# A Novel Adaptive Sparse Deep Feature Selection Method for Enhanced Gene-based Cancer Classification

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

*Received: 3 February 2025 | Revised: 19 February 2025 | Accepted: 27 February 2025*

 *DOI: https://doi.org/10.48084/etasr.10440*

## ABSTRACT

**The Adaptive Sparse Deep Feature Selection (ASDFS) method introduces a novel deep learning-based approach to enhance gene-based cancer classification. Designed to address the high dimensionality and complexity of genomic data, ASDFS leverages sparse autoencoders for dimensionality reduction and a Dual-Target Deep Neural Network (DT-DNN) to refine and identify a minimal yet biologically significant subset of genes. The method achieved outstanding classification accuracies of 99.9%, 100%, and 99.8% for ovarian, prostate, and lung cancers, respectively, achieving superior results compared to state-of-the-art techniques, including Principal Component Analysis with Grey Wolf Optimizer (PCA-GWO), Recursive Feature Elimination (RFE), and Minimum Redundancy Maximum Relevance (mRMR)-based hybrid methods. ASDFS was validated on microarray datasets with detailed characteristics: ovarian cancer (15,154 genes, 253 samples), prostate cancer (12,600 genes, 102 samples), and lung cancer (12,533 genes, 181 samples). This demonstrates its robust performance and ability to achieve significant gene reduction. Additionally, pathway enrichment analysis validated the biological relevance of the selected genes, highlighting their roles in critical cancer pathways.**

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

## I. INTRODUCTION

Gene expression profiling is advancing cancer research by revealing transcriptional patterns associated with cancer progression [1]. However, its effectiveness in cancer classification is hampered by high dimensionality, overfitting, and limited biological interpretability [2-4]. Overcoming these challenges requires advanced feature selection methods that balance accuracy, efficiency, and biological relevance. Current methods face limitations: Filter-based techniques such as ReliefF overlook complex gene interactions [5-7]; wrapper-based approaches, such as Recursive Feature Elimination (RFE), are computationally intensive and prone to overfitting [8]; and embedded methods, though efficient, struggle to model non-linear gene relationships [9]. Moreover, many methods fail to validate selected genes against known cancer pathways, reducing their clinical value.

This study introduces Adaptive Sparse Deep Feature Selection (ASDFS) to address these gaps. ASDFS employs deep learning with sparsity constraints to capture complex gene interactions while isolating biologically relevant features [10, 11]. It operates in two stages: sparse autoencoders filter noise and reduce dimensionality, whereas a Dual-Target Neural Network (DT-DNN) optimizes classification accuracy and biological relevance by aligning selected genes with key pathways like PI3K-Akt and MAPK [12, 13]. The research objectives are to design a feature selection framework that balances accuracy with biological relevance, enhance computational efficiency through sparsity, validate selected genes via pathway enrichment analysis, and benchmark ASDFS against existing methods on ovarian, prostate, and lung cancer datasets. This work provides a biologically informed, computationally efficient solution for gene-based cancer classification to advance precision oncology.

## II. METHOD

In this study, the ASDFS method is systematically developed to optimize gene selection for cancer classification while balancing accuracy with interpretability. Leveraging the power of deep learning in capturing complex gene interactions and sparsity constraints to isolate relevant genes, ASDFS is structured into three key stages: Data preprocessing, feature extraction with sparse autoencoder, and supervised fine-tuning with DT-DNN analysis [12, 13].

### *A. Dataset Description and Preprocessing*

All datasets underwent consistent preprocessing steps, including log transformation, centering, and scaling [14]. The ComBat algorithm was applied to correct for batch effects and to minimize variability from experimental conditions [15]. Outliers were removed using z-score filtering to ensure data integrity and reliability [16]. A 70/30 stratified split was applied for training and testing, while preserving class proportions. The datasets used are:

- Ovarian cancer dataset: Collected by Petricoin et al. [17], this dataset includes gene expression profiles from 253 ovarian tissue samples (162 cancerous, 91 normal) comprising 15,154 genes. The dataset is available from the DBC repository of the University of Granada: https://leo.ugr.es/elvira/DBCRepository/OvarianCancer/OvarianCancer-NCI-PBSII.html.
- Prostate cancer dataset: Published by Singh et al. [18], this dataset contains 102 prostate tissue samples (52 tumor, 50 normal) with expression profiles for 12,600 genes. It is available from the University of Granada: https://leo.ugr.es/elvira/DBCRepository/ProstateCancer/ProstateCancer.html, with documentation from Carnegie Mellon University: https://www.stat.cmu.edu/~jiashun/Research/software/HCClassification/Prostate/Readme.txt.
- Lung cancer dataset: Provided by Gordon et al. [19], this dataset includes 181 samples (31 MPM, 150 ADCA) with 12,533 gene expression profiles obtained via Affymetrix Human GeneAtlas U95Av2 microarrays. It is available from Mendeley Data: https://data.mendeley.com/datasets/572khd33by/1 and from the University of Granada: https://leo.ugr.es/elvira/DBCRepository/LungCancer/LungCancer-Harvard2.html.

### *B. Adaptive Sparse Deep Feature Selection*

ASDFS consists of three stages, each contributing to optimized feature selection and classification. These stages are as follows.

1. Unsupervised deep feature extraction with sparse autoencoder: A sparse autoencoder [20] is an unsupervised neural network designed for feature extraction, particularly effective for high-dimensional gene expression data. It compresses the input data into a latent space at the bottleneck layer, retaining only essential features. Sparsity is enforced via L1 regularization, which drives irrelevant weights to zero, filters noise, and enhances interpretability by prioritizing critical genes. Bayesian hyperparameter tuning optimized parameters such as bottleneck size and learning rate, balancing dimensionality reduction with biological relevance. This process distilled the data into fewer than 100 key genes, providing a strong foundation for classification and analysis.
2. Supervised fine-tuning with DT-DNN: The DT-DNN is a deep learning model designed to optimize both feature selection and classification [21]. It incorporates a custom gene-weighting layer that dynamically assigns importance to genes, focusing on biologically relevant features. Its dual loss function combines cross-entropy loss for classification accuracy with an adaptive sparsity penalty to select only the most relevant genes, ensuring both precision and interpretability. The Adam optimizer was used for efficient training, whereas early stopping prevented overfitting by halting training when validation performance plateaued, thereby enhancing generalizability.
3. Gene subset extraction and pathway analysis: Pathway enrichment analysis validated the biological relevance of the selected genes. Following Stage 2, where DT-DNN prioritized key genes, these high-weight genes were analyzed using KEGG and Reactome [22-24]. KEGG provides insights into biological functions and metabolic pathways, whereas Reactome details molecular interactions. This analysis identified critical cancer-related pathways, including PI3K-Akt (cell growth and survival), MAPK (proliferation and differentiation), and EGFR (cell division and survival). Mapping of selected genes to these pathways confirmed their role in cancer progression and reinforced their biological significance.

### *C. Classification and Evaluation*

The effectiveness of ASDFS-selected genes was evaluated using Support Vector Machine (SVM), a robust classifier suitable for high-dimensional data [24, 25]. Its compatibility with stage 1 gene subsets and stage 2 refined features made it ideal for this study [26]. Performance was assessed by fivefold cross-validation using key metrics:

- Accuracy: Measures correctly classified samples, validating feature selection and model training effectiveness [27].
- Precision: Ensures that predicted cancerous samples are truly cancerous, reducing false positives and unnecessary treatments [28].
- Recall: Evaluates sensitivity in detecting true positives, minimizing missed diagnoses [28].
- F1-score: Balances precision and recall, crucial for imbalanced datasets [29].
- Area Under the Curve (AUC): Assesses classifier separability; a high value confirms SVM robustness across thresholds [29].

### *D. Proposed Algorithm*

The ASDFS algorithm consists of three main stages: feature extraction using a sparse autoencoder, fine-tuning with a DT-DNN, and pathway-driven interpretability analysis. The following pseudocode outlines the ASDFS steps.

```
Algorithm 1: Adaptive Sparse Deep Feature
Selection (ASDFS)
 - Input: Gene expression matrix X with
   samples N and genes G, cancer label y.
 - Output: Selected gene subset S and
   classification model.
Data preprocessing
 - Normalize X by log-transformation and
   scaling.
 - Apply batch effect correction (e.g.,
   ComBat) if necessary.
 - Remove outliers using z-score
   filtering.
Stage 1: Sparse autoencoder for feature
extraction
 - Initialize sparse autoencoder
   parameters, including the bottleneck
```

```
    layer size and regularization
    coefficient.
  - Train the sparse autoencoder on X to
    obtain the latent representation Z.
  - Select top features in Z based on
    their contribution to the latent
    space, retaining fewer than 100 genes.
Stage 2: DT-DNN for feature refinement
  - Input: Selected features from Z.
  - Initialize DT-DNN with adaptive
    sparsity in the gene-weighting layer.
  - Define dual-target loss function [30].
  - Optimize the DT-DNN parameters using
    the Adam optimizer and apply early
    stopping based on validation
    performance.
Stage 3: Gene subset selection and pathway
analysis
  - Extract genes with the highest weights
    in the DT-DNN.
  - Perform pathway enrichment analysis
    (e.g., KEGG, Reactome) to validate
    biological relevance.
Classification and evaluation
  - Train SVM and Random Forest (RF)
    classifiers using the selected gene
    subset S.
  - Evaluate using fivefold cross-
    validation and report accuracy,
    precision, recall, F1-score, and AUC.
End algorithm
```

## III. RESULTS AND DISCUSSION

The ASDFS method, as outlined above, was applied to datasets representing ovarian, prostate, and lung cancers. This approach aimed to reduce dimensionality, enhance classification accuracy, and validate the biological significance of selected genes through pathway analysis. The following sections present the results of these analyses, beginning with the evaluation of feature selection and classification performance, followed by pathway enrichment analysis to establish biological relevance.

### *A. Feature Selection Results*

The ASDFS method achieved substantial dimensionality reduction, retaining fewer than 31 genes across all datasets while preserving high classification performance. Table I shows over 99.7% gene reduction in ovarian, prostate, and lung cancer datasets, demonstrating ASDFS's ability to isolate essential genes and enhance computational efficiency and interpretability. Reducing thousands of genes to fewer than 31 underscores ASDFS's effectiveness in handling high-dimensional genomic data. This reduction supports clinical and research objectives by enabling manageable gene panels that retain vital biological information.

### *B. Classification Performance*

Table II presents the classification performance of SVM, RF, K-Nearest Neighbors (KNN), Decision Tree (DT), and Linear Discriminant Analysis (LDA) in terms of accuracy, precision, recall, F1-score, and AUC. SVM consistently outperformed other classifiers across all datasets, leveraging ASDFS-selected features with L1 regularization and a dual-target loss function to maximize classification outcomes.

TABLE I. GENE REDUCTION BY ASDFS

| Dataset | Initial genes | Selected genes | Percentage reduction |
|---|---|---|---|
| Ovarian cancer | 15,154 | 29 | 99.8% |
| Prostate cancer | 12,600 | 25 | 99.8% |
| Lung cancer | 12,533 | 31 | 99.7% |

TABLE II. CLASSIFICATION RESULTS FOR ASDFS-SELECTED GENES VIA DIFFERENT CLASSIFIERS

| Dataset | Classifier | Acc. (%) | Pre. (%) | Recall (%) | F1-score | AUC |
|---|---|---|---|---|---|---|
| Ovarian cancer | SVM | 99.9 | 99.6 | 99.8 | 0.99 | 0.99 |
| | RF | 98.9 | 98.8 | 99.0 | 0.98 | 0.98 |
| | KNN | 98.5 | 98.4 | 98.7 | 0.98 | 0.98 |
| | DT | 97.8 | 97.6 | 98.0 | 0.97 | 0.97 |
| | LDA | 97.0 | 96.8 | 97.2 | 0.97 | 0.96 |
| Prostate cancer | SVM | 100 | 100 | 100 | 1.00 | 1.00 |
| | RF | 99.5 | 99.4 | 99.7 | 0.99 | 0.99 |
| | KNN | 98.5 | 98.4 | 98.7 | 0.98 | 0.98 |
| | DT | 97.8 | 97.6 | 98.0 | 0.97 | 0.97 |
| | LDA | 97.0 | 96.8 | 97.2 | 0.97 | 0.96 |
| Lung cancer | SVM | 99.8 | 99.6 | 100.0 | 0.99 | 0.99 |
| | RF | 99.5 | 99.4 | 99.6 | 0.99 | 0.99 |
| | KNN | 98.5 | 98.4 | 98.7 | 0.98 | 0.98 |
| | DT | 97.8 | 97.6 | 98.0 | 0.97 | 0.97 |
| | LDA | 97.0 | 96.8 | 97.2 | 0.97 | 0.96 |

For ovarian cancer, SVM outperformed all other classifiers with 99.9% accuracy, 99.6% precision, 99.8% recall, 0.99 F1-score, and 0.99 AUC. RF followed with 98.9% accuracy and 0.98 F1-score, whereas KNN showed a similar trend. DT and LDA had lower performance, with LDA scoring the lowest AUC (0.96).

In prostate cancer, SVM achieved perfect classification with 100% for all metrics, demonstrating superior generalization. RF performed well with 99.5% accuracy and an F1-score of 0.99, but fell slightly short of SVM. KNN, DT, and LDA maintained similar performance trends, with LDA again showing the lowest AUC.

For lung cancer, SVM achieved 99.8% accuracy, 99.6% precision, 100% recall, 0.99 F1-score, and 0.99 AUC. RF followed closely, matching the F1-score and AUC, but with slightly lower recall (99.6%). KNN, DT, and LDA showed the expected decrease in performance, with LDA having the lowest AUC (0.96).

SVM outperformed the other classifiers in all metrics, particularly excelling in recall and AUC, making it the most reliable model for cancer classification. RF remained competitive but was slightly less effective at feature reduction. KNN struggled with high-dimensional data, whereas DT suffered from overfitting. LDA, constrained by its linear assumptions, consistently had the lowest AUC.

### *C. Biological Relevance of Selected Genes*

The biological relevance of the genes selected by ASDFS was validated through their involvement in key cancer-related pathways. Table III highlights these enriched pathways and notable genes for each cancer type, which are further interpreted below.

TABLE III. ENRICHED PATHWAYS BY ASDFS

| Dataset | Pathway | Notable genes |
|---|---|---|
| Ovarian cancer | PI3K-Akt, MAPK | AKT1, MAPK3, PTEN |
| Prostate cancer | Androgen receptor, Cell cycle | AR, TP53, CCND1 |
| Lung cancer | Wnt, EGFR | EGFR, APC, WNT7B |

The enriched pathways identified across cancer types highlight key molecular mechanisms underlying tumor progression. For ovarian cancer, the PI3K-Akt and MAPK signaling pathways emerged as significant, with key genes such as AKT1, MAPK3, and PTEN prominently featured. The PI3K-Akt pathway regulates cell growth, survival, and metabolism, frequently driving tumor progression through aberrant activation. Similarly, the MAPK pathway governs cell proliferation and differentiation, often contributing to therapy resistance. The presence of PTEN, a well-known tumor suppressor, underscores the disruption of tumor-suppressing mechanisms in ovarian cancer development.

Building on these findings, prostate cancer showed enrichment in androgen receptor signaling and cell cycle regulation pathways, highlighting genes such as AR, TP53, and CCND1. The androgen receptor pathway is central to prostate cancer progression, promoting cell proliferation under hormonal stimulation. TP53, which is frequently mutated in prostate cancer, plays a key role in apoptosis and DNA repair, whereas CCND1 facilitates cell cycle progression, driving uncontrolled cancer cell division.

Extending this analysis, lung cancer was linked to Wnt and EGFR signaling pathways, with EGFR, APC, and WNT7B as key players. EGFR mutations and overexpression drive aggressive tumor behavior in lung adenocarcinoma, whereas the Wnt pathway, involving APC and WNT7B, contributes to tumor growth and metastasis through aberrant signaling. The interplay between these pathways underscores their importance in lung cancer biology and their potential as therapeutic targets.

### *D. Comparative Analysis with Baseline Methods*

ASDFS was evaluated against four baseline feature selection methods: PCA with Grey Wolf Optimizer (PCA-GWO) [31], Random Multi-Subspace Based ReliefF (RBEFF) [32], RFE [33], and Information Gain with Minimum Redundancy Maximum Relevance (mRMR-IG) [34]. In the performance comparison presented in Table IV, ovarian, prostate, and lung cancer datasets were used with classifiers including SVM, RF, KNN, DT, and LDA.

For ovarian cancer, PCA-GWO achieved 99.2% accuracy and 0.99 F1-score with 29 genes using SVM. RBEFF performed best with KNN, reaching 99.5% accuracy and 0.99 F1-score with 30 genes. RFE peaked at 98.5% accuracy with KNN using 30 genes, whereas mRMR-IG reached 99.0% accuracy with LDA using 35 genes.

TABLE IV. PERFORMANCE COMPARISON WITH TRADITIONAL METHODS

| Method | Dataset | Classifier | Accuracy (%) | Selected genes | F1-score |
|---|---|---|---|---|---|
| PCA-GWO [31] | Ovarian cancer | SVM | 99.2 | 29 | 0.99 |
| | | RF | 98.9 | 29 | 0.98 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Prostate cancer | SVM | 99.2 | 31 | 0.92 |
| | | RF | 92.5 | 20 | 0.91 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Lung cancer | SVM | 89.1 | 20 | 0.90 |
| | | RF | 88.6 | 20 | 0.89 |
| | | KNN | 99.5 | 30 | 0.99 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| RBEFF [32] | Ovarian cancer | SVM | 88.6 | 33 | 0.89 |
| | | RF | 87.9 | 22 | 0.88 |
| | | KNN | 99.5 | 30 | 0.99 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Prostate cancer | SVM | 91.0 | 45 | 0.91 |
| | | RF | 90.1 | 15 | 0.90 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 100 | 35 | 1.00 |
| | Lung cancer | SVM | 85.7 | 73 | 0.87 |
| | | RF | 84.3 | 47 | 0.86 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| RFE [33] | Ovarian cancer | SVM | 92.3 | 32 | 0.92 |
| | | RF | 91.5 | 27 | 0.92 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Prostate cancer | SVM | 94.7 | 48 | 0.94 |
| | | RF | 93.8 | 15 | 0.93 |
| | | KNN | 99.5 | 30 | 0.99 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Lung cancer | SVM | 91.2 | 22 | 0.91 |
| | | RF | 90.7 | 31 | 0.90 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| mRMR-IG [34] | Ovarian cancer | SVM | 95.0 | 23 | 0.95 |
| | | RF | 94.3 | 28 | 0.95 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 99.0 | 35 | 0.99 |
| | Prostate cancer | SVM | 96.0 | 37 | 0.96 |
| | | RF | 95.2 | 26 | 0.95 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |
| | Lung cancer | SVM | 99.7 | 41 | 0.97 |
| | | RF | 93.5 | 35 | 0.94 |
| | | KNN | 98.5 | 30 | 0.98 |
| | | DT | 97.8 | 32 | 0.97 |
| | | LDA | 97.0 | 35 | 0.96 |

For ovarian cancer, PCA-GWO achieved 99.2% accuracy and 0.99 F1-score with 29 genes using SVM. RBEFF performed best with KNN, reaching 99.5% accuracy and 0.99 F1-score with 30 genes. RFE peaked at 98.5% accuracy with KNN using 30 genes, whereas mRMR-IG reached 99.0% accuracy with LDA using 35 genes.

For prostate cancer, mRMR-IG achieved 96.0% accuracy and 0.96 F1-score with SVM using 37 genes. PCA-GWO reached 99.2% accuracy with SVM but had a lower F1-score (0.92). RBEFF performed best with LDA, achieving 100% accuracy and 1.00 F1-score using 35 genes, but underperformed with other classifiers. RFE peaked with KNN, reaching 99.5% accuracy with 30 genes, whereas its SVM performance was lower (94.7% accuracy).

For lung cancer, mRMR-IG achieved 99.7% accuracy and 0.97 F1-score with SVM using 41 genes. PCA-GWO peaked with KNN at 99.5% accuracy and 0.99 F1-score using 30 genes, but had lower SVM results (89.1% accuracy). RBEFF also performed best with KNN (98.5% accuracy, 0.98 F1-score), but required 73 genes and had weak SVM results (85.7% accuracy). RFE achieved 98.5% accuracy with KNN using 30 genes, while its SVM performance was lower (91.2% accuracy).

ASDFS outperformed all methods by achieving higher or matching top accuracies and F1-scores while selecting fewer genes. It maintained stability across classifiers, especially with SVM, while avoiding the high number of features observed in RBEFF and mRMR-IG.

### *E. Discussion*

The ASDFS method advances gene-based cancer classification by integrating machine learning with biological validation, focusing on pathway enrichment, regularization, and a DT-DNN to achieve high accuracy with minimal gene sets. ASDFS delivers strong performance, achieving 99.92%, 100.00%, and 99.82% accuracy for ovarian, prostate, and lung cancers (SVM). Its DT-DNN and L1 regularization enhance interpretability, whereas early stopping and Bayesian optimization increase efficiency and prevent overfitting.

ASDFS consistently outperforms existing methods. For ovarian cancer, it outperforms the method in [35] (99.48%) by integrating pathway enrichment to identify key genes. For prostate cancer, it outperforms PCA-based [36] (96.46%) and swarm optimization [37] (99.48%), achieving 100.00% (SVM) and 99.50% (RF) accuracy using only 25 genes. For lung cancer, it outperforms STFT [38] (96.13%) and swarm intelligence [39] (99.10%), reaching 99.82% (SVM) with 31 genes by capturing complex gene interactions.

ASDFS highlights key cancer pathways: PI3K-Akt and MAPK (ovarian), androgen receptor (prostate), and Wnt/EGFR (lung). Selecting biologically relevant genes ensures high accuracy with minimal features. Despite its strengths, ASDFS requires significant computational resources which limits its use in real time. Future work will focus on lightweight architectures, cloud-based deployment, and validating performance on diverse cohorts with multi-omics integration. Incorporating explainability frameworks (such as SHAP values) will further enhance model transparency and clinical trust.

## IV. CONCLUSION

The effectiveness of Adaptive Sparse Deep Feature Selection (ASDFS) is driven by its sparsity-enforcing L1 regularization, which selects biologically significant genes, and its dual-target loss function, which balances accuracy and sparsity. This design reduces computational overhead while enabling real-time data processing. By incorporating pathway enrichment analysis, ASDFS maps selected genes to key cancer pathways, such as PI3K-Akt, MAPK, and androgen receptor signaling, offering critical insights for diagnosis and therapy.

Compared to methods such as Principal Component Analysis with Grey Wolf Optimizer (PCA-GWO), Recursive Feature Elimination (RFE), and Minimum Redundancy Maximum Relevance (mRMR), ASDFS demonstrates superior performance by capturing non-linear gene interactions with fewer features. Despite its computational requirements, future enhancements such as lightweight architectures and cloud-based solutions can further improve scalability. In addition, validating ASDFS with external datasets and integrating multi-omics data, such as transcriptomics and proteomics, will broaden its applicability. Independent cohort testing and user-friendly diagnostic tools will strengthen its clinical utility. In conclusion, ASDFS represents a next-generation solution for gene-based cancer classification, offering both diagnostic precision and biological insight, with promising potential for advancements in precision medicine.

## REFERENCES

[1] M. Khalsan *et al.*, "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," *IEEE Access*, vol. 10, pp. 27522–27534, 2022, https://doi.org/10.1109/ACCESS.2022.3146312.

[2] S. Z. Ahammed, R. Baskar, and G. Nalinipriya, "Detection of Lung Cancer Using Multi-Stage Image Processing and Advanced Deep Learning InceptiMultiLayer-Net Model," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 4, pp. 714–727, Aug. 2024, https://doi.org/10.22266/ijies2024.0831.54.

[3] S. H. Bouazza, "A Deep Ensemble Gene Selection and Attention-guided Classification Framework for Robust Cancer Diagnosis from Microarray Data," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20235–20241, Feb. 2025, https://doi.org/10.48084/etasr.9476.

[4] W. L. Al-Yaseen, A. Jehad, Q. A. Abed, and A. K. Idrees, "The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, pp. 190–200, Apr. 2021, https://doi.org/10.22266/ijies2021.0430.17.

[5] 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, Art. no. 952709, https://doi.org/10.3389/fphys.2022.952709.

[6] E. Capobianco, "High-dimensional role of AI and machine learning in cancer research," *British Journal of Cancer*, vol. 126, no. 4, pp. 523–532, Mar. 2022, https://doi.org/10.1038/s41416-021-01689-z.

[7] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, Jan. 2022, Art. no. 105051, https://doi.org/10.1016/j.compbiomed.2021.105051.

[8] M. A. Siddiqi and W. Pak, "Optimizing Filter-Based Feature Selection Method Flow for Intrusion Detection System," *Electronics*, vol. 9, no.

12, Dec. 2020, Art. no. 2114, https://doi.org/10.3390/electronics9122114.

[9] H. Liu and R. Setiono, "Feature Selection and Classification – A Probabilistic Wrapper Approach," in *Industrial and Engineering Applications or Artificial Intelligence and Expert Systems*, T. Tanaka, S. Ohsuga, and A. Moonis, Eds. Boca Raton, FL, USA: CRC Press, 1997, pp. 419–424, https://doi.org/10.1201/9780429332111-72.

[10] S. H. Bouazza, "Optimized Machine Learning for Cancer Classification via Three-Stage Gene Selection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21093-21099, 2025, https://doi.org/10.48084/etasr.9473.

[11] 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.

[12] M. Akçakaya, B. Yaman, H. Chung, and J. C. Ye, "Unsupervised Deep Learning Methods for Biological Image Reconstruction and Enhancement: An overview from a signal processing perspective," *IEEE Signal Processing Magazine*, vol. 39, no. 2, pp. 28–44, Mar. 2022, https://doi.org/10.1109/MSP.2021.3119273.

[13] H. A. Gündüz *et al.*, "A self-supervised deep learning method for data-efficient training in genomics," *Communications Biology*, vol. 6, no. 1, pp. 1–12, Sep. 2023, https://doi.org/10.1038/s42003-023-05310-2.

[14] J. Sun and Y. Xia, "Pretreating and normalizing metabolomics data for statistical analysis," *Genes & Diseases*, vol. 11, no. 3, May 2024, Art. no. 100979, https://doi.org/10.1016/j.gendis.2023.04.018.

[15] M. Ligero *et al.*, "Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis," *European Radiology*, vol. 31, no. 3, pp. 1460–1470, Mar. 2021, https://doi.org/10.1007/s00330-020-07174-0.

[16] M. V. Bhargavi and V. Sireesha, "A COMPARATIVE STUDY FOR STATISTICAL OUTLIER DETECTION USING COLON CANCER DATA," *Advances and Applications in Statistics*, vol. 72, no. 1, pp. 41–54, Jan. 2022, https://doi.org/10.17654/0972361722003.

[17] E. F. Petricoin *et al.*, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002, https://doi.org/10.1016/S0140-6736(02)07746-2.

[18] 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.

[19] 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, Sep. 2002.

[20] A. Aldhahab, S. Ibrahim, and W. B. Mikhael, "Stacked Sparse Autoencoder and Softmax Classifier Framework to Classify MRI of Brain Tumor Images," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 3, pp. 268–279, Jun. 2020, https://doi.org/10.22266/ijies2020.0630.25.

[21] J. Li, M. Zhang, Q. Zhang, and D. Wang, "DT-LNS: Digital-Twin-Based Low-Risk Network Slicing Using Safe Reinforcement Learning," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 39606–39625, Dec. 2024, https://doi.org/10.1109/JIOT.2024.3445177.

[22] M. Yousef, E. Ülgen, and O. U. Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis," *PeerJ Computer Science*, vol. 7, Feb. 2021, Art. no. e336, https://doi.org/10.7717/peerj-cs.336.

[23] H. Nguyen, V.-D. Pham, H. Nguyen, B. Tran, J. Petereit, and T. Nguyen, "CCPA: cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and Reactome," *Briefings in Bioinformatics*, vol. 25, no. Supplement_1, Jul. 2024, Art. no. bbae222, https://doi.org/10.1093/bib/bbae222.

[24] R. S. Khairy, A. S. Hussein, and H. TH. S. ALRikabi, "The Detection of Counterfeit Banknotes Using Ensemble Learning Techniques of AdaBoost and Voting," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 1, pp. 326–339, Feb. 2021, https://doi.org/10.22266/ijies2021.0228.31.

[25] D. Yifan, L. Jialin, and F. Boxi, "Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost," in *2021 International Conference on Communications, Information System and Computer Engineering*, Beijing, China, 2021, pp. 716–719, https://doi.org/10.1109/CISCE52179.2021.9445847.

[26] P. E. Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021, https://doi.org/10.1109/ACCESS.2021.3123090.

[27] A. Yaqoob, R. Musheer Aziz, and N. K. verma, "Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review," *Human-Centric Intelligent Systems*, vol. 3, no. 4, pp. 588–615, Dec. 2023, https://doi.org/10.1007/s44230-023-00041-3.

[28] E. Goceri, "Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets," *International Journal of Imaging Systems and Technology*, vol. 33, no. 5, pp. 1727–1744, Sep. 2023, https://doi.org/10.1002/ima.22890.

[29] A. T. Alhasani, H. Alkattan, A. A. Subhi, E.-S. M. El-Kenawy, and M. M. Eid, "A Comparative Analysis of Methods for Detecting and Diagnosing Breast Cancer Based on Data Mining," *Journal of Artificial Intelligence and Metaheuristics*, vol. 4, no. 2, pp. 08–17, 2023, https://doi.org/10.54216/JAIM.040201.

[30] F. Hou, M. Fang, T. Xianghuan Luo, X. Fan, and Y. Guo, "Dual-Task GPR Method: Improved Generative Adversarial Clutter Suppression Network and Adaptive Target Localization Algorithm in GPR Image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024, https://doi.org/10.1109/TGRS.2024.3425890.

[31] H. Basak, R. Kundu, S. Chakraborty, and N. Das, "Cervical Cytology Classification Using PCA and GWO Enhanced Deep Features Selection," *SN Computer Science*, vol. 2, no. 5, Jul. 2021, Art. no. 369, https://doi.org/10.1007/s42979-021-00741-2.

[32] B. Zhang, Y. Li, and Z. Chai, "A novel random multi-subspace based ReliefF for feature selection," *Knowledge-Based Systems*, vol. 252, Sep. 2022, Art. no. 109400, https://doi.org/10.1016/j.knosys.2022.109400.

[33] R. K. Sachdeva, P. Bathla, P. Rani, V. Kukreja, and R. Ahuja, "A Systematic Method for Breast Cancer Classification using RFE Feature Selection," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering*, Greater Noida, India, 2022, pp. 1673–1676, https://doi.org/10.1109/ICACITE53722.2022.9823464.

[34] 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, Apr. 2021, Art. no. e0249094, https://doi.org/10.1371/journal.pone.0249094.

[35] S. K. Prabhakar and S.-W. Lee, "An Integrated Approach for Ovarian Cancer Classification With the Application of Stochastic Optimization," *IEEE Access*, vol. 8, pp. 127866–127882, 2020, https://doi.org/10.1109/ACCESS.2020.3006154.

[36] N. K, H. Rajaguru, and P. Rajkumar, "Microarray Prostate Cancer Classification using Eminent Genes," in *2021 Smart Technologies, Communication and Robotics*, Sathyamangalam, India, 2021, pp. 1–5, https://doi.org/10.1109/STCR51658.2021.9588811.

[37] S. K. Prabhakar and S.-W. Lee, "Transformation Based Tri-Level Feature Selection Approach Using Wavelets and Swarm Computing for Prostate Cancer Classification," *IEEE Access*, vol. 8, pp. 127462–127476, 2020, https://doi.org/10.1109/ACCESS.2020.3006197.

[38] M. S. Karthika, H. Rajaguru, and A. R. Nair, "Analysis of Machine Learning Classifiers for the Detection of Lung Cancer from Micro Array Gene Data," in *2023 Third International Conference on Smart Technologies, Communication and Robotics*, Sathyamangalam, India, 2023, pp. 1–6, https://doi.org/10.1109/STCR59085.2023.10396899.

[39] S. K. Prabhakar, H. Rajaguru, and D.-O. Won, "A Holistic Performance Comparison for Lung Cancer Classification Using Swarm Intelligence Techniques," *Journal of Healthcare Engineering*, vol. 2021, no. 1, 2021, Art. no. 6680424, https://doi.org/10.1155/2021/6680424.