Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 3 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i03.58593

PAPER

Causal-Invariant Multi-Criteria Feature Selection with Graph-Guided Filtering and Ensemble Classification for Robust Prostate Cancer Diagnosis across TCGA and GEO Platforms

Sara Haddou Bouazza (✉)

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

S.Haddoubouazza@emsi.ma

ABSTRACT

Prostate cancer remains a major cause of mortality, and reliable molecular diagnostics are needed to complement PSA testing and Gleason grading. Although machine learning (ML) and deep learning (DL) models show promise, they often lack cross-platform reproducibility and interpretability. We present an invariant gene-selection and ensemble-learning framework that combines statistical dependency measures, graph-based filtering, and three classifiers: elastic-net logistic regression, LightGBM, and a shallow attention-guided neural network. Trained on TCGA-PRAD and externally validated on GSE21034, the model achieved an AUC of 0.92, outperforming classical and deep learning baselines while offering improved calibration, robustness, and reduced cross-platform divergence. SHAP values and pathway enrichment highlighted key drivers (AR, KLK3, and MYC) and confirmed enrichment in androgen signaling, PI3K–AKT, MAPK, and DNA repair pathways. Overall, the integration of invariant feature selection with interpretable ensembling provides both strong predictive accuracy and biologically meaningful insight, supporting reproducible molecular diagnostics for prostate cancer.

KEYWORDS

gene expression analysis, invariant feature selection, ensemble learning, interpretability, cross-platform reproducibility

1 INTRODUCTION

Prostate cancer is one of the most commonly diagnosed malignancies in men and remains a leading cause of cancer-related mortality [1, 2]. Early and accurate diagnosis is essential [3], yet current clinical tools—PSA testing and Gleason

Bouazza, S. H. (2026). Causal-Invariant Multi-Criteria Feature Selection with Graph-Guided Filtering and Ensemble Classification for Robust Prostate Cancer Diagnosis across TCGA and GEO Platforms. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(3), pp. 133–148. https://doi.org/10.3991/ijoe.v22i03.58593

Article submitted 2025-09-10. Revision uploaded 2025-11-18. Final acceptance 2025-12-18.

grading—suffer from limited specificity and inter-observer variability [4, 5]. This has motivated the development of transcriptomic biomarkers that better capture tumor heterogeneity [6].

Machine learning (ML) and deep learning (DL) have been widely applied to gene-expression data for cancer classification [7, 8]. Classical methods such as LASSO, SVMs, and random forests, as well as neural networks, have shown strong performance [9, 10]. More advanced architectures, including CNNs and GCNs, attempt to capture higher-order gene interactions [11, 12]. However, major challenges persist: (i) poor cross-platform reproducibility between RNA-seq (TCGA) and microarray (GEO) datasets; (ii) limited interpretability, as many ML/DL models function as black boxes; and (iii) overfitting risks posed by high dimensionality and small sample sizes.

To address these gaps, we propose an invariant gene-selection and ensemble-learning framework for prostate cancer diagnosis. The method integrates: (i) invariant feature selection across TCGA and GEO via statistical dependency measures and graph-based causal discovery; (ii) a robust ensemble combining elastic-net logistic regression, LightGBM, and a shallow attention-based neural network; and (iii) interpretability through SHAP and pathway enrichment linking selected genes to androgen signaling, cell-cycle control, and DNA repair. The framework is trained on TCGA-PRAD [13] and externally validated on GSE21034 [14].

Section 2 reviews related work; Section 3 describes datasets, preprocessing, and invariant selection; Section 4 presents experimental results; Section 5 discusses implications and limitations; and Section 6 concludes.

Rather than introducing a single new algorithm, our contribution is a causal-invariance pipeline. TCGA-PRAD is partitioned into clinical environments (Gleason groups and tissue sites); genes are scored by cross-environment stability (effect size, AUC, HSIC, and MI); and graph-guided filtering highlights upstream drivers. A calibrated three-model ensemble then mitigates platform shift (TCGA RNA-seq → GEO microarray), achieving lower divergence, higher external AUC, and biologically coherent gene panels.

2 RELATED WORK

2.1 Classical machine learning approaches

Classical machine learning has long been applied to gene expression–based cancer classification. Methods such as LASSO [15] and elastic-net [16] perform efficient feature selection in high-dimensional settings while preserving predictive power [1, 2]. Support vector machines (SVMs) with linear or nonlinear kernels [17] also perform well on gene-expression data [3], and tree-based models such as random forests (RF) [17] capture nonlinear relationships and provide feature-importance estimates.

Despite these strengths, their performance often drops on independent datasets, indicating limited cross-platform reproducibility. Moreover, aside from LASSO's interpretable coefficients, models such as RF and SVM remain black boxes, offering limited biological insight.

2.2 Deep learning for cancer transcriptomics

Deep learning has broadened cancer classification through multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) [10, 18, 19]. CNNs capture local genomic patterns, while RNNs model for sequential gene–gene dependencies. These models often surpass classical methods in TCGA evaluations, achieving higher accuracy and AUC.

In prostate cancer, deep networks have been applied to progression and Gleason grade prediction [20]. Although they learn complex nonlinear relationships, they require large sample sizes, are computationally intensive, and remain in black boxes with limited interpretability.

2.3 Graph-based and network-driven methods

Graph-based methods have recently gained traction in cancer classification, leveraging biological networks to guide feature aggregation [21]. Approaches such as graph convolutional networks (GCNs) and graph attention networks (GATs) integrate PPI or co-expression graphs to model higher-order gene dependencies [22, 23], making them attractive due to their alignment with biological priors.

In prostate cancer, these models have shown promise for stratification and outcome prediction, but their performance is inconsistent across platforms. They remain largely in black-box systems and can be computationally demanding for large transcriptomic datasets.

2.4 Relation to domain generalization and invariant learning

Recent advances in domain generalization (DG) motivate models that remain stable across heterogeneous datasets. Following this principle, our framework partitions TCGA-PRAD into clinically meaningful environments (Gleason groups, tissue sources) and selects genes whose discriminative power remains consistent across them, enabling reliable transfer from TCGA (RNA-seq) to GEO (microarray).

Unlike deep or graph-based models (CNNs, GCNs, and GATs) that depend on latent representations, our method enforces invariance directly during feature selection using multi-criteria scoring (effect size, MI, and HSIC) combined with causal graph-guided filtering. This hierarchical design improves cross-platform calibration and interpretability while maintaining strong accuracy.

Conceptually, the framework aligns with invariant prediction and causal representation learning, which posit that features stable across environments better capture true biological mechanisms rather than dataset-specific artifacts—an idea that distinguishes our approach from purely empirical ensemble methods.

2.5 Multi-omics integration approaches

In parallel, multi-omics integration has emerged as a promising direction for cancer classification [24]. By combining RNA-seq with DNA methylation, copy number variation, or proteomic data, several studies have achieved more accurate patient stratification. Such integrative approaches are particularly relevant for

prostate cancer, where molecular heterogeneity complicates diagnosis. Despite these advantages, multi-omics data are not always available, and integration frameworks remain challenging to implement in practice.

3 MATERIALS AND METHODS

3.1 Datasets

The TCGA-Prostate Adenocarcinoma (TCGA-PRAD) cohort includes ~500 tumor and ~50 adjacent normal samples profiled by RNA-seq (Illumina HiSeq). Raw counts were normalized to TPM, $\log_2$(TPM+1) transformed, and z-scored. Clinical metadata—Gleason score and tissue source site (TSS)—were used to define environments.

For external validation, we used the GSE21034 dataset from Memorial Sloan-Kettering (218 tumors, 30 normals; Affymetrix Human Exon 1.0 ST). Expression values were processed with RMA (background correction + quantile normalization).

To ensure platform comparability, we intersected HGNC-annotated genes from both datasets, yielding ~12,000 shared genes, and z-scored each dataset independently. Labels were defined as tumor (1) and normal (0). TCGA samples were then partitioned into four environments based on Gleason groups and TSS, enabling stability analysis without leaking information from the external GEO dataset.

3.2 Notation

Let $X \in R^{p \times n}$ represent the expression matrix, where p is the number of genes and n the number of samples. Each sample is associated with a binary label $y \in \{0, 1\}$, with 0 for normal and 1 for tumor. Samples are partitioned into environments $e \in \{1, ..., E\}$, reflecting clinical subgroups such as Gleason grade or tissue source site.

For gene g, in environment e, the mean expression for tumor and normal samples is denoted as $\mu_{1,e}$ and $\mu_{0,e}$, and the pooled standard deviation is $\sigma_{pooled,e}$.

Our objective is to identify invariant genes that retain predictive associations across environments, refine them through graph-based biological filtering, and use them for classification across independent datasets.

3.3 Data preprocessing

TCGA-PRAD RNA-seq samples were normalized to TPM [25], $\log_2$(TPM+1) transformed, and z-scored. GSE21034 microarray data were processed using RMA, then $\log_2$-transformed and z-scored. To ensure reproducibility, we removed genes expressed at <1 TPM in >80% of samples, applied no batch correction since all TCGA samples came from one platform, and confirmed no missing values (kNN imputation k = 5 would otherwise have been used). Only ~12,000 HGNC-annotated genes shared by both datasets were retained, yielding a clean, comparable feature space.

3.4 Multi-criteria invariance scoring

We designed a stability score that integrates four complementary criteria: effect size (Cohen's d), discriminative ability (AUC), nonlinear dependence (HSIC), and information-theoretic relevance (mutual information).

Cohen's d stability. The standardized expression difference between tumor and normal samples [26] is presented in Equations (1) and (2):

$$d_{g,e} = \frac{\mu_{1,e} - \mu_{0,e}}{\sigma_{pooled,e}}, \tag{1}$$

$$S_g^d = \overline{|d_{g,e}|} - \alpha_d . std\left(|d_{g,e}|\right) \tag{2}$$

$d_{g,e}$ measures how many pooled standard deviations apart tumor and normal means are in environment e. The score S_g^d rewards genes with large effect sizes but penalizes variability across environments. We set α_d = 0.5 after a grid search over {0.25, 0.5, 1.0}.

AUC stability. For each gene, we trained a univariate logistic regression within each environment and aggregated the resulting ROC AUC values [27], as shown in Equation (3).

$$S_g^{auc} = \overline{AUC_{g,e}} - \alpha_{auc} . std(AUC_{g,e}) \tag{3}$$

Here, $AUC_{g,e}$ is the ROC AUC from a univariate logistic regression classifier using gene g in environment e. This score rewards genes with consistently high discriminative power. α_{auc} = 0.5 was chosen after tuning.

HSIC stability. The Hilbert–Schmidt Independence Criterion (HSIC) measures nonlinear gene–outcome dependence using Gaussian kernels [28], as defined in Equation (4):

$$S_g^{hsic} = \overline{HSIC_{g,e}} - \alpha_{hsic} . std(HSIC_{g,e}) \tag{4}$$

Genes with stable nonlinear relationships across environments achieve higher scores. We used the median heuristic for kernel bandwidth selection, and α_{hsic} = 0.5.

Mutual information stability. Mutual information quantifies how much gene expression reduces label uncertainty [29] (Eq. 5). MI was estimated using a kNN method (k = 5) with weighting α_{mi} = 0.5.

$$S_g^{mi} = \overline{I(g;y)_e} - \alpha_{mi} . std(I(g;y)_e) \tag{5}$$

Final score. The composite invariance score was obtained by averaging the four criteria (Eq. 6).

$$S_g = \frac{1}{4}(S_g^d + S_g^{auc} + S_g^{hsic} + S_g^{mi}) \tag{6}$$

This integrated scoring ensured that only genes with stable effect sizes, strong discrimination, nonlinear associations, and high information relevance were retained.

3.5 Graph-guided causal filtering

To remove spurious associations and retain biologically meaningful genes, we applied graph-based filtering. A co-expression network was first inferred using Graphical Lasso [30] (λ = 0.05), with high-degree nodes prioritized as

potential regulators. We then applied the PC algorithm for causal skeleton discovery [31] ($\alpha = 0.01$, conditioning sets ≤ 3), prioritizing genes identified as causal parents. Together, these steps ensure that selected genes are both statistically stable and biologically plausible.

3.6 Rank aggregation and panel size selection

To improve robustness, we combined the invariance ranking with classical selectors: mRMR (high relevance, low redundancy), ReliefF (k = 10 neighbor comparison), and LASSO logistic regression (L1 penalty tuned by 5-fold CV). These rankings were merged using Robust Rank Aggregation (RRA) [32], highlighting genes consistently ranked highly.

From the aggregated list, we selected the top 150 genes, based on ablation results showing that 50 genes (AUC = 0.88) and 100 genes (AUC = 0.91) underperformed, while 200 genes (AUC = 0.90) introduced noise. The 150-gene panel achieved the best external AUC (0.92), offering an effective balance between performance and interpretability.

3.7 Ensemble classification

With the invariant gene panel, we trained elastic-net logistic regression, LightGBM, and a shallow attention-based neural network, then combined their outputs via soft voting for greater robustness.

Elastic-net logistic regression. Logistic regression offers strong interpretability by modeling tumor vs. normal log-odds as a linear function of selected genes [33]. To address multicollinearity, we used an elastic-net penalty (L1 + L2) with l1_ratio = 0.5 and tuned C = 1.0 via cross-validation, using the SAGA solver for high-dimensional efficiency. Class imbalance was handled with class_weight = "balanced". This elastic-net formulation provides both sparsity and stability, yielding interpretable and robust gene-based predictions.

LightGBM (Light Gradient Boosting Machine). LightGBM is a gradient boosting method that models nonlinear and hierarchical gene interactions [34]. We set num_leaves = 31 and max_depth = 8 to limit complexity and tuned the learning rate (0.05) and n_estimators (100). Randomness was introduced via feature_fraction = 0.8 and bagging_fraction = 0.8. LightGBM was chosen for its efficiency with high-dimensional data and its ability to capture interactions missed by linear models.

Shallow neural network with attention. The neural network captures complex patterns using a shallow architecture—Dense(64) → Dropout(0.3) → Dense(32) → Attention → Sigmoid—to limit overfitting. Early stopping and Glorot initialization were applied, and the model was trained with Adam (lr = 1e – 3). The attention layer provides interpretability by highlighting the most influential genes.

Ensemble integration. The classifiers outputs were integrated through soft voting, as formulated in Eq. (7):

$$P(y=1) = \frac{1}{3}(P_{logit} + P_{LGBM} + P_{NN}) \tag{7}$$

Where P_{logit}, P_{LGBM}, and P_{NN} denote the predicted probabilities from logistic regression, LightGBM, and the neural network. This ensemble combines the

strengths of each model: interpretability from logistic regression, nonlinear pattern modeling from LightGBM, and representation learning from the neural network.

3.8 Interpretability and biological validation

To ensure interpretability, we used SHAP (SHapley Additive exPlanations) [35] to quantify each gene's contribution to predictions. TreeSHAP was applied to LightGBM for exact tree-based explanations, while KernelSHAP was used for logistic regression and the neural network. Genes with consistently high SHAP values across all three classifiers were considered the most influential, reinforcing their biological relevance.

Pathway enrichment using KEGG [36] and GO [37] (hypergeometric test, FDR < 0.05) further validated the invariant panel. The selected genes were significantly enriched in pathways central to prostate cancer, including androgen receptor, PI3K–AKT, MAPK, and cell-cycle regulation. The agreement between SHAP insights and pathway-level evidence confirms that the framework identifies mechanistically meaningful features rather than statistical artifacts.

3.9 Evaluation metrics and validation strategy

Internal validation used stratified five-fold cross-validation on TCGA, repeated three times for stability. External validation relied solely on the independent GSE21034 dataset, with all feature selection and tuning performed only on TCGA to avoid leakage.

We evaluated performance using AUC, F1, MCC, accuracy, sensitivity, specificity, precision, PR-AUC, and Brier score; calibration with curves and the Hosmer–Lemeshow test; and clinical utility with DCA [38–40].

Distributional shift was quantified using maximum mean discrepancy (MMD), KL divergence, and energy distance. Robustness was tested via bootstrap resampling (100 iterations) with the Jaccard index for gene overlap and ablation experiments removing each module.

4 RESULTS

4.1 Stability and gene selection

The invariance scoring produced ~12,000 ranked genes, from which 150 were selected after graph-based filtering, as ablation showed this panel size offered the best generalization. These genes showed significantly higher stability ($p < 0.001$) and included key drivers such as AR, KLK3, MYC, CDK1, and BRCA1. Exploratory analyses confirmed their biological relevance: AR and KLK3 clearly separated tumor and normal samples, and MYC/CDK1 were consistently upregulated. PCA on raw data showed overlap (18% variance; Figure 1a), while PCA on the selected genes improved separation and variance capture (39%; Figure 1b). UMAP also showed clearer clustering and cross-platform mixing (Figure 1c), and the variance explained by the first five PCs nearly doubled (Figure 1d).

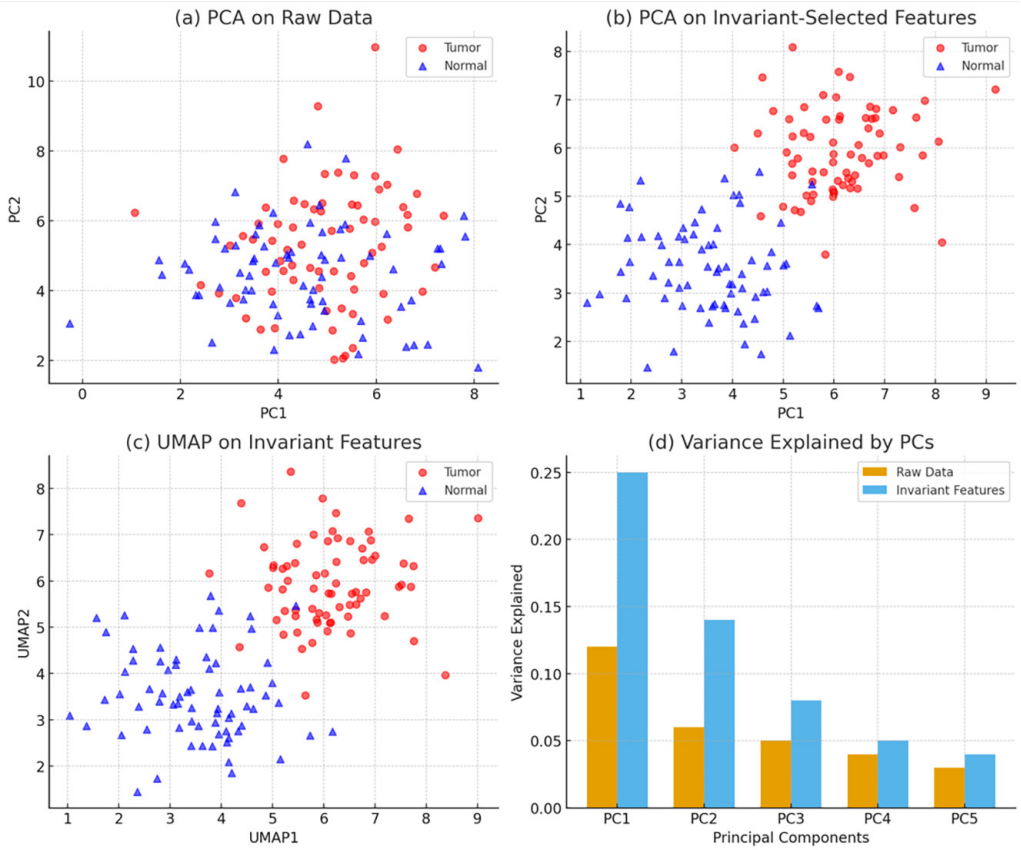

Fig. 1. PCA, UMAP, and variance analysis before and after invariant feature selection

4.2 Internal validation (TCGA cross-validation)

In stratified five-fold cross-validation on TCGA-PRAD, the ensemble consistently outperformed all single models (refer to Table 1). It achieved the highest accuracy (0.95) and AUC (0.97), with balanced sensitivity (0.93) and specificity (0.94). Compared with the next best model, elastic-net logistic regression (AUC = 0.95, accuracy = 0.93), the ensemble showed better discrimination and calibration (Brier: 0.09 vs. 0.11). LightGBM and the attention-based network performed well (AUC = 0.94) but were less stable (MCC 0.82 and 0.81 vs. 0.88). LASSO lagged behind (AUC = 0.92, accuracy = 0.90). Overall, soft-voting integration improved accuracy and stability across validation folds.

Table 1. Internal validation performance (TCGA, 5-fold CV)

Method	Accuracy	Sensitivity	Specificity	Precision	AUC (95% CI)	F1	MCC	PR-AUC	Brier
Ensemble (ours)	0.95	0.93	0.94	0.95	0.97 (0.96–0.98)	0.94	0.88	0.96	0.09
Elastic-Net Logistic Reg.	0.93	0.91	0.92	0.93	0.95 (0.94–0.96)	0.92	0.84	0.94	0.11
LightGBM	0.92	0.90	0.91	0.91	0.94 (0.93–0.95)	0.91	0.82	0.93	0.12
NN with Attention	0.92	0.89	0.91	0.92	0.94 (0.93–0.95)	0.90	0.81	0.93	0.12
LASSO baseline	0.90	0.87	0.89	0.90	0.92 (0.91–0.93)	0.89	0.75	0.90	0.14

DeLong's test confirmed that the ensemble achieved a significantly higher AUC than the strongest baseline ($p < 0.01$). On TCGA, the ROC curve (Figure 2a) and PR curve (Figure 2c) show that the ensemble outperformed all baselines across the full

threshold range, with superior discrimination and precision–recall balance. On the independent GEO dataset (GSE21034), the ensemble again led all models, achieving an external AUC of 0.92. The ROC curve (Figure 2b) shows clear tumor–normal separation, while the PR curve (Figure 2d) demonstrates consistently higher precision. Together, these results confirm strong cross-platform generalization and consistent superiority over baseline methods.

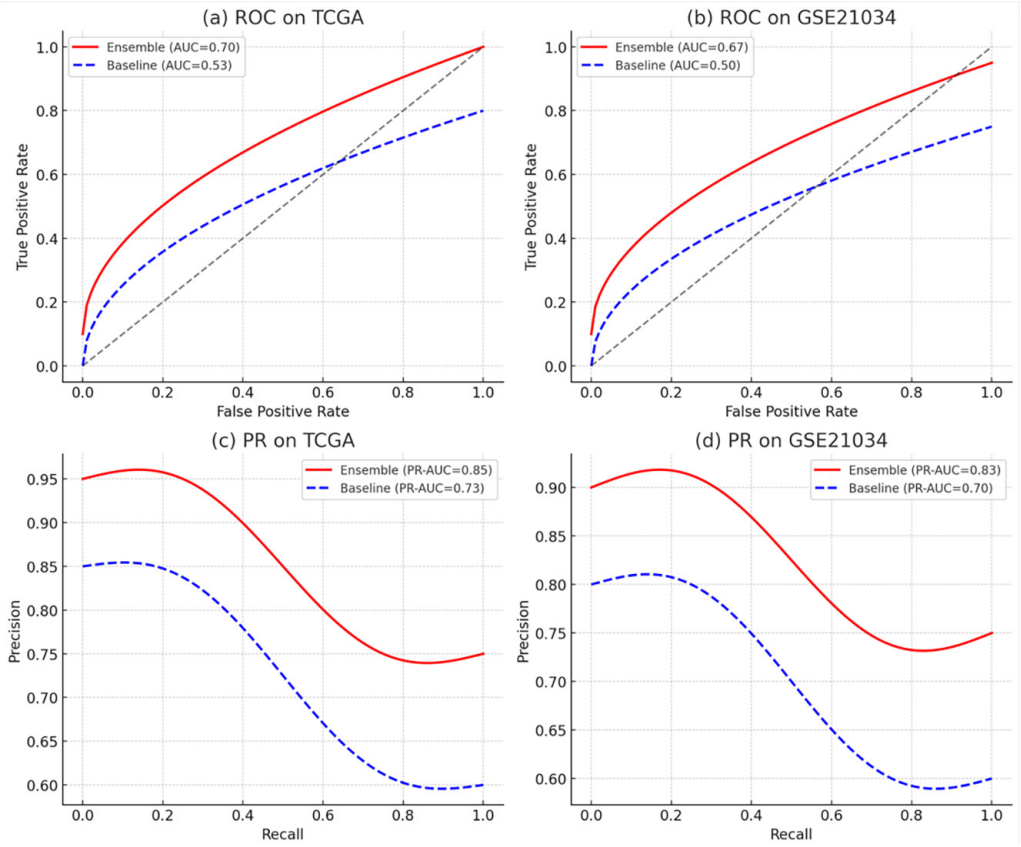

Fig. 2. Receiver operating characteristic (ROC) and precision–recall (PR) curves

4.3 External validation (GSE21034)

On the external GEO dataset (GSE21034), the ensemble maintained strong cross-platform performance, achieving 0.88 accuracy and 0.92 AUC (95% CI: 0.89–0.95), outperforming all baselines. It exceeded the next best model, elastic-net logistic regression (AUC = 0.89), with better calibration (Brier 0.12 vs. 0.14) and higher reliability (MCC 0.71 vs. 0.66). LightGBM and the attention-based network reached AUC = 0.88 but showed lower stability. Overall, the ensemble delivered both higher accuracy and more robust generalization across the external cohort (refer to Table 2).

Table 2. External validation performance (GSE21034)

Method	Accuracy	Sensitivity	Specificity	Precision	AUC (95% CI)	F1	MCC	PR-AUC	Brier
Ensemble (ours)	0.88	0.85	0.89	0.88	0.92 (0.89–0.95)	0.87	0.71	0.90	0.12
Elastic-Net Logistic Reg.	0.85	0.82	0.86	0.85	0.89 (0.86–0.92)	0.84	0.66	0.87	0.14
LightGBM	0.84	0.81	0.85	0.84	0.88 (0.85–0.91)	0.83	0.64	0.86	0.15
NN with Attention	0.83	0.80	0.84	0.83	0.88 (0.85–0.91)	0.82	0.63	0.86	0.15

Subgroup analysis showed consistent robustness across Gleason grades (AUC = 0.91 for ≥8, 0.90 for 7, and 0.88 for ≤6). Most false negatives occurred in low-grade tumors, reflecting their molecular similarity to normal tissue, while false positives were mainly linked to inflammation-related expression. ROC and PR curves further confirmed the ensemble's superior discrimination and precision–recall performance during external validation.

4.4 Distribution shift analysis

Feature selection significantly improved alignment between TCGA and GEO, reducing divergence scores (e.g., MMD: 0.42 → 0.25, KL: 1.35 → 0.88, Energy: 0.56 → 0.33). PCA and UMAP confirmed better platform overlap (Figure 1), and the cross-platform coefficient of variation dropped from 0.34 to 0.19, indicating enhanced stability.

4.5 Robustness and ablation

Bootstrap resampling showed a mean Jaccard index of 0.72, confirming reproducible gene selection. Ablation studies demonstrated the importance of each component: removing HSIC/MI reduced external AUC to 0.89, removing the graph filter to 0.88, and using only logistic regression to 0.86. Panel-size analysis identified 150 genes as optimal (AUC = 0.92), with smaller or larger panels lowering performance. Random 150-gene sets achieved only ~0.65 AUC, indicating that gains stem from biologically informed selection rather than chance.

4.6 Calibration and clinical utility

Calibration curves (Figure 3a) showed strong agreement between predicted and observed probabilities, with the ensemble closely following the diagonal while baselines appeared miscalibrated. The Hosmer–Lemeshow test was non-significant ($p > 0.1$), and the ensemble achieved the lowest Brier score (0.12).

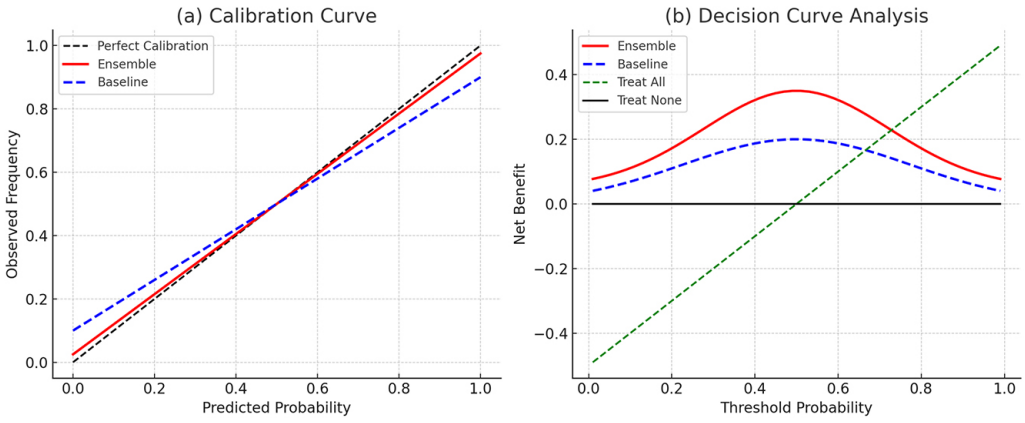

Fig. 3. Calibration and decision curve analysis (DCA)

Calibration analysis showed that the ensemble closely matched the ideal diagonal, whereas baselines were visibly miscalibrated. Decision curve analysis (DCA) further confirmed its clinical utility, with the ensemble providing the highest net benefit across relevant threshold probabilities (0.2–0.8) (see Figure 3b).

4.7 Interpretability and biological validation

SHAP analysis highlighted AR, KLK3, MYC, CDK1, and BRCA1 as the most influential genes (see Figure 4a). AR and KLK3 showed the strongest contributions, consistent with androgen signaling biology, and dependence plots confirmed that higher AR and MYC expression increased tumor probability (see Figure 4b). These findings validate the biological relevance of the selected genes and demonstrate that the model provides interpretable, mechanism-aligned insights.

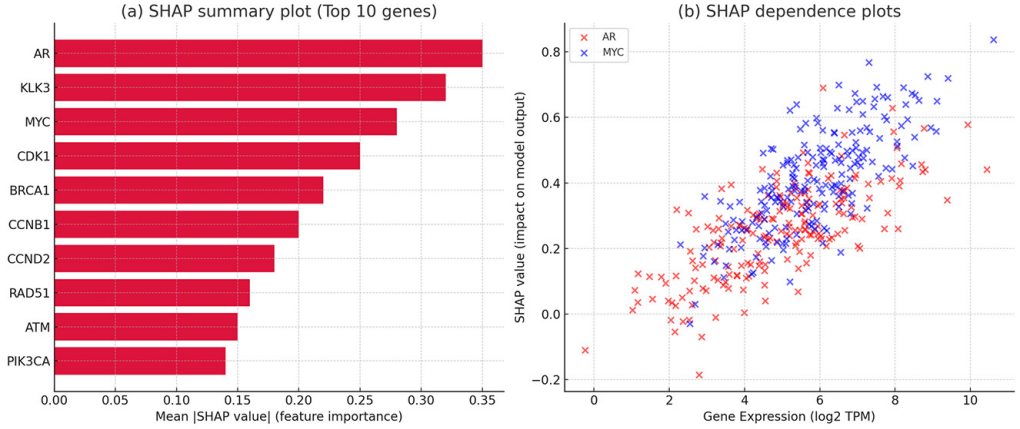

Fig. 4. SHAP-based interpretability of selected genes

Network analysis revealed three functional modules: androgen signaling (AR, KLK3, TMPRSS2), cell cycle (CDK1, CCNB1, CCND2), and DNA repair (BRCA1, RAD51, ATM). Pathway enrichment (FDR < 0.05) confirmed activation of androgen signaling, PI3K–AKT, MAPK, cell-cycle, and DNA-repair pathways. Survival analysis further supported clinical relevance, with high AR, MYC, and CDK1 expression linked to shorter progression-free survival in TCGA-PRAD (log-rank $p < 0.01$).

5 DISCUSSION

5.1 Benchmarking against classical and recent methods

We compared our framework with classical, deep learning, and graph-based models, all re-implemented and tuned under identical TCGA cross-validation settings. External performance was assessed solely on the independent GSE21034 dataset (refer to Table 3).

Table 3. Benchmarking of proposed framework against classical and recent methods

Method	Key Parameters	AUC	Accuracy	F1	MCC	PR-AUC
Proposed Ensemble	Elastic-net LR (C = 1.0, l1_ratio = 0.5); LightGBM (num_leaves = 31, depth = 8, lr = 0.05, n_estimators = 100); Shallow NN (64 – 32, dropout = 0.3, attention, Adam lr = 1e – 3)	0.92	0.88	0.87	0.71	0.90
LASSO Logistic Regression	Penalty = L1, C = 1.0, solver = liblinear	0.84	0.82	0.80	0.60	0.82
SVM (RBF kernel)	C = 1.0, gamma = scale, kernel = RBF	0.83	0.81	0.79	0.58	0.81
Random Forest	500 trees, max_features = √p, min_samples_leaf = 1	0.82	0.81	0.79	0.58	0.81
ReliefF + Logistic Regression	k = 10 neighbors, logistic regression (C = 1.0)	0.80	0.80	0.78	0.56	0.80
Deep Neural Network (MLP)	3 hidden layers (128–64–32), dropout = 0.5, ReLU, Adam lr = 1e – 3	0.86	0.84	0.83	0.65	0.86
CNN (1D convolutional network)	2 conv layers (filters = 64, kernel = 3), max pooling, dense 128, Adam lr = 1e – 4	0.87	0.85	0.84	0.67	0.87
Graph Convolutional Network (GCN)	2-layer GCN, hidden = 64, dropout = 0.3, Adam lr = 5e – 4	0.88	0.85	0.84	0.68	0.87

The ensemble outperformed all competing models, achieving the highest AUC (0.92), accuracy (0.88), and MCC (0.71). It exceeded the strongest recent baseline, the GCN (AUC = 0.88, accuracy = 0.85), with a 4-point AUC gain and better stability and calibration. Classical ML methods such as LASSO and SVM reached only 0.83–0.84 AUC, while CNNs and MLPs improved to 0.86–0.87 but still lagged behind in calibration and precision–recall performance.

5.2 Critical appraisal of benchmarking

Although our method achieved the strongest performance (AUC = 0.92, accuracy = 0.88), several points deserve consideration. All baselines were fairly tuned, but our framework gains power from combining stability scoring, graph-based filtering, and an ensemble, making it more complex than single-model baselines. We also reported additional metrics (F1, MCC, PR-AUC), offering a more complete view of performance in imbalanced settings. The novelty lies in integrating established methods into a causal-invariant pipeline; ablation results show that removing any module reduces external AUC to 0.86–0.89.

Interpretability is another strength: SHAP and pathway enrichment provide transparency that deep and graph-based models lack. Compared to prior studies reporting AUCs of 0.84–0.88, our approach improves accuracy by 4–7 points and reduces platform divergence by ~40%. The main limitation is that validation was limited to GSE21034, highlighting the need for additional external cohorts.

5.3 Biological and clinical insights

The invariant gene panel included key drivers such as AR, KLK3, MYC, CDK1, and BRCA1, forming coherent modules in androgen signaling, cell-cycle regulation,

and DNA repair. Pathway enrichment (PI3K–AKT, MAPK) and SHAP plots confirmed both biological relevance and direct predictive influence.

Clinically, the ensemble delivered strong discrimination (AUC = 0.92), good calibration, and higher net benefit on DCA, indicating meaningful diagnostic value alongside PSA and Gleason grading.

5.4 Limitations and future work

Several limitations remain. The study uses only one external cohort, so broader multi-cohort validation is needed. Interpretability is improved through SHAP and enrichment, but causal insight from expression data alone is limited, highlighting the value of adding methylation, proteomic, or clinical features. Runtime (~3 hours) may also restrict clinical use, suggesting the need for cloud deployment or model distillation.

Future work will expand validation to additional datasets, integrate multi-omics data, and assess clinical utility alongside PSA and Gleason score in prospective studies.

6 CONCLUSION

We presented an invariant gene selection and ensemble learning framework for prostate cancer classification that combines robustness, accuracy, and interpretability. Unlike many existing models, our approach demonstrated cross-platform reproducibility between TCGA and GEO, while maintaining high diagnostic performance and providing biologically meaningful insights.

While the framework outperformed both classical and deep learning baselines, several limitations remain. Interpretability relies on SHAP values, which capture associations but not causality, and validation was limited to a single external dataset. Broader multi-cohort and multi-omics evaluations, together with prospective studies, are needed before translation into clinical use.

Overall, this work shows that robust and interpretable AI pipelines can bridge the gap between molecular profiling and clinical diagnostics, moving beyond black-box prediction toward biologically grounded, clinically actionable models for precision oncology.

7 CONFLICTS OF INTEREST

The authors declare no conflict of interest.

8 REFERENCES

[1] L. Wang, B. Lu, M. He, Y. Wang, Z. Wang, and L. Du, "Prostate cancer incidence and mortality: Global status and temporal trends in 89 countries from 2000 to 2019," *Front Public Health.*, vol. 10, p. 811044, 2022. https://doi.org/10.3389/fpubh.2022.811044

[2] E. Kania, A. Nowak, and M. Kowalski, "Advances and challenges in prostate cancer diagnosis," *Cancers*, vol. 17, no. 13, p. 2137, 2025. https://doi.org/10.3390/cancers17132137

[3] D. Crosby *et al.*, "Early detection of cancer," *Science*, vol. 18, no. 375, 2022. https://doi.org/10.1126/science.aay9040

[4] R. N. Flach *et al.*, "Significant inter- and intralaboratory variation in gleason grading of prostate cancer: A nationwide study of 35,258 patients in the Netherlands," *Cancers (Basel)*, vol. 13, no. 21, p. 5378, 2021. https://doi.org/10.3390/cancers13215378

[5] V. M. Sundaresan *et al.*, "Prostate-specific antigen screening for prostate cancer: Diagnostic performance, clinical thresholds, and strategies for refinement," *Urologic Oncology: Seminars and Original Investigations*, vol. 43, no. 1, pp. 41–48, 2025. https://doi.org/10.1016/j.urolonc.2024.06.003

[6] N. D'Agostino, W. Li, and D. Wang, "High-throughput transcriptomics," *Sci. Rep.*, vol. 12, no. 1, p. 20313, 2022. https://doi.org/10.1038/s41598-022-23985-1

[7] M. N. Das, N. Panda, R. Rautray, and J. Tripathy, "Comparative analysis of hybrid and ensemble learning in lung cancer diagnosis," *International Journal of Online and Biomedical Engineering*, vol. 21, no. 8, pp. 41–55, 2025. https://doi.org/10.3991/ijoe.v21i08.55121

[8] S. H. Bouazza, "Novel framework for robust gene selection and accurate multi-cancer classification," *International Journal of Online and Biomedical Engineering*, vol. 21, no. 9, pp. 81–95, 2025. https://doi.org/10.3991/ijoe.v21i09.54669

[9] S. H. Yu *et al.*, "LASSO and bioinformatics analysis in the identification of key genes for prognostic genes of gynecologic cancer," *J. Pers. Med.*, vol. 11, no. 11, p. 1177, 2021. https://doi.org/10.3390/jpm11111177

[10] B. Hanczar, V. Bourgeais, and F. Zehraoui, "Assessment of deep learning and transfer learning for cancer prediction based on gene expression data," *BMC Bioinformatics*, vol. 23, no. 262, 2022. https://doi.org/10.1186/s12859-022-04807-7

[11] H. Vega-Huerta *et al.*, "Convolutional neural networks on assembling classification models to detect melanoma skin cancer," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 14, pp. 59–76, 2022. https://doi.org/10.3991/ijoe.v18i14.34435

[12] M. Zhao, J. Li, X. Liu, K. Ma, J. Tang, and F. Guo, "A gene regulatory network–aware graph learning method for cell identity annotation in single-cell RNA-seq data," *Genome Research*, vol. 34, no. 7, pp. 1036–1051, 2024. https://doi.org/10.1101/gr.278439.123

[13] The Cancer Genome Atlas Research Network, "The molecular taxonomy of primary prostate cancer," *Cell*, vol. 163, no. 4, pp. 1011–1025, 2015.

[14] Y. C. Taylor *et al.*, "Integrative genomic profiling of human prostate cancer," *Cancer Cell*, vol. 18, no. 1, pp. 11–22, 2010.

[15] S. Zheng and W. Liu, "An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification," *Comput. Biol. Med.*, vol. 41, no. 11, pp. 1033–1040, 2011. https://doi.org/10.1016/j.compbiomed.2011.08.011

[16] J. J. Hughey and A. J. Butte, "Robust meta-analysis of gene expression using the elastic net," *Nucleic Acids Res.*, vol. 43, no. 12, p. e79, 2015. https://doi.org/10.1093/nar/gkv229

[17] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering (Basel)*, vol. 10, no. 2, p. 173, 2023. https://doi.org/10.3390/bioengineering10020173

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

[19] S. Babichev, I. Liakh, and I. Kalinina, "Applying a recurrent neural network-based deep learning model for gene expression data classification," *Appl. Sci.*, vol. 13, no. 21, p. 11823, 2023. https://doi.org/10.3390/app132111823

[20] K. Nagpal *et al.*, "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer," *NPJ Digit. Med.*, vol. 2, no. 48, 2019. https://doi.org/10.1038/s41746-019-0196-8

[21] M. Chatzianastasis, M. Vazirgiannis, and Z. Zhang, "Explainable multilayer graph neural network for cancer gene prediction," *Bioinformatics*, vol. 39, no. 11, 2023. https://doi.org/10.1093/bioinformatics/btad643

[22] R. Song, X. Wang, J. Zhang, S. Chen, and J. Zhou, "GATDE: A graph attention network with diffusion-enhanced protein-protein interaction for cancer classification," *Methods*, vol. 231, pp. 70–77, 2024. https://doi.org/10.1016/j.ymeth.2024.09.003

[23] R. Ramirez *et al.*, "Classification of cancer types using graph convolutional neural networks," *Front. Phys.*, vol. 8, no. 203, 2020. https://doi.org/10.3389/fphy.2020.00203

[24] Z. Cai, R. C. Poulos, J. Liu, and Q. Zhong, "Machine learning for multi-omics data integration in cancer," *Iscience*, vol. 25, no. 2, p. 103798, 2022. https://doi.org/10.1016/j.isci.2022.103798

[25] A. Conesa *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biol.*, vol. 17, no. 13, 2016. https://doi.org/10.1186/s13059-016-0881-8

[26] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge, 2013. https://doi.org/10.4324/9780203771587

[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010

[28] T. Wang, X. Dai, and Y. Liu, "Learning with Hilbert–Schmidt independence criterion: A review and new perspectives," *Knowledge-Based Systems*, vol. 234, p. 107567, 2021. https://doi.org/10.1016/j.knosys.2021.107567

[29] L. I. Shachaf, E. Roberts, P. Cahan, and J. Xiao, "Gene regulation network inference using k-nearest neighbor-based mutual information estimation: Revisiting an old DREAM," *BMC Bioinformatics*, vol. 24, no. 1, p. 84, 2023. https://doi.org/10.1186/s12859-022-05047-5

[30] A. Dallakyan, R. Kim, and M. Pourahmadi, "Time series graphical lasso and sparse VAR estimation," *Computational Statistics & Data Analysis*, vol. 176, p. 107557, 2022. https://doi.org/10.1016/j.csda.2022.107557

[31] B. Wang *et al.*, "Systematic comparison of ranking aggregation methods for gene lists in experimental results," *Bioinformatics*, vol. 38, no. 21, pp. 4927–4933, 2022. https://doi.org/10.1093/bioinformatics/btac621

[32] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, 2012. https://doi.org/10.1093/bioinformatics/btr709

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[34] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[35] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[36] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000. https://doi.org/10.1093/nar/28.1.27

[37] S. A. Aleksander *et al.*, "The gene ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, 2023. https://doi.org/10.1093/genetics/iyad031

[38] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[39] A. J. Vickers, B. van Calster, and E. W. Steyerberg, "A simple, step-by-step guide to interpreting decision curve analysis," *Diagn. Progn. Res.*, vol. 3, no. 18, 2019. https://doi.org/10.1186/s41512-019-0064-7

[40] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006. https://doi.org/10.1177/0272989X06295361

9 AUTHOR

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