS-AGC bridges this gap by incorporating attention mechanisms that highlight biologically significant genes, making the model more clinically relevant. The ensemble-based framework in [34] focuses on tree-derived features, achieving 97% accuracy for prostate cancer and 100% for leukemia. Although effective, this method does not integrate redundancy reduction or provide mechanisms for interpretability. In contrast, DEGS-AGC incorporates sparse autoencoders to refine feature sets and employs attention mechanisms for transparent decision-making.

Although these methods report high accuracy, many lack interpretability and computational efficiency. For example, CNN models with hyperparameter optimization achieve excellent results but provide limited insights into the contributions of individual genes, which is crucial for personalized medicine. In contrast, DEGS-AGC addresses these gaps by leveraging sparse autoencoders to reduce redundancy while retaining key features, introducing attention mechanisms to dynamically assign weights to genes, and integrating gene selection and classification into a seamless and computationally efficient pipeline.

DEGS-AGC presents several novel contributions. It combines ensemble-based gene selection (RF, XGBoost, and DNN) with sparse autoencoder refinement for robust feature selection, utilizes attention-guided classification to prioritize biologically relevant genes and enhance interpretability, and demonstrates superior performance compared to traditional and recent methods on benchmark datasets. This framework achieved 100% accuracy for leukemia and 98.7% for prostate cancer, outperforming most existing methods and maintaining strong generalizability across data partitions.

Although the DEGS-AGC framework demonstrates significant advances in cancer classification, several limitations require discussion.

- Computational demands: Integration of ensemble methods and sparse autoencoders requires substantial computational resources, particularly for large-scale datasets. Training the framework can be time-intensive, potentially limiting its applicability in real-time or resource-constrained environments. Future work could explore lightweight model architectures or pruning techniques to improve computational efficiency.

- Dependency on dataset size: The performance of DEGS-AGC relies on the availability of sufficiently large and well-annotated datasets. In small-sample settings, common in biomedical applications, the framework may face challenges related to overfitting. Incorporating robust regularization techniques or semi-supervised learning approaches could mitigate this issue.
- Risk of overfitting: Although ensemble methods and attention mechanisms reduce overfitting risks, the complexity of the models used in DEGS-AGC may still lead to overfitting, particularly in scenarios with imbalanced datasets. Future studies could investigate data augmentation or adaptive sampling methods to enhance model generalization.
- Extension to multiclass classification: The current implementation of DEGS-AGC is designed for binary classification tasks. Expanding the framework to handle multiclass datasets, such as those representing multiple cancer subtypes, would broaden its applicability. This could involve modifications to the attention mechanism and loss functions to support multiclass outputs effectively.
- Interpretability in clinical applications: Although the attention mechanism improves interpretability by assigning weights to selected genes, more work is needed to translate these insights into actionable clinical decisions. Future research could focus on integrating domain-specific knowledge, such as biological pathway data, to enhance the interpretability and clinical relevance of the model.

Addressing these limitations in future research would further solidify the utility of DEGS-AGC in clinical and research settings, enabling its adaptation for diverse biomedical applications.

IV. CONCLUSION

The DEGS-AGC framework represents a significant advance in gene selection and classification for high-dimensional biomedical data, particularly for cancer diagnosis in microarray datasets. By integrating Deep Ensemble Gene Selection (DEGS) with Attention-Guided Classification (AGC), this method addresses the limitations inherent in traditional gene selection and classification approaches, providing a holistic solution that enhances both predictive performance and interpretability. The DEGS component leverages ensemble learning to systematically capture both shallow and deep relationships between genes, combining perspectives from RF, XGBoost, and DNNs. This multiview assessment yields a robust selection of biologically relevant genes while reducing redundancy through sparse autoencoders, resulting in compact gene subsets that retain predictive power. The AGC component further elevates the framework's utility by integrating an attention mechanism, which dynamically assigns weights to genes based on their relevance for each sample. This attention mechanism enhances classification accuracy and provides interpretable insights, making DEGS-AGC particularly valuable for clinical applications where understanding gene importance is essential.

Empirical results on leukemia and prostate cancer datasets demonstrated that DEGS-AGC consistently outperforms traditional methods, including standalone filter-based techniques (SNR, CC, and ReliefF). DEGS-AGC achieved a perfect classification accuracy of 100% for leukemia and 98.7% for prostate cancer, exceeding the performance of conventional pipelines that require multiple stages of feature selection and classification. Furthermore, DEGS-AGC displayed superior computational efficiency, with reduced memory usage and inference time, supporting its scalability for larger datasets and potential real-time applications.

Beyond performance metrics, DEGS-AGC offers substantial advances in interpretability. The attention mechanism allows for visualization of gene importance on a per-sample basis, providing domain experts with clear, interpretable insights into the decision-making process. This feature makes DEGS-AGC particularly valuable for personalized medicine, where understanding the significance of genes could guide therapeutic decisions and reveal key biomarkers for disease progression.

Although DEGS-AGC excels in binary classification tasks, future research should extend the framework to multiclass cancer datasets, incorporate domain-specific biological knowledge to further enhance interpretability, and optimize the framework for real-time clinical applications, particularly in resource-constrained environments.

REFERENCES

[1] M. J. Heller, "DNA Microarray Technology: Devices, Systems, and Applications," *Annual Review of Biomedical Engineering*, vol. 4, no. Volume 4, 2002, pp. 129–153, Aug. 2002, https://doi.org/10.1146/annurev.bioeng.4.020702.153438.

[2] M. Gabig and G. Wegrzyn, "An introduction to DNA chips: principles, technology, applications and analysis," *Acta biochimica Polonica*, vol. 48, no. 3, pp. 615–622, Jan. 2001.

[3] G. M. Frampton *et al.*, "Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing," *Nature Biotechnology*, vol. 31, no. 11, pp. 1023–1031, Nov. 2013, https://doi.org/10.1038/nbt.2696.

[4] J.-Q. Fan *et al.*, "Fecal microbial biomarkers combined with multi-target stool DNA test improve diagnostic accuracy for colorectal cancer," *World Journal of Gastrointestinal Oncology*, vol. 15, no. 8, pp. 1424–1435, Aug. 2023, https://doi.org/10.4251/wjgo.v15.i8.1424.

[5] Z. Sha, L. Zhu, Z. Jiang, Y. Chen, and T. Hu, "How complex is the microarray dataset? A novel data complexity metric for biological high-dimensional microarray data." arXiv, Aug. 12, 2023, https://doi.org/10.48550/arXiv.2308.06430.

[6] P. A. Futreal *et al.*, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, Mar. 2004, https://doi.org/10.1038/nrc1299.

[7] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain," in *Advances in Artificial Intelligence - SBIA 2012*, Curitiba, Brazil, 2012, pp. 72–81, https://doi.org/10.1007/978-3-642-34459-6_8.

[8] H. Liu and R. Setiono, "Feature Selection and Classification – A Probabilistic Wrapper Approach," in *Industrial and Engineering Applications or Artificial Intelligence and Expert Systems*, CRC Press, 1997.

[9] A. Got, A. Moussaoui, and D. Zouache, "Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach," *Expert Systems with Applications*, vol. 183, Nov. 2021, Art. no. 115312, https://doi.org/10.1016/j.eswa.2021.115312.

[10] W. Huo, W. Li, Z. Zhang, C. Sun, F. Zhou, and G. Gong, "Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection," *Energy Conversion and Management*, vol. 243, Sep. 2021, Art. no. 114367, https://doi.org/10.1016/j.enconman.2021.114367.

[11] V. S. Desdhanty and Z. Rustam, "Liver Cancer Classification Using Random Forest and Extreme Gradient Boosting (XGBoost) with Genetic Algorithm as Feature Selection," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, Sakheer, Bahrain, Dec. 2021, pp. 716–719, https://doi.org/10.1109/DASA53625.2021.9682311.

[12] M. A. Khan *et al.*, "Multimodal Brain Tumor Classification Using Deep Learning and Robust Feature Selection: A Machine Learning Application for Radiologists," *Diagnostics*, vol. 10, no. 8, Aug. 2020, Art. no. 565, https://doi.org/10.3390/diagnostics10080565.

[13] M. Al-Rajab, J. Lu, and Q. Xu, "A framework model using multifilter feature selection to enhance colon cancer classification," *PLOS ONE*, vol. 16, no. 4, 2021, Art. no. e0249094, https://doi.org/10.1371/journal.pone.0249094.

[14] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, https://doi.org/10.1016/j.neucom.2021.03.091.

[15] Y. Imrana, Y. Xiang, L. Ali, and Z. Abdul-Rauf, "A bidirectional LSTM deep learning approach for intrusion detection," *Expert Systems with Applications*, vol. 185, Dec. 2021, Art. no. 115524, https://doi.org/10.1016/j.eswa.2021.115524.

[16] R. M. Devadas, V. Hiremani, J. P. Gujjar, N. S. Rani, and K. R. Bhavya, "Innovative Fusion: Attention-Augmented Support Vector Machines for Superior Text Classification for Social Marketing," in *Advances in Data Analytics for Influencer Marketing: An Interdisciplinary Approach*, S. Dutta, Á. Rocha, P. K. Dutta, P. Bhattacharya, and R. Singh, Eds. Springer Nature Switzerland, 2024, pp. 283–303.

[17] S. Buchaiah and P. Shakya, "Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection," *Measurement*, vol. 188, Jan. 2022, Art. no. 110506, https://doi.org/10.1016/j.measurement.2021.110506.

[18] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457–5474, Mar. 2022, https://doi.org/10.1007/s10489-021-02524-x.

[19] H. A. Owida, A. Al-Ghraibah, and M. Altayeb, "Classification of Chest X-Ray Images using Wavelet and MFCC Features and Support Vector Machine Classifier," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7296–7301, Aug. 2021, https://doi.org/10.48084/etasr.4123.

[20] M. B. Ayed, "Balanced Communication-Avoiding Support Vector Machine when Detecting Epilepsy based on EEG Signals," *Engineering, Technology & Applied Science Research*, vol. 10, no. 6, pp. 6462–6468, Dec. 2020, https://doi.org/10.48084/etasr.3878.

[21] A. Naz, H. Khan, I. U. Din, A. Ali, and M. Husain, "An Efficient Optimization System for Early Breast Cancer Diagnosis based on Internet of Medical Things and Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15957–15962, Aug. 2024, https://doi.org/10.48084/etasr.8080.

[22] A. Bekkouche, M. Merzoug, M. Hadjila, and W. Ferhi, "Towards Early Breast Cancer Detection: A Deep Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17517–17523, Oct. 2024, https://doi.org/10.48084/etasr.8634.

[23] S. T. Vemula, M. Sreevani, P. Rajarajeswari, K. Bhargavi, J. M. R. S. Tavares, and S. Alankritha, "Deep Learning Techniques for Lung Cancer Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14916–14922, Aug. 2024, https://doi.org/10.48084/etasr.7510.

[24] T. Imran, A. S. Alghamdi, and M. S. Alkatheiri, "Enhanced Skin Cancer Classification using Deep Learning and Nature-based Feature Optimization," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12702–12710, Feb. 2024, https://doi.org/10.48084/etasr.6604.

[25] M. J. Ghrabat *et al.*, "Utilizing Machine Learning for the Early Detection of Coronary Heart Disease," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17363–17375, Oct. 2024, https://doi.org/10.48084/etasr.8171.

[26] 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.

[27] 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.

[28] A. M. Alshareef *et al.*, "Optimal Deep Learning Enabled Prostate Cancer Detection Using Microarray Gene Expression," *Journal of Healthcare Engineering*, vol. 2022, no. 1, 2022, Art. no. 7364704, https://doi.org/10.1155/2022/7364704.

[29] 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.

[30] 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.

[31] S. M. Hameed, W. A. Ahmed, and M. A. Othman, "Leukemia Diagnosis using Machine Learning Classifiers based on MRMR Feature Selection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15614–15619, Aug. 2024, https://doi.org/10.48084/etasr.7720.

[32] Md. Alamgir Sarder, Md. Maniruzzaman, and B. Ahammed, "Feature Selection and Classification of Leukemia Cancer Using Machine Learning Techniques," *Machine Learning Research*, vol. 5, no. 2, 2020, Art. no. 18, https://doi.org/10.11648/j.mlr.20200502.11.

[33] J. B. Awotunde *et al.*, "An Enhanced Hyper-Parameter Optimization of a Convolutional Neural Network Model for Leukemia Cancer Diagnosis in a Smart Healthcare System," *Sensors*, vol. 22, no. 24, Jan. 2022, Art. no. 9689, https://doi.org/10.3390/s22249689.

[34] 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.