Hybrid Statistical-Spectral Sparse Feature Selection with Optimization for Robust and Generalizable Lung Adenocarcinoma Classification

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

Received: 9 April 2025 | Revised: 28 April 2025 and 12 May 2025 | Accepted: 15 May 2025

 DOI: https://doi.org/10.48084/etasr.11388

ABSTRACT

High dimensionality, redundant features, and poor cross-dataset generalization hinder Lung Adenocarcinoma (LUAD) classification using gene expression data. This study proposes Hybrid Statistical-Spectral Sparse Feature Selection with Optimization (HS3FS+), a novel framework that integrates Mutual Information (MI) and Kullback-Leibler (KL) divergence for feature ranking, Kernel Principal Component Analysis (KPCA) for nonlinear transformation, pathway-guided filtering for biological validation, and Genetic Algorithm (GA)-based optimization for feature selection. The framework was validated on four independent datasets: The Cancer Genome Atlas (TCGA)-LUAD, Gene Expression Omnibus (GEO) datasets GSE19188 and GSE37745, and TCGA-Lung Squamous Cell Carcinoma (TCGA-LUSC), ensuring robust cross-platform evaluation. HS3FS+ achieved classification accuracy of 98.3% on TCGA-LUAD, 97.1% on GSE19188, 96.0% on GSE37745, and 94.8% on TCGA-LUSC. The selected gene signatures exhibit strong concordance with established LUAD biomarkers, supporting both biological relevance and model interpretability. Additionally, the method demonstrated a fivefold reduction in computational time compared to Deep Learning (DL)–based feature selection approaches. These findings confirm HS3FS+ as a robust, interpretable, and scalable solution for LUAD classification, with potential applications in biomarker discovery and precision oncology.

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

I. INTRODUCTION

Lung Adenocarcinoma (LUAD) [1] is the most common subtype of Non-Small Cell Lung Cancer (NSCLC) [2], accounting for approximately 40% of all lung cancer cases worldwide [3]. Despite advancements in targeted therapies and immunotherapy, early and accurate diagnosis remains a critical challenge, largely due to LUAD's complex molecular heterogeneity [4]. Gene expression profiling has emerged as a powerful tool for LUAD classification and biomarker discovery, offering the potential for early detection and personalized treatment [5]. However, the high dimensionality of gene expression data presents significant challenges in developing robust Machine Learning (ML) models, as only a small subset of genes is truly informative for classification [6-7]. Feature selection is a key step in gene-based cancer classification, as it reduces dimensionality, improves classification accuracy, and enhances biological interpretability [8]. Existing feature selection techniques can be categorized into filter-based, wrapper-based, and embedded methods [9]. Filter methods [10, 11], such as Mutual Information (MI), ReliefF, and minimum Redundancy Maximum Relevance (mRMR), independently rank genes based on statistical criteria. These methods are computationally efficient but fail to capture complex gene dependencies. Wrapper methods, such as Recursive Feature Elimination (RFE) with Support Vector Machines (SVMs), iteratively refine gene subsets based on classification performance, but they suffer from high computational costs and overfitting risks [12, 13]. Embedded methods, including Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regression (RR), incorporate feature selection into model training but may not prioritize biologically meaningful genes [14, 15]. In recent years, hybrid feature selection methods have been developed to combine the strengths of different approaches [16], integrating filter and wrapper methods, ensuring that selected features are statistically relevant and classifier dependent. Genetic Algorithm (GA) [17] and Particle Swarm Optimization (PSO) [18] have been employed to refine selected feature subsets by optimizing classification accuracy and feature sparsity. However, these approaches often require extensive hyperparameter tuning and computational resources.

Recent Deep Learning (DL)-based feature selection approaches attempted to overcome these limitations by capturing nonlinear interactions among genes using autoencoders and attention-based models [19-21]. While promising, these methods require large, well-annotated datasets, lack interpretability, and impose no explicit biological constraints, limiting their clinical applicability. Moreover, a critical shortcoming across many existing approaches is poor generalization to external datasets. Models often achieve high accuracy on training data but fail to replicate performance on independent validation datasets obtained using different platforms, such as Ribonucleic Acid Sequencing (RNA-seq) and microarrays [6].

To address these challenges, this study proposes Hybrid Statistical-Spectral Sparse Feature Selection with Optimization (HS3FS+), a novel feature selection framework that integrates statistical, spectral, biological, and optimization-based techniques to enhance LUAD classification. The key innovations of HS3FS+ include:

- A hybrid multi-step feature selection strategy that combines MI and Kullback-Leibler (KL) divergence for feature ranking, Kernel Principal Component Analysis (KPCA) for nonlinear transformation, pathway-guided filtering for biological validation, and GA-based optimization for feature subset refinement.
- A robust cross-platform validation approach, evaluating HS3FS+ across four independent datasets: The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD), Gene Expression Omnibus (GEO) datasets GSE19188 and GSE37745, and TCGA Lung Squamous Cell Carcinoma (TCGA-LUSC), to assess generalization and classification stability.
- A biologically interpretable feature selection process, ensuring that the identified genes align with known LUAD biomarkers, making the model more relevant for clinical applications.

Experimental results confirm that HS3FS+ significantly outperforms traditional and DL-based feature selection methods, achieving higher classification accuracy, improved generalization across datasets, and lower computational complexity.

II. METHODS

The HS3FS+ framework is designed to enhance LUAD classification by integrating statistical, spectral, biological, and optimization-based feature selection techniques. Unlike conventional selection methods that rely solely on statistical measures or DL-based approaches requiring large labelled datasets, HS3FS+ optimally balances classification accuracy, biological interpretability, and computational efficiency. The framework consists of six key stages:

1. Data preprocessing and normalization.
2. Feature ranking using information-theoretic measures.
3. Spectral feature transformation with KPCA.
4. Pathway-guided feature filtering.
5. GA-based feature optimization.
6. Classification and performance evaluation.

A. Data Preprocessing and Normalization

This study utilizes four independent datasets to ensure robust evaluation across different experimental platforms and lung cancer subtypes:

- TCGA-LUAD (RNA-seq) [22].
- GEO datasets GSE19188 and GSE37745 (microarray-based gene expression) [23, 24].
- TCGA-LUSC (RNA-seq) [25].

Since these datasets originate from distinct platforms, batch effect correction, normalization, and feature alignment were applied.

1) Handling Missing Values and Outliers

To ensure data quality and reliability, missing values within the gene expression profiles were addressed through K-Nearest Neighbors (KNN) [26] imputation with k=5, effectively preserving local patterns and relationships among gene expression levels. Additionally, outlier detection was performed using the Interquartile Range (IQR) method [27], allowing for the identification and exclusion of extreme values. Specifically, data points exceeding three standard deviations from the mean were treated as noise and removed, thereby enhancing the robustness and reliability of subsequent analyses.

2) Normalization Techniques

To ensure consistency and comparability across datasets, distinct normalization strategies were employed based on the data type. For RNA-seq datasets (TCGA-LUAD and TCGA-LUSC), a $\log_2$ transformation was first applied to stabilize variance across gene expression levels [28], followed by quantile normalization to align distributional properties across samples [29]. Finally, Z-score standardization was utilized to scale features, promoting comparability across genes [30]. In contrast, for microarray datasets (GSE19188 and GSE37745), Robust Multi-Array Normalization (RMA) normalization was utilized to correct background noise and normalize probe-level intensities. Quantile normalization was then applied to harmonize expression distributions across arrays. To further mitigate potential platform-specific biases, ComBat batch effect correction was employed, enhancing the integration and reliability of multi-platform data.

3) Feature Alignment Across Datasets

To ensure a consistent feature space across all datasets, genes present in at least three out of four datasets were retained. This threshold balances data completeness and cross-platform consistency, avoiding excessive gene loss due to platform-specific dropout effects. Genes missing from one dataset but highly informative in others were retained to prevent loss of significant biomarkers. The final dataset consists of 12,150 genes.

B. Feature Ranking Using Information-Theoretic Measures

The feature selection process begins with an information-theoretic ranking that combines MI and KL divergence. MI quantifies the dependency between gene expression and cancer status, ensuring that highly informative genes are prioritized [31]:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where X is the random variable representing gene expression values, Y is the random variable representing cancer status (e.g., tumor vs. normal), $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$, $p(y)$ are the marginal probability distributions of X and Y.

KL-divergence measures the difference in gene expression distributions between tumor and normal samples [32]. Higher values indicate genes that exhibit significantly distinct expression patterns:

$$D_{KL}(P\|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (2)$$

where $P(i)$ and $Q(i)$ are the probability distribution of gene expression values in tumor samples and normal samples, respectively. Genes are ranked based on a combined MI-KL score, and the top 500 genes are retained for further analysis.

C. Spectral Feature Transformation with KPCA

To capture nonlinear dependencies among genes, KPCA is applied using a Radial Basis Function (RBF) kernel [33]:

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (3)$$

A grid search strategy was used to tune kernel width σ across {0.1, 0.5, 1, 5, 10}, with σ = 1 yielding the highest classification performance. The top 200 transformed features were retained for the next stage of selection.

D. Pathway-guided Feature Filtering

To ensure biological relevance, genes were filtered based on their documented associations with LUAD using curated databases, including Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome Pathway Database (RPD), and Gene Ontology (GO) biological processes. Genes without known LUAD associations were excluded. Notable retained biomarkers include Epidermal Growth Factor Receptor (EGFR), Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS), Tumor Protein p53 (TP53), Anaplastic Lymphoma Kinase (ALK), and Mesenchymal-Epithelial Transition (MET), which play crucial roles in LUAD progression.

E. GA-based Feature Optimization

A GA-based optimization strategy was employed to refine feature selection by optimizing a fitness function that balances classification accuracy and feature sparsity [34]:

$$Fitness = Accuracy - \lambda \cdot |S| \quad (4)$$

where $|S|$ represents the subset size, and λ is a penalty term controlling gene count.

To optimize the feature selection process, a GA was utilized with finely tuned hyperparameters. The population size was set at 50, providing a balanced trade-off between diversity and computational efficiency. A mutation rate of 0.1 was applied to introduce random variations and enhance exploration, while a crossover rate of 0.8 enabled effective combination of genetic material between parent solutions. The algorithm employed an early stopping criterion, terminating if no improvement in classification accuracy was observed over 10 consecutive generations, preventing overfitting and unnecessary iterations. After approximately 50 generations, the GA consistently converged to an optimal gene subset containing between 50 and 100 features, yielding a compact yet informative feature space for robust LUAD classification.

F. Classification and Statistical Significance Testing

The final gene subset selected by HS3FS$^+$ was evaluated using three classification models: Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Deep Neural Networks (DNNs). Classification performance was assessed using the following standard metrics [35]:

- Accuracy: The proportion of correctly classified samples:

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

where TP and TN are true positives and true negatives, while FP and FN represent false positives and false negatives.

- Generalization Loss (GL): Assesses the model's ability to generalize across datasets:

$$GL = Acc_{TGCA} - Acc_{TGSE} \quad (6)$$

where Acc_{TGCA} and Acc_{TGSE} represent classification accuracy on TCGA-LUAD and an external dataset (GSE19188, GSE37745, or TCGA-LUSC), respectively. Lower GL values indicate better cross-dataset generalization.

To evaluate the statistical significance of HS3FS+ performance, a paired t-test was conducted comparing its classification accuracy against baseline feature selection methods. The null hypothesis stated that no significant difference exists between HS3FS+ and the alternative methods, while the alternative hypothesis posited a statistically significant improvement. A significance threshold of $p<0.05$ was used.

III. RESULTS AND DISCUSSION

A. Classification Performance and Statistical Significance Analysis

Table I summarizes the classification performance of HS3FS+ and other baseline feature selection methods (mRMR, PCA, RFE, sparse autoencoders, GA) across four datasets: TCGA-LUAD, GSE19188, GSE37745, and TCGA-LUSC. The results demonstrate that HS3FS+ consistently outperforms all competing methods, achieving higher accuracy, superior AUC values.

A paired t-test was conducted to compare HS3FS+ with the GA method (best-performing baseline) across all datasets. The results confirm that HS3FS+ achieves a statistically significant

improvement (p-value < 0.01), indicating that the observed accuracy gains are unlikely to result from random variation.

TABLE I. PERFORMANCE COMPARISON OF FEATURE SELECTION METHODS (MEAN ± STANDARD DEVIATION OVER 5 RUNS)

Feature Selection Method	TCGA-LUAD Accuracy (%)	GSE19188 Accuracy (%)	GSE37745 Accuracy (%)	TCGA-LUSC Accuracy (%)	Generalization Loss (GL) (%)	95% Confidence Intervals (CI)
mRMR	87.2 ± 0.4	83.1 ± 0.5	82.7 ± 0.6	81.4 ± 0.6	-4.5 ± 0.3	(85.9, 88.5)
PCA	89.5 ± 0.3	85.4 ± 0.4	84.1 ± 0.4	82.9 ± 0.5	-3.9 ± 0.3	(88.2, 90.8)
RFE	91.8 ± 0.3	88.6 ± 0.4	87.3 ± 0.4	85.5 ± 0.5	-3.2 ± 0.3	(90.5, 93.1)
Sparse autoencoder	95.3 ± 0.2	93.0 ± 0.3	91.5 ± 0.4	89.2 ± 0.4	-2.3 ± 0.2	(94.0, 96.6)
GA	97.1 ± 0.2	95.5 ± 0.3	94.2 ± 0.3	92.3 ± 0.4	-1.6 ± 0.2	(96.0, 98.2)
HS3FS+ (Proposed)	**98.3 ± 0.2**	**97.1 ± 0.3**	**96.0 ± 0.3**	**94.8 ± 0.4**	**-0.7 ± 0.2**	**(97.5, 99.1)**

B. Generalization and Robustness Across Datasets

Compared to baseline methods, HS3FS+ achieves the lowest GL of -0.7%, confirming its ability to maintain high predictive power across datasets. Conventional approaches such as mRMR, PCA, and RFE exhibit generalization losses between -3.2% and -4.5%, indicating higher accuracy degradation when transitioning to external datasets. The improved generalization of HS3FS+ suggests that biologically validated features contribute to enhanced cross-dataset classification.

C. Computational Efficiency of HS3FS+ Across All Datasets

Computational efficiency is a critical consideration in feature selection, especially for large-scale genomic studies. Table II compares the average runtime and memory usage of various feature selection methods across the four datasets. All experiments were conducted in a controlled Linux-based high-performance computing environment, utilizing a 12-core Intel Xeon processor with 64 GB RAM. Where applicable, methods were executed with parallel processing enabled to exploit multi-threading capabilities and ensure fair evaluation. The proposed HS3FS+ model achieved the third-lowest average runtime (23.1 seconds) and the fourth-lowest average memory usage (420 MB), demonstrating a strong balance between computational cost and performance.

TABLE II. COMPUTATIONAL RESOURCE UTILIZATION OF FEATURE SELECTION METHODS ACROSS ALL DATASETS

Feature Selection Method	Average Runtime (s)	Average Memory Usage (MB)
mRMR	5.0 ± 0.2	130
PCA	7.9 ± 0.3	180
RFE	28.9 ± 0.4	300
Sparse autoencoder	119.1 ± 0.5	1500
GA	41.2 ± 0.4	890
HS3FS+ (Proposed)	23.1 ± 0.3	420

D. Error Analysis and Misclassification Rates Across All Datasets

To assess classification reliability, FPR and False Negative Rate (FNR) were recorded for each dataset. Table III presents the error rates across the four datasets. The proposed model across all datasets achieved an FPR ranging from 1.2-2.5% and an FNR ranging from 0.7-1.8%.

TABLE III. ERROR ANALYSIS OF HS3FS+ ACROSS ALL DATASETS

Dataset	False Positives (FP%)	False Negatives (FN%)
TCGA-LUAD	1.2 ± 0.1%	0.7 ± 0.1%
GSE19188	1.9 ± 0.2%	1.3 ± 0.2%
GSE37745	2.0 ± 0.2%	1.5 ± 0.2%
TCGA-LUSC	2.5 ± 0.3%	1.8 ± 0.3%

IV. CONCLUSION

This study proposed Hybrid Statistical-Spectral Sparse Feature Selection with Optimization (HS3FS+), a novel framework for gene-based Lung Adenocarcinoma (LUAD) classification. The method integrates Mutual Information (MI) and Kullback-Leibler (KL) divergence for feature ranking, Kernel Principal Component Analysis (KPCA) for nonlinear transformation, pathway-guided filtering for biological validation, and Genetic Algorithm (GA)-based optimization for feature selection. Unlike existing feature selection techniques, HS3FS+ ensures high classification accuracy while maintaining biological interpretability and computational efficiency.

Experimental evaluations across four independent datasets, The Cancer Genome Atlas (TCGA)-LUAD, GSE19188, GSE37745, and TCGA-Lung Squamous Cell Carcinoma (TCGA-LUSC), demonstrated that HS3FS+ consistently outperforms traditional and Deep Learning (DL)-based feature selection methods, achieving 98.3% accuracy on TCGA-LUAD, 97.1% on GSE19188, 96.0% on GSE37745, and 94.8% on TCGA-LUSC, with a minimal Generalization Loss (GL) of -0.7%. Additionally, the method significantly reduces False Positive Rate (FPR) and False Negative Rate (FNR), confirming its robustness in distinguishing cancerous from normal samples. Computational efficiency tests show that HS3FS+ is very competitive compared to DL-based selection methods, making it scalable for large-scale genomic studies.

Despite its advantages, HS3FS+ has certain limitations. The framework relies on predefined pathway databases for feature filtering, which may exclude novel biomarkers not yet annotated in existing repositories. Additionally, the method currently focuses on gene expression data alone, while multi-omics integration (e.g., DNA methylation, proteomics) could further enhance its predictive power. Moreover, further evaluation using larger patient cohorts and cross-laboratory

datasets would strengthen its generalizability and clinical applicability.

Future research should explore multi-omics data fusion to improve classification performance, as well as incorporate advanced explainability techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to enhance clinical interpretability. Additionally, optimization strategies such as reinforcement learning-based selection and federated learning could further improve feature selection efficiency and applicability in real-world clinical settings.

In conclusion, HS3FS+ provides a robust, interpretable, and computationally efficient solution for LUAD classification, offering potential applications in biomarker discovery and precision oncology. By integrating statistical, spectral, and biological selection techniques, the proposed framework offers a strong foundation for generalizable and biologically meaningful feature selection in cancer classification.

REFERENCES

[1] R. Fujikawa *et al.*, "Clinicopathologic and Genotypic Features of Lung Adenocarcinoma Characterized by the International Association for the Study of Lung Cancer Grading System," *Journal of Thoracic Oncology*, vol. 17, no. 5, pp. 700–707, May 2022, https://doi.org/10.1016/j.jtho.2022.02.005.

[2] J. W. Chen and J. Dhahbi, "Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods," *Scientific Reports*, vol. 11, no. 1, Jun. 2021, Art. no. 13323, https://doi.org/10.1038/s41598-021-92725-8.

[3] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, "The global burden of lung cancer: current status and future trends," *Nature Reviews Clinical Oncology*, vol. 20, no. 9, pp. 624–639, Sep. 2023, https://doi.org/10.1038/s41571-023-00798-3.

[4] D. Huang, Z. Li, T. Jiang, C. Yang, and N. Li, "Artificial intelligence in lung cancer: current applications, future perspectives, and challenges," *Frontiers in Oncology*, vol. 14, Dec. 2024, https://doi.org/10.3389/fonc.2024.1486310.

[5] S. Srivastava *et al.*, "Unveiling the potential of proteomic and genetic signatures for precision therapeutics in lung cancer management," *Cellular Signalling*, vol. 113, Jan. 2024, Art. no. 110932, https://doi.org/10.1016/j.cellsig.2023.110932.

[6] 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, Apr. 2025, https://doi.org/10.48084/etasr.9473.

[7] 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.

[8] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, and M. Oussalah, "Graph-based relevancy-redundancy gene selection method for cancer diagnosis," *Computers in Biology and Medicine*, vol. 147, Aug. 2022, Art. no. 105766, https://doi.org/10.1016/j.compbiomed.2022.105766.

[9] M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi, and R. Sarkar, "A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification," *Sensors*, vol. 21, no. 16, Aug. 2021, Art. no. 5571, https://doi.org/10.3390/s21165571.

[10] M. Farsi, "Filter-Based Feature Selection and Machine-Learning Classification of Cancer Data," *Intelligent Automation & Soft Computing*, vol. 28, no. 1, pp. 83–92, 2021, https://doi.org/10.32604/iasc.2021.015460.

[11] S. H. Bouazza and J. H. Bouazza, "Advanced Cancer Classification Using AI and Pattern Recognition Techniques," *ITM Web of Conferences*, vol. 69, 2024, Art. no. 02001, https://doi.org/10.1051/itmconf/20246902001.

[12] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, "A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches," *Complexity*, vol. 2022, no. 1, Jan. 2022, https://doi.org/10.1155/2022/8190814.

[13] N. S. Azman *et al.*, "Support Vector Machine – Recursive Feature Elimination for Feature Selection on Multi-omics Lung Cancer Data," *Progress In Microbes & Molecular Biology*, vol. 6, no. 1, Apr. 2023, https://doi.org/10.36877/pmmb.a0000327.

[14] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Computing and Applications*, vol. 33, no. 22, pp. 15091–15118, Nov. 2021, https://doi.org/10.1007/s00521-021-06406-8.

[15] A. B. Buriro and S. Kumar, "The Fisher Component-based Feature Selection Method," *Engineering, Technology & Applied Science Research*, vol. 12, no. 4, pp. 9023–9027, Aug. 2022, https://doi.org/10.48084/etasr.5137.

[16] D. K. Singh and M. Shrivastava, "Evolutionary Algorithm-based Feature Selection for an Intrusion Detection System," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7130–7134, Jun. 2021, https://doi.org/10.48084/etasr.4149.

[17] S. H. Bouazza and J. H. Bouazza, "Optimized colon cancer classification via feature selection and machine learning," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1476–1485, Apr. 2025, https://doi.org/10.11591/eei.v14i2.9270.

[18] L. Meenachi and S. Ramakrishnan, "Review on hybrid feature selection and classification of microarray gene expression data," *Data Fusion Techniques and Applications for Smart Healthcare*, pp. 319–340, 2024, https://doi.org/10.1016/b978-0-44-313233-9.00020-5.

[19] I. Zafar *et al.*, "Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine," *Biomedical Signal Processing and Control*, vol. 86, Sep. 2023, Art. no. 105263, https://doi.org/10.1016/j.bspc.2023.105263.

[20] G. Sokar, Z. Atashgahi, M. Pechenizkiy, and D. C. Mocanu, "Where to Pay Attention in Sparse Training for Feature Selection?" arXiv, 2022, https://doi.org/10.48550/ARXIV.2211.14627.

[21] R. Shanmugavelu and V. Ravi, "Enhancing Security in Healthcare Frameworks using Optimal Deep Learning-based Attack Detection and Classification for Medical Wireless Sensor Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21197–21202, Apr. 2025, https://doi.org/10.48084/etasr.9741.

[22] *Lung Adenocarcinoma*. (2025), National Cancer Institute. [Online]. Available: https://portal.gdc.cancer.gov/projects/TCGA-LUAD.

[23] *GSE19188*. (2009), NCBI GEO. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19188.

[24] *GSE37745*. (2012), NCBI GEO. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37745.

[25] *Lung Squamous Cell Carcinoma*. (2025), National Cancer Institute. [Online]. Available: https://portal.gdc.cancer.gov/projects/TCGA-LUSC.

[26] P. Keerin and T. Boongoen, "Improved KNN Imputation for Missing Values in Gene Expression Data," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 4009–4025, 2022, https://doi.org/10.32604/cmc.2022.020261.

[27] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, Mar. 2023, Art. no. 100164, https://doi.org/10.1016/j.dajour.2023.100164.

[28] M. Rezapour, S. J. Walker, D. A. Ornelles, P. M. McNutt, A. Atala, and M. N. Gurcan, "Analysis of gene expression dynamics and differential expression in viral infections using generalized linear models and quasi-likelihood methods," *Frontiers in Microbiology*, vol. 15, Apr. 2024, https://doi.org/10.3389/fmicb.2024.1342328.

[29] Y. Xia, "Statistical normalization methods in microbiome data with application to microbiome cancer research," *Gut Microbes*, vol. 15, no. 2, Dec. 2023, https://doi.org/10.1080/19490976.2023.2244139.

[30] R. D. Tihagam and S. Bhatnagar, "A multi-platform normalization method for meta-analysis of gene expression data," *Methods*, vol. 217, pp. 43–48, Sep. 2023, https://doi.org/10.1016/j.ymeth.2023.06.012.

[31] Z. Jandaghi, "Mutual Information-based Machine Learning with Microarray Cancer Data," Ph.D. dissertation, University of Georgia, Georgia, USA, 2022.

[32] S. Liu and W. Yao, "Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, https://doi.org/10.1186/s12859-022-04689-9.

[33] M. Ahsan, M. Mashuri, H. Khusna, and Wibawati, "Kernel principal component analysis (PCA) control chart for monitoring mixed non-linear variable and attribute quality characteristics," *Heliyon*, vol. 8, no. 6, Jun. 2022, Art. no. e09590, https://doi.org/10.1016/j.heliyon.2022.e09590.

[34] Z. Song, H. Wang, B. Xue, and M. Zhang, "Balancing Different Optimization Difficulty Between Objectives in Multiobjective Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 6, pp. 1824–1837, Dec. 2024, https://doi.org/10.1109/tevc.2023.3334233.

[35] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," in *Lecture Notes in Networks and Systems*, Cham, 2023, pp. 15–25, https://doi.org/10.1007/978-3-031-35314-7_2.