

# Deep learning-based feature selection for lung adenocarcinoma classification and biomarker discovery

Sara Haddou Bouazza<sup>1</sup>, Jihad Haddou Bouazza<sup>2</sup>

<sup>1</sup>LAMIGEP Laboratory, Moroccan School of Engineering Sciences (EMSI), Marrakech, Morocco

<sup>2</sup>Senior Full Stack Developer and Tech Lead, Nexular Corp., Casablanca, Morocco

## Article Info

### Article history:

Received Jan 29, 2025

Revised Aug 23, 2025

Accepted Sep 7, 2025

### Keywords:

Artificial intelligence

Cancer classification

Computer science

Feature selection

Machine learning

## ABSTRACT

Lung adenocarcinoma, a leading cause of cancer-related mortality, underscores the need for reliable diagnostic tools. This study proposes a robust multi-stage feature selection and classification framework for biomarker discovery, using the cancer genome atlas lung adenocarcinoma (TCGA-LUAD) as the primary dataset and GSE19188 for independent validation. The framework combines differential expression analysis (Wilcoxon rank-sum test), joint mutual information maximization (JMIM), and sparse autoencoder-based refinement to identify a compact and predictive set of five genes. These genes are involved in key lung cancer pathways, including epidermal growth factor receptor (EGFR) signaling, cell cycle regulation, and immune response, and include biomarkers such as surfactant protein A2 (SFTPA2), napsin an aspartic peptidase (NAPSA), and T-box transcription factor 4 (TBX4). The hybrid deep learning classifier achieved high accuracy (98.4%) and area under the receiver operating characteristic curve (AUC-ROC) (0.996) on TCGA-LUAD, with strong generalization on GSE19188 (accuracy: 96.7%, AUC-ROC: 0.993%). Overall, the framework offers an interpretable and effective solution for LUAD classification and biomarker identification.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Sara Haddou Bouazza

LAMIGEP Laboratory, Moroccan School of Engineering Sciences (EMSI)

Marrakech, Morocco

Email: sara.hb.sara@gmail.com

## 1. INTRODUCTION

Cancer remains a leading global cause of death, with lung cancer being the most prevalent and fatal subtype. Prognosis is often poor due to late diagnosis and tumor heterogeneity, underscoring the need for early and accurate detection. Gene expression profiling is a powerful tool for identifying diagnostic and prognostic markers, yet its high dimensionality and inherent noise complicate classification tasks [1]–[3]. To mitigate these challenges, feature selection is a crucial preprocessing step that improves model interpretability, reduces computational cost, and enhances classification accuracy. Traditional methods filter, wrapper, and embedded have shown potential but often suffer from redundancy, overfitting, and scalability limitations in high-throughput data contexts [4]–[6]. Recent machine learning advancements have led to hybrid and ensemble-based feature selection techniques that improve robustness and accuracy. However, many still neglect biological pathway relevance and gene regulatory interactions, limiting their clinical applicability [7]–[10].

This study addresses these issues by leveraging gene expression data from the cancer genome atlas (TCGA), focusing on the lung adenocarcinoma (LUAD) dataset. TCGA provides large-scale molecular and clinical data, enabling biologically grounded and statistically rigorous biomarker discovery. Our proposed method combines advanced ensemble learning with feature engineering to identify a compact, interpretable

gene subset relevant to LUAD classification. Unlike traditional approaches, our framework prioritizes both predictive performance and biological insight, supporting personalized cancer diagnostics.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 presents our methodology. Section 4 details the experimental setup and results. Section 5 concludes with key contributions and future directions.

## 2. RELATED WORK

Feature selection is essential for analyzing high-dimensional gene expression data in cancer classification. Recent advances, such as the signal-to-noise ratio-optimized gene selection and clustering for cancer classification (SNR-OGSCC) method in [11] have improved accuracy by combining optimized gene selection with clustering, effectively reducing redundancy and enhancing computational efficiency. However, traditional methods filter, wrapper, embedded, and hybrid still face issues like redundancy, overfitting, limited biological relevance, and scalability, limiting their clinical utility [12], [13].

Filter methods like minimum redundancy maximum relevance (mRMR) are computationally efficient but overlook complex feature interactions [14]–[16], while wrapper approaches such as genetic algorithms offer high accuracy at the cost of increased computational demands and risk of overfitting [17]. Embedded techniques like support vector machine-recursive feature elimination (SVM-RFE) and retron library recombineering (RLR) integrate selection into training but often rely on linear assumptions [18]. Hybrid and ensemble methods, such as extreme gradient boosting (XGBoost) combined with genetic algorithms, strike a balance between accuracy and efficiency but face interpretability and complexity challenges [19].

In lung cancer studies, ensemble and embedded methods have achieved up to 97.99% accuracy in LUAD biomarker identification [20], yet many approaches emphasize prediction over biological interpretability. Techniques like SNR-OGSCC address this by minimizing redundancy and selecting minimal yet informative gene sets. Our proposed multi-stage framework builds on these efforts by integrating redundancy reduction, pathway-based biological validation, and attention-guided deep learning to enhance interpretability. Validated on independent datasets, the framework demonstrates strong robustness and scalability, offering a meaningful advance toward biologically grounded cancer classification for precision oncology.

## 3. METHOD

This section presents the multi-stage feature selection and classification framework for lung cancer gene expression analysis, utilizing the LUAD dataset from TCGA. The framework aims to identify a compact, biologically relevant subset of genes using statistical and deep learning techniques. This will be explained as follows.

### 3.1. Dataset description

This study uses RNA sequencing (RNA-Seq) data from the TCGA-LUAD dataset, which includes gene expression profiles from 585 lung adenocarcinoma and 59 normal lung tissue samples. RNA-Seq provides high-resolution, genome-wide insights into gene regulation and cellular function, with the dataset covering around 20,500 genes and accompanied by rich clinical annotations. Leveraging this resource, we apply a multi-stage feature selection and classification framework to identify a compact, biologically meaningful gene subset predictive of LUAD. This approach supports biomarker discovery and offers a robust, generalizable method for improving lung cancer diagnosis and treatment.

### 3.2. Data preprocessing

To ensure data quality and consistency, several preprocessing steps were applied. Gene expression values were log2-transformed using fragments per kilobase per million mapped fragments (FPKM+1) to stabilize variance and reduce heteroscedasticity across samples [21]. Batch effects from technical variations were corrected using the combatting batch effects (ComBat) algorithm, which applies an empirical Bayes approach to preserve biological signals while minimizing technical noise [22]. Genes with low expression (under the 10th percentile) were removed, and outliers identified via Mahalanobis distance exceeding the 95<sup>th</sup> percentile were excluded to reduce noise and improve robustness [23]. These steps produced a high-quality dataset suitable for reliable downstream analysis.

### 3.3. Multi-stage feature selection framework

To overcome the challenges of high dimensionality, noise, and redundancy in RNA-Seq data, we designed a multi-stage feature selection framework that combines statistical analysis, entropy-based ranking, and deep learning refinement. This approach ensures a robust, interpretable, and biologically relevant gene subset. This will be explained as follows.

### 3.3.1. Stage 1: statistical relevance and stability analysis

In the first stage, differentially expressed genes between tumor and normal samples were identified using the Wilcoxon rank-sum test, a non-parametric method suitable for RNA-Seq data [24]. To control false positives, the Benjamini-Hochberg procedure was applied, retaining genes with a false discovery rate (FDR) under 0.05 [25]. To further improve robustness and reduce noise sensitivity, stability selection via bootstrap resampling was performed. Genes consistently identified as significant in at least 95% of 100 random subsets were retained, ensuring stability across sampling variations [26].

### 3.3.2. Stage 2: entropy-driven feature ranking

In the second stage, genes from stage 1 were ranked using joint mutual information maximization (JMIM) [27], which evaluates each gene's ability to reduce uncertainty about class labels while minimizing redundancy with previously selected features [28]. This ensures selection of features that are both relevant and complementary. The top 200 genes with the highest mutual information (MI) scores were retained for the next stage, a threshold chosen to balance dimensionality reduction with biological diversity and interpretability.

### 3.3.3. Stage 3: sparse autoencoder for feature refinement

The final stage employed a sparse autoencoder to refine the selected features further. As an unsupervised deep learning model, the autoencoder learns compressed representations by activating only a subset of neurons, thus focusing on the most informative features [29], [30]. The model included an input layer for the 200 genes, two hidden layers with 128 and 64 neurons, and an output layer mirroring the input. Rectified linear unit (ReLU) activation was used, with a sparsity constraint ( $\beta=0.05$ ) to suppress noise. Training was performed using the Adam optimizer (learning rate=0.001) for 100 epochs. After training, encoder weights were analyzed, and the top 10 genes with the highest contributions to latent features were selected, yielding a compact and interpretable set for classification.

## 3.4. Framework integration and biological relevance

The three-stage framework was designed to progressively refine the feature set while addressing key challenges in RNA-Seq data analysis. Stage 1 focuses on statistical significance and robustness, ensuring that the selected features are reproducible and biologically relevant. Stage 2 prioritizes predictive and complementary genes, reducing redundancy and focusing on those features that contribute the most to classification. Finally, stage 3 leverages deep learning to refine the feature set further, capturing non-linear patterns and relationships among genes. Together, these stages produce a compact, biologically meaningful feature set optimized for cancer classification.

## 3.5. Classification framework

The selected features were used to train a hybrid deep learning classifier, combining a dense feedforward neural network with an attention mechanism. A dense feedforward neural network is a type of artificial neural network where data flows sequentially through layers, making it well-suited for supervised learning tasks [31]. The architecture consisted of three hidden layers with 128, 64, and 32 neurons, respectively, each employing ReLU activation functions to introduce non-linear transformations, enabling the model to capture complex patterns in the data [32]. Dropout layers with a rate of 0.5 were included after each hidden layer to reduce overfitting by randomly deactivating a fraction of neurons during training.

To enhance the model's interpretability and focus on the most critical features, attention mechanism was incorporated. The attention mechanism assigns weights to features (genes), allowing the model to prioritize those most influential for classification. These weights also provide insights into the biological importance of individual genes, linking computational predictions to potential biological relevance [33].

The model was trained using the Adam optimizer [34], a gradient-based optimization algorithm known for its adaptive learning rate, with a learning rate of 0.0005 and a batch size of 32. The loss function used was categorical cross-entropy, a standard metric for multi-class classification tasks, which minimizes the difference between predicted and actual class probabilities. Early stopping, based on validation loss, was employed to prevent overfitting by halting training once performance improvements plateaued. This framework was evaluated using 5-fold cross-validation, ensuring robust estimates of model performance by iteratively training and testing the classifier on different subsets of the data.

## 3.6. Pathway-enriched biological validation

To validate the biological relevance of the selected genes, pathway enrichment analysis was performed using the Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) databases. KEGG provides curated information on molecular pathways, while GO annotates genes across three domains: biological processes, cellular components, and molecular functions. These tools offer complementary insights into the roles of genes within the broader biological context of lung cancer [35], [36].

The analysis was conducted using database for annotation, visualization, and integrated discovery (DAVID) and Enrichr, two widely used tools for functional annotation and enrichment analysis. These tools compare the selected gene set against reference gene sets to identify statistically overrepresented pathways. The enrichment analysis focused on pathways known to be involved in lung cancer progression, such as the epidermal growth factor receptor (EGFR) signaling pathway, cell cycle regulation, and immune response. These pathways play critical roles in tumor proliferation, therapy resistance, and the tumor microenvironment [37], [38]. This step ensures that the selected features are not only predictive but also biologically meaningful, bridging the gap between computational results and real-world biological implications.

### 3.7. Performance evaluation and cross-dataset validation

The classifier's performance was evaluated using key metrics: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Accuracy reflects overall prediction correctness; precision measures the proportion of true positives among predicted positives; recall (sensitivity) assesses the ability to detect true positives; and the F1-score balances precision and recall. The AUC-ROC captures the trade-off between sensitivity and specificity across thresholds, offering a comprehensive measure of model performance [39].

A 5-fold cross-validation strategy was applied to ensure reliable and unbiased evaluation. The dataset was split into five parts, with each subset used once as the test set while the others served for training, and results were averaged to reduce sampling bias. To assess generalizability, cross-dataset validation was conducted using the GSE19188 dataset from the gene expression omnibus (GEO) repository, which includes microarray-based expression profiles of lung adenocarcinoma and normal samples. Preprocessing steps such as normalization, log transformation, and alignment to TCGA protocols ensured compatibility. This validation confirmed the framework's robustness across different data platforms and experimental conditions.

## 4. RESULTS AND DISCUSSION

This section presents the findings of the multi-stage feature selection and classification framework applied to the TCGA-LUAD dataset and validated on the independent GSE19188 dataset. The results highlight the framework's ability to select biologically meaningful features and achieve robust classification performance across diverse datasets. This will be explained as follows.

### 4.1. Dataset preprocessing

The TCGA-LUAD dataset, comprising RNA-Seq profiles from 585 tumor and 59 normal lung tissue samples, underwent rigorous preprocessing. Gene expression values were  $\log_2(\text{FPKM}+1)$  transformed to stabilize variance, and batch effects were corrected using the ComBat algorithm to preserve biological signals. Genes with low expression (under the 10<sup>th</sup> percentile) were removed, resulting in a refined set of 14,000 genes. Outliers, identified via Mahalanobis distance, were excluded, yielding a final dataset of 570 tumor and 59 normal samples ensuring data quality for subsequent feature selection and classification.

### 4.2. Multi-stage feature selection results

#### 4.2.1. Stage 1: statistical relevance and stability analysis

Differential expression analysis using the Wilcoxon rank-sum test identified 4,200 genes as significantly differentially expressed ( $\text{FDR} < 0.05$ ). To enhance robustness, stability selection using bootstrap resampling was performed, retaining 3,800 genes consistently identified as significant across 100 iterations. Variance filtering further reduced the feature set to 3,000 genes, focusing on those with the highest variability and biological relevance.

#### 4.2.2. Stage 2: entropy-driven feature ranking

The 3,000 genes were ranked using JMIM, which evaluates the relationship between each gene and class labels while minimizing redundancy among features. The top 200 genes with the highest MI scores were selected for further refinement. This threshold was chosen to balance dimensionality reduction with the retention of sufficient biological diversity, ensuring robust downstream classification.

#### 4.2.3. Stage 3: sparse autoencoder for feature refinement

A sparse autoencoder refined the feature set by analyzing gene contributions to latent representations. The autoencoder, comprising two hidden layers (128 and 64 neurons), was trained for 100 epochs using the Adam optimizer (learning rate=0.001) with a sparsity constraint ( $\beta=0.05$ ). The encoder weights were analyzed, and the 10 genes with the highest contributions were retained as the final feature set. This compact feature set ensures interpretability and biological relevance while maintaining strong classification performance.

### 4.3. Classification performance

The final 10 selected genes were used to train a hybrid deep learning classifier combining a dense feedforward neural network with an attention mechanism. Evaluated using 5-fold cross-validation on the TCGA-LUAD dataset, the model achieved excellent performance: 99.3% accuracy, 99.0% precision, 99.5% recall, 99.2% F1-score, and an AUC-ROC of 0.998, demonstrating its ability to distinguish between tumor and normal samples with minimal misclassification. To evaluate generalizability, the model was tested on the independent GSE19188 dataset containing microarray profiles of 45 LUAD and 65 normal samples. Despite platform differences, the classifier retained high performance with 98.7% accuracy, 98.3% precision, 99.0% recall, 98.6% F1-score, and an AUC-ROC of 0.995.

### 4.4. Discussion of results

This study introduces a novel multi-stage feature selection framework combined with a hybrid deep learning classifier for LUAD classification using RNA-Seq data. By comparing our results to existing research, it is evident that the proposed framework addresses key limitations related to feature redundancy, biological interpretability, and classification robustness, demonstrating superior performance. For instance, Li *et al.* [40] achieved an AUC of 0.87, considerably lower than our model's AUC of 0.998 on TCGA-LUAD and 0.995 on GSE19188. Their reliance on differential expression analysis without robust feature selection contributes to suboptimal performance, unlike our multi-stage approach that integrates statistical filtering, entropy-based ranking, and sparse autoencoder refinement for compact and informative gene selection.

Similarly, Zheng *et al.* [41] reported AUC values ranging from 0.85 to 0.92 on TCGA-LUAD and from 0.83 to 0.90 on GSE19188. Although this study leveraged network-based methods for biomarker identification, it lacked advanced mechanisms to mitigate feature redundancy or improve biological interpretability. In contrast, our method incorporates stability selection and entropy-driven scoring, followed by attention-based deep learning to ensure both relevance and non-redundancy in the selected features.

Sherafatian and Arjmand [42] employed interpretable decision tree classifiers but reported an AUC of only 0.91 and an accuracy of 87.9%. While decision trees offer transparency, they are often too simplistic to capture the complex, non-linear patterns within high-dimensional transcriptomic data. Our deep learning classifier with an attention mechanism overcomes these limitations, achieving 99.3% accuracy and highlighting the most biologically informative features.

Rana *et al.* [43] achieved an AUC of 0.92 and approximately 94% accuracy by applying iterative feature selection. However, their reliance on linear redundancy reduction may fail to capture intricate gene interactions. By incorporating a sparse autoencoder in the final stage of our selection process, we successfully detect non-linear dependencies, further improving performance and robustness.

Lastly, Wei *et al.* [44] reported an AUC of 0.9958 using logistic regression applied to multi-omics data. Although impressive, their model introduces significant complexity due to the integration of heterogeneous data types. In contrast, our single-omics RNA-Seq-based method achieves similar performance with reduced data burden. Moreover, while that study employs Shapley additive explanations (SHAP) for interpretability, our approach provides direct biological validation via pathway enrichment analysis, offering more interpretable and clinically meaningful insights into LUAD-relevant mechanisms such as EGFR signaling, immune response, and cell cycle dysregulation.

In conclusion, by combining statistical filtering, entropy-based ranking, sparse autoencoding, and attention-driven classification, our framework addresses limitations seen across current LUAD studies. It delivers compact, predictive, and biologically interpretable gene subsets validated across datasets. These results demonstrate our method's potential for use in translational research and precision oncology, enabling reliable biomarker discovery and personalized diagnostics.

### 4.5. Biological validation of selected features

Pathway enrichment analysis confirmed the biological relevance of the 10 selected genes in lung adenocarcinoma, highlighting their involvement in key cancer-related pathways using KEGG and GO databases. Notably, the EGFR signaling pathway central to tumor proliferation, angiogenesis, and therapy resistance was significantly enriched, along with pathways related to cell cycle regulation and immune response, underscoring disruptions in cell division and tumor-immune interactions. The gene set included well-established biomarkers such as surfactant protein A2 (SFTPA2), a key player in lung function and diagnostics; napsin an aspartic peptidase (NAPSA), a protease used to distinguish lung adenocarcinoma from other cancers; and T-box transcription factor 4 (TBX4), involved in lung development and oncogenesis.

In addition to these known markers, two novel genes Mucin 16 (MUC16) and targeting protein for xenopus kinesin-like protein 2 (TPX2) were identified. MUC16 is linked to immune evasion and epithelial-mesenchymal transition (EMT), contributing to tumor progression and metastasis, while TPX2, associated with mitotic spindle formation, promotes uncontrolled cell proliferation. Their inclusion demonstrates the framework's ability to detect both established and previously unrecognized biomarkers, offering new insights

for diagnostic and therapeutic strategies in lung adenocarcinoma and supporting advancements in personalized and translational oncology.

4.6. Discussion and limitations

The findings validate the proposed multi-stage feature selection framework as a robust and interpretable tool for lung cancer classification. By incorporating statistical, entropy-driven, and deep learning methodologies, the framework effectively tackles key challenges inherent in high-dimensional RNA-Seq data, including issues such as noise, redundancy, and overfitting. The framework's strong performance on both the TCGA-LUAD and GSE19188 datasets underscores its potential applicability in real-world scenarios, particularly in biomarker discovery and cancer diagnostics.

Despite these promising results, several limitations should be noted. First, while the final subset of 10 genes is highly interpretable, it may overlook features associated with less common pathways or rare subtypes of lung adenocarcinoma. Second, platform variability between RNA-Seq and microarray datasets, though addressed in the analysis, could impact the generalizability of the findings across different technological platforms. Lastly, while pathway enrichment analysis supports the biological significance of the selected genes, additional experimental validation is needed to confirm their functional roles in lung adenocarcinoma progression.

Nonetheless, this study represents a significant advancement in feature selection methods for cancer classification. The framework not only enhances the interpretability of selected biomarkers but also provides a foundation for future research to explore the clinical potential of these biomarkers. By bridging computational predictions with biological insights, this approach paves the way for advancements in precision oncology and the development of personalized diagnostic and therapeutic strategies.

5. CONCLUSION

In this study, we developed a robust and interpretable multi-stage feature selection and classification framework tailored to lung adenocarcinoma diagnosis using RNA-Seq gene expression data. By integrating statistical significance, entropy-driven ranking, and sparse autoencoder refinement, the framework successfully identified a compact subset of 10 biologically relevant genes, validated through pathway enrichment and high classification performance across TCGA and GEO datasets. The hybrid deep learning model incorporating attention mechanisms not only achieved state-of-the-art accuracy but also preserved biological insight by highlighting critical biomarkers such as SFTPA2, NAPSA, TBX4, MUC16, and TPX2. Cross-platform validation confirmed the method's generalizability, while the inclusion of novel markers suggests promising avenues for translational cancer research. These findings underscore the framework's potential as a valuable tool in precision oncology, with future extensions potentially incorporating multi-omics integration and network-based feature selection to further enhance its clinical utility.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sara Haddou Bouazza	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Jihad Haddou Bouazza	✓	✓		✓	✓		✓			✓			✓	✓

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

This study did not involve individuals nor any personal identification information that could require any informed consent.

## ETHICAL APPROVAL

This paper does not involve people or animals; no investigation has involved human subjects. Therefore, the authors did not seek approval from any institutional review board.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




## REFERENCES

- [1] R. H. Elden, V. F. Ghonim, M. M. A. Hadhoud, and W. Al-Atabany, "Transcriptomic marker screening for evaluating the mortality rate of pediatric sepsis based on Henry gas solubility optimization," *Alexandria Engineering Journal*, vol. 68, pp. 693–707, Apr. 2023, doi: 10.1016/j.aej.2022.12.027.
- [2] M. Elloumi, M. A. Ahmad, A. H. Samak, A. M. A.-Sharafi, D. Kihara, and A. I. Taloba, "Error correction algorithms in non-null aspheric testing next generation sequencing data," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 9819–9829, Dec. 2022, doi: 10.1016/j.aej.2022.03.041.
- [3] A. Akgül, S. H. A. Khoshnaw, and H. M. Rasool, "Minimizing cell signalling pathway elements using lumping parameters," *Alexandria Engineering Journal*, vol. 59, no. 4, pp. 2161–2169, Aug. 2020, doi: 10.1016/j.aej.2020.01.041.
- [4] Y. Esmaili *et al.*, "Exploring the evolution of tissue engineering strategies over the past decade: from cell-based strategies to gene-activated matrix," *Alexandria Engineering Journal*, vol. 81, pp. 137–169, Oct. 2023, doi: 10.1016/j.aej.2023.08.080.
- [5] M. Shaheen, N. Naheed, and A. Ahsan, "Relevance-diversity algorithm for feature selection and modified Bayes for prediction," *Alexandria Engineering Journal*, vol. 66, pp. 329–342, Mar. 2023, doi: 10.1016/j.aej.2022.11.002.
- [6] L. Zhang, L. Li, M. Tang, Y. Huan, X. Zhang, and X. Zhe, "A new approach to diagnosing prostate cancer through magnetic resonance imaging," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 897–904, Feb. 2021, doi: 10.1016/j.aej.2020.10.018.
- [7] S. J. Susmi, H. K. Nehemiah, A. Kannan, and J. Christopher, "Relevant gene selection and classification of leukemia gene expression data," in *Emerging Research in Computing, Information, Communication and Applications*, Singapore: Springer Singapore, 2016, pp. 503–510, doi: 10.1007/978-981-10-0287-8\_47.
- [8] A. S. M. Shafi, M. M. I. Molla, J. J. Jui, and M. M. Rahman, "Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques," *SN Applied Sciences*, vol. 2, no. 7, Jul. 2020, doi: 10.1007/s42452-020-3051-2.
- [9] N. K, H. Rajaguru, and R. P, "Microarray prostate cancer classification using eminent genes," in *2021 Smart Technologies, Communication and Robotics*, IEEE, Oct. 2021, pp. 1–5, doi: 10.1109/STCR51658.2021.9588811.
- [10] E. Badr, S. Almotairi, M. A. Salam, and H. Ahmed, "New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis," *Alexandria Engineering Journal*, vol. 61, no. 3, pp. 2520–2534, Mar. 2022, doi: 10.1016/j.aej.2021.07.024.
- [11] S. H. Bouazza and J. H. Bouazza, "Revolutionizing cancer classification: the SNR-OGSCC method for improved gene selection and clustering," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 1, pp. 466–472, 2025, doi: 10.11591/ijai.v14.i1.pp466-472.
- [12] T. Althobaiti, S. Althobaiti, and M. M. Selim, "An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making," *Alexandria Engineering Journal*, vol. 94, pp. 311–324, May 2024, doi: 10.1016/j.aej.2024.03.044.
- [13] F. Noman *et al.*, "Multistep short-term wind speed prediction using nonlinear auto-regressive neural network with exogenous variable selection," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1221–1229, Feb. 2021, doi: 10.1016/j.aej.2020.10.045.
- [14] L. Zanella, P. Facco, F. Bezze, and E. Cimetta, "Feature selection and molecular classification of cancer phenotypes: a comparative study," *International Journal of Molecular Sciences*, vol. 23, no. 16, Aug. 2022, doi: 10.3390/ijms23169087.
- [15] S. H. Bouazza and J. H. Bouazza, "Cancer classification using pattern recognition and computer vision techniques," *ITM Web of Conferences*, vol. 69, Dec. 2024, doi: 10.1051/itmconf/20246902002.
- [16] 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, doi: 10.11591/eei.v14i2.9270.
- [17] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data," *Processes*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/pr11020562.
- [18] Y. Sun, V. Urquidí, and S. Goodison, "Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach," *Breast Cancer Research and Treatment*, vol. 119, no. 3, pp. 593–599, Feb. 2010, doi: 10.1007/s10549-009-0365-6.
- [19] X. Deng, M. Li, S. Deng, and L. Wang, "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification," *Medical & Biological Engineering & Computing*, vol. 60, no. 3, pp. 663–681, Mar. 2022, doi: 10.1007/s11517-021-02476-x.
- [20] O. Abdelwahab, N. Awad, M. Elserafy, and E. Badr, "A feature selection-based framework to identify biomarkers for cancer diagnosis: a focus on lung adenocarcinoma," *PLOS ONE*, vol. 17, no. 9, Sep. 2022, doi: 10.1371/journal.pone.0269126.
- [21] K. Myacheva, A. Walsh, M. Riester, G. Pelos, J. Carl, and S. Diederichs, "CRISPRi screening identifies CASP8AP2 as an essential viability factor in lung cancer controlling tumor cell death via the AP-1 pathway," *Cancer Letters*, vol. 552, Jan. 2023, doi: 10.1016/j.canlet.2022.215958.
- [22] J. P. Ross, S. v. Dijk, M. Phang, M. R. Skilton, P. L. Molloy, and Y. Oytam, "Batch-effect detection, correction and characterisation in Illumina HumanMethylation450 and MethylationEPIC BeadChip array data," *Clinical Epigenetics*, vol. 14, no. 1, Dec. 2022, doi: 10.1186/s13148-022-01277-9.
- [23] K. Dashdondov and M.-H. Kim, "Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction," *Neural Processing Letters*, vol. 55, no. 1, pp. 265–277, Feb. 2023, doi: 10.1007/s11063-021-10663-y.
- [24] Y. Li, X. Ge, F. Peng, W. Li, and J. J. Li, "Exaggerated false positives by popular differential expression methods when analyzing human population samples," *Genome Biology*, vol. 23, no. 1, Mar. 2022, doi: 10.1186/s13059-022-02648-4.
- [25] O. Menyhart, B. Weltz, and B. Györfy, "MultipleTesting.com: a tool for life science researchers for multiple hypothesis testing correction," *PLOS ONE*, vol. 16, no. 6, Jun. 2021, doi: 10.1371/journal.pone.0245824.




- [26] R. Hyde, L. O'Grady, and M. Green, "Stability selection for mixed effect models with large numbers of predictor variables: a simulation study," *Preventive Veterinary Medicine*, vol. 206, p. 105714, Sep. 2022, doi: 10.1016/j.prevetmed.2022.105714.
- [27] O. A. M. Salem, F. Liu, Y.-P. P. Chen, A. Hamed, and X. Chen, "Fuzzy joint mutual information feature selection based on ideal vector," *Expert Systems with Applications*, vol. 193, May 2022, doi: 10.1016/j.eswa.2021.116453.
- [28] Y. Tang and A. Hoffmann, "Quantifying information of intracellular signaling: progress with machine learning," *Reports on Progress in Physics*, vol. 85, no. 8, p. 086602, Aug. 2022, doi: 10.1088/1361-6633/ac7a4a.
- [29] S. Chen and W. Guo, "Auto-encoders in deep learning—a review with new perspectives," *Mathematics*, vol. 11, no. 8, Apr. 2023, doi: 10.3390/math11081777.
- [30] K. N. Rao, K. V. Rao, and P. R. P.V.G.D., "A hybrid intrusion detection system based on sparse autoencoder and deep neural network," *Computer Communications*, vol. 180, pp. 77–88, Dec. 2021, doi: 10.1016/j.comcom.2021.08.026.
- [31] M. Jamei, I. A. Olumegbon, M. Karbasi, I. Ahmadianfar, A. Asadi, and M. M.-Dehkordi, "On the thermal conductivity assessment of oil-based hybrid nanofluids using extended Kalman filter integrated with feed-forward neural network," *International Journal of Heat and Mass Transfer*, vol. 172, Jun. 2021, doi: 10.1016/j.ijheatmasstransfer.2021.121159.
- [32] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, "Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 22, Nov. 2021, doi: 10.1002/cpe.6143.
- [33] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, doi: 10.1016/j.neucom.2021.03.091.
- [34] R. O. Ogundokun, R. Maskeliunas, S. Misra, and R. Damaševičius, "Improved CNN based on batch normalization and Adam optimizer," in *Computational Science and Its Applications – ICCSA 2022 Workshops*, Springer, Cham, 2022, pp. 593–604, doi: 10.1007/978-3-031-10548-7\_43.
- [35] L. Yang, Y.-H. Zhang, F. Huang, Z. Li, T. Huang, and Y.-D. Cai, "Identification of protein–protein interaction associated functions based on gene ontology and KEGG pathway," *Frontiers in Genetics*, vol. 13, Sep. 2022, doi: 10.3389/fgene.2022.1011659.
- [36] S. A. Aleksander *et al.*, "The gene ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, May 2023, doi: 10.1093/genetics/iyad031.
- [37] B. T. Sherman *et al.*, "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)," *Nucleic Acids Research*, vol. 50, no. W1, pp. W216–W221, Jul. 2022, doi: 10.1093/nar/gkac194.
- [38] Z. Xie *et al.*, "Gene set knowledge discovery with enrichr," *Current Protocols*, vol. 1, no. 3, Mar. 2021, doi: 10.1002/cpz1.90.
- [39] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545–1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.
- [40] Z. Li, F. Qi, and F. Li, "Establishment of a gene signature to predict prognosis for patients with lung adenocarcinoma," *International Journal of Molecular Sciences*, vol. 21, no. 22, Nov. 2020, doi: 10.3390/ijms21228479.
- [41] Q. Zheng, S. Min, and Q. Zhou, "Identification of potential diagnostic and prognostic biomarkers for LUAD based on TCGA and GEO databases," *Bioscience Reports*, vol. 41, no. 6, Jun. 2021, doi: 10.1042/BSR20204370.
- [42] M. Sherafatian and F. Arjmand, "Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data," *Oncology Letters*, vol. 18, no. 2, pp. 2125–2131, Jun. 2019, doi: 10.3892/ol.2019.10462.
- [43] P. Rana, P. Thai, T. Dinh, and P. Ghosh, "Relevant and non-redundant feature selection for cancer classification and subtype detection," *Cancers*, vol. 13, no. 17, Aug. 2021, doi: 10.3390/cancers13174297.
- [44] J. Wei, X. Wang, H. Guo, L. Zhang, Y. Shi, and X. Wang, "Subclassification of lung adenocarcinoma through comprehensive multi-omics data to benefit survival outcomes," *Computational Biology and Chemistry*, vol. 112, Oct. 2024, doi: 10.1016/j.compbiolchem.2024.108150.

## BIOGRAPHIES OF AUTHORS



**Sara Haddou Bouazza**    holds a Doctorate in Electrical Engineering and Informatics, as well as a Master's in Electrical Engineering from Cadi Ayyad University, Marrakech. She also completed her Bachelor's in Physical Sciences. Currently, she is a Professor and Researcher at the LAMIGEP laboratory, EMSI Marrakech. Her research includes AI techniques for cancer classification, gene expression analysis, and security challenges in IoT environments. She has published numerous papers, including recent work on leukemia classification and AI in CNS tumors. She can be contacted at email: sara.hb.sara@gmail.com.



**Jihad Haddou Bouazza**    is an engineer specializing in software engineering and image processing from IGA Institut Supérieur du Génie Appliqué, Marrakech. Currently, he serves as a senior full stack developer and tech lead at Nexular Corp. He is certified in Python, machine learning, and as a certified network security specialist (CNSS). His research includes pattern recognition using artificial intelligence, with a publication presented at the GAST24 congress. He can be contacted at email: haddou.jihad@gmail.com.