JJEE

Volume 12 | Number 2 | June 2026 | Pages 248-260

Jordan Journal of Electrical Engineering

ISSN (print): 2409-9600, ISSN (online): 2409-9619

Homepage: jjee.ttu.edu.jo

Pathway Guided Identification of Prognostic Biomarkers in Lung Adenocarcinoma

Nrupal Sankpal[1*], Dr. Sandeep S Musale[2], Dr. Supriya Mangale[3]

[1, 2, 3] MKSSS Cummins College of Engineering for Women, Pune

E-mail: nrupal.sankpal@cumminscollege.in

Received: Sep 10, 2025 *Revised: Dec 18, 2025* *Accepted: Apr 22, 2026* *Available online: Jun 15, 2026*

***Abstract*—** Lung adenocarcinoma (LUAD), a major subtype of non-small cell lung cancer, poses a persistent health challenge due to the lack of precise biomarkers that can guide early detection and treatment decisions. Major challenge in this multiomics integration lies in missing values and lack of paired patient samples, across different patient cohorts. In this study we combined miRNA and RNA expression profiles for LUAD patients. Applied GAIN to impute missing values without discarding the unpaired data. Our pathway level integration framework uses unpaired datasets, while maintaining its biological significance. LASSO based feature selection was implemented to identify stable genes associated with LUAD for RNA and miRNA separately. KEGG enrichment analysis of selected miRNA and RNA was performed and overlapping pathways were identified to select candidate genes. Five pathway supported candidate genes (CYP17A1, VANGL1, NRAS, PRICKLE2, and RAC1) demonstrated strong association with LUAD in LinkedOmics dataset, based on Kaplan Mier plot and cox proportional hazard models. External validation of these candidate genes was performed using two independent GEO datasets. PubMed literature count of the genes was also included to give biological relevance from literature. Overall, this framework gives robust biomarker discovery from unpaired data.

***Keywords*—** LUAD; Prognostic biomarkers; GAIN imputation; KEGG pathways; Multi-omics integration; Kaplan–Meier Survival; Machine learning.

1. INTRODUCTION

Lung adenocarcinoma (LUAD), the most widespread variant of non-small cell lung cancer (NSCLC), remains a major cause of cancer-related deaths worldwide. Despite advances in diagnostics and therapeutics, early detection and personalized treatment of LUAD continue to be significant clinical challenges. High-throughput technologies have generated vast multi-omics datasets, including transcriptomics, epigenomics, proteomics, and metabolomics, offering unprecedented opportunities to dissect the molecular complexity of LUAD. However, integrating these datasets into actionable clinical insights remains difficult.

Recent studies have leveraged multi-omics and computational approaches to improve biomarker discovery in LUAD. Zhao et al. [1] utilized Wasserstein distance to quantify inter-omics relationships, while Zhang et al. [2] highlighted mitochondrial gene signatures associated with patient prognosis. Han et al. [3] employed machine learning approaches to integrate multi-omics datasets for classifying LUAD subtypes, and Luo et al. [4] reported hypoxia-related transcriptional patterns with validated clinical relevance. Several studies have also included experimental or cohort-level validation: Wu et al. [5] identified immune-related prognostic candidates supported by immunological assays; Srivastava et al. [6] demonstrated

* Corresponding author

DOI: https://doi.org/10.5455/jjee.204-1757512496

proteogenomic markers with potential therapeutic implications; and Xu et al. [7] used internal and external validation to establish ligand–receptor gene pairs associated with survival. Furthermore, Bourbonne et al. [8] emphasized multi-omics signatures for immunotherapy response, Yang et al. [9] characterized mutation-defined LUAD subtypes, and Kong et al. [11] provided miRNA-based biomarkers for stage-specific diagnosis. Additional work has expanded understanding of epigenetic and metabolic influences on LUAD progression [12–15].

In addition, broader computational frameworks have contributed to methodological advances beyond LUAD. MOFA (Multi-Omics Factor Analysis) [16] and iClusterPlus [17] are among the most widely used methods for unsupervised multi-omics integration and latent-space modeling; however, both require paired multi-omics samples and typically discard unmatched data, limiting applicability to TCGA datasets with incomplete omics coverage. Recent deep learning–based multimodal integration frameworks, including autoencoders and graph neural networks, have improved cancer subtype classification and prognostic modeling by integrating multi-omics data [18-20].

However, these methods often rely on fully paired datasets and produce latent representations that are difficult to interpret biologically, motivating pathway-guided strategies for unpaired omics data that preserve interpretability while maintaining predictive performance.

Abu-Doleh & Al Fahoum [21] introduced XgCPred, an interpretable XGBoost-based clinical prediction model emphasizing transparent feature attribution; however, it is not designed for high-dimensional multi-omics integration or missing-data scenarios typical in TCGA datasets.

Although these studies collectively demonstrate the promise of multi-omics approaches, three methodological limitations remain insufficiently addressed:

1. Limited integration across heterogeneous omics layers, especially between miRNAs and RNAs where most studies analyze each layer independently or rely on paired datasets that exclude unmatched samples.
2. Feature-selection pipelines often prioritize predictive accuracy over biological interpretability, resulting in statistically significant features whose mechanistic roles remain unclear.
3. Pathway-level convergence between different omics signals is rarely evaluated systematically, even though convergent pathway involvement may better reflect biologically meaningful markers than single-gene associations.

To address these gaps, we developed a structured, biologically guided multi-omics workflow integrating miRNA and RNA expression data from LUAD patients. Missing values were imputed using GAIN, followed by a LASSO based feature selection. Predicted miRNA targets and selected RNA features underwent KEGG pathway enrichment, and intersected pathways were used to prioritize biologically coherent gene sets.

Candidate genes were subsequently evaluated using survival analysis, including Kaplan–Meier estimation and Cox proportional hazards models, to assess their association with overall survival. Validation was supported through curated biological databases and recent literature to reinforce biological plausibility.

By combining robust statistical selection with pathway-level biological convergence and survival-based evaluation, this study provides a transparent and interpretable framework for identifying prognostic biomarkers in LUAD.

2. METHODOLOGY

2.1. Data Gathering and Preprocessing

miRNA and RNA LUAD expression profiles, along with its associated clinical data, were obtained from the LinkedOmics database [22]. The expression matrices of RNA and miRNA were transposed to get rows as samples and genes as columns. Invalid or missing values were converted to NaN for preprocessing.

Prior to imputation, both datasets were normalized to the [0,1] range using Min–Max scaling. Binary mask matrices were generated to indicate observed and missing entries. Small Gaussian noise N(0,1) was added only at missing positions during model initialization, while original observed values were retained to allow reconstruction back to the original scale following imputation.

2.2. Missing Value Imputation using GAIN

Missing values were imputed using a Generative Adversarial Imputation Network (GAIN) [23, 24]. The GAIN model consists of a generator and discriminator trained adversarially, incorporating a hint mechanism that partially reveals missingness information to stabilize training.

Both miRNA-seq and RNA-seq datasets were normalized to the [0,1] range prior to training. The generator and discriminator architectures consisted of two hidden layers with 256 units each and ReLU activation, followed by a sigmoid output layer. The generator input comprised the concatenation of the data matrix and its corresponding mask.

Model training was carried out using the Adam optimizer (learning rate = 0.001), with a hint rate of 0.9 and reconstruction loss weight α=10. The loss function combined binary cross-entropy for adversarial learning and mean squared error for reconstruction. Training was conducted for 1,000 epochs(miRNA) and 100 epochs (RNA) with a batch size of 128. Convergence was determined by observing stable discriminator and generator losses, and missing entries were replaced with generator outputs and rescaled to the original data range.

2.3. Integration of Clinical Survival Data

Overall survival (OS) data was retrieved using patient-specific identifiers from the clinical metadata. OS was denoted as the duration death/last follow up from the initial diagnosis, along with corresponding vital status (dead or alive). Only patients with complete survival time and event status were included in downstream analyses. Survival time was retained on its original scale and used directly in Cox proportional hazards and Kaplan–Meier analyses.

2.4. Feature Selection using LASSO Regression Model

Independent feature selection was performed separately for RNA-seq and miRNA-seq datasets using LASSO regression as an exploratory filtering step. Overall survival was treated as a continuous variable and log-transformed using the log1p function to reduce skewness. Expression matrices were merged with corresponding clinical data and standardized using z-score normalization prior to model fitting.

LASSO regression with five-fold cross-validation was implemented using the LassoCV algorithm from the scikit-learn library (version 1.3.0), with a maximum of 50,000 iterations and a fixed random seed (random state = 42) to ensure reproducibility. Features with non-zero regression coefficients were retained as candidate biomarkers and forwarded for downstream pathway enrichment, multi-omic integration, and survival-based analyses.

2.5. KEGG Enrichment Analysis of RNA and miRNA

Functional enrichment analysis was conducted using the gseapy Enrichr tool with the KEGG 2021 Human pathway database. For RNA-seq, enrichment was performed on a small gene set (16 genes) and is therefore considered exploratory; pathways were reported without multiple-testing correction.

For miRNA-seq, five miRNAs identified through LASSO were mapped to experimentally validated target genes using miRTarBase[24], yielding approximately 4,000 target genes. KEGG enrichment was performed on this gene set, and statistically significant pathways exhibiting a false discovery rate (FDR) below 0.05 were considered were selected.

2.6. Identification of Common Pathways and Biomarkers

KEGG pathways enriched from RNA-seq and miRNA-derived target gene analyses were intersected to identify shared biological processes. For each shared pathway, overlapping genes between RNA-derived and miRNA-derived gene sets were extracted. Genes supported by both omic layers within the same pathway were prioritized as candidate multi-omic biomarkers and forwarded for prognostic modeling.

2.7. Prognostic Signature and Survival Analysis

Univariate Cox proportional hazards regression was applied to pathway-derived candidate features. Features passing Benjamini–Hochberg FDR < 0.05 were retained for multivariable Cox modeling. A prognostic risk score for each patient was calculated as a linear combination of feature expression values weighted by Cox regression coefficients.

Patients were stratified into high and low risk groups depending on the median risk score. Kaplan–Meier survival analysis, time-dependent ROC analysis, and concordance indices (C-index) were used to assess prognostic performance.

2.8. External Validation Using Independent GEO Cohorts

The prognostic signature was validated in independent GEO datasets [25] GSE42127 and GSE72094. Expression profiles were processed, probes mapped to gene symbols, and multiple probes per gene were averaged. Risk scores were calculated using coefficients obtained from LinkedOmics LUAD cohort. Survival analysis was performed using univariate Cox regression, Kaplan–Meier analysis, and C-index estimation.

2.9. Literature Validation

The biological and clinical significance of the identified biomarker genes was assessed through a review of published literature available in the PubMed database [26]. Each identified gene was queried using terms “Lung Adenocarcinoma” and “Lung Cancer”. The

publication count for each gene was considered as an indicator of prior research evidence. This approach helped us to find survival related biomarker genes with limited literature coverage.

3. RESULTS

3.1. Overview of the Study Design and Datasets

Independent RNA and miRNA datasets with associated clinical survival information were analyzed using a pathway-guided multi-omics framework. As the datasets were unpaired, all analyses were performed separately for each omic layer, followed by convergence at the pathway level. Overall survival (OS) was used as the primary clinical endpoint.

3.2. Feature Selection in RNA-seq and miRNA-seq Datasets

Independent feature selection was performed for RNA-seq and miRNA-seq datasets using LASSO regression with cross-validation in an exploratory setting. In the RNA-seq dataset, 16 genes were selected as candidate prognostic features: C15orf28, C1orf172, CLVS2, CYP17A1, DHDDS, HPX, HSDL1, INVS, LASS4, LDLRAD3, MAGEB18, NTS, SGK196, USP50, WFDC2, and ZNF121. In the miRNA-seq dataset, five miRNAs were selected: hsa-mir-122, hsa-mir-1290, hsa-mir-29c, hsa-mir-374a, and hsa-mir-876. These selected features were forwarded for pathway enrichment and integrative survival analysis.

3.3. Pathway Enrichment Analysis and Identification of Common Pathways

KEGG pathway enrichment analysis was conducted separately for RNA-derived genes and miRNA-associated target genes. The RNA-based analysis identified a limited number of significantly enriched pathways, primarily related to metabolic and hormone biosynthesis processes, including steroid hormone biosynthesis and prolactin signaling.

In contrast, the miRNA-based analysis yielded 105 significantly enriched KEGG pathways, encompassing cancer-related signaling mechanisms such as pathways in cancer, Wnt signaling, focal adhesion, cellular senescence, and immune-related pathways.

Intersection of RNA- and miRNA-derived KEGG results identified two common pathways (Table 1) shared across both omics layers:

(i) Wnt signaling pathway and
(ii) Prolactin signaling pathway.

Genes participating in these shared pathways were extracted and prioritized for downstream prognostic evaluation.

Table 1. Common KEGG pathways identified across RNA- and miRNA-based analyses.

Pathway	Omics layers
Wnt signaling pathway	RNA & miRNA
Prolactin signaling pathway	RNA & miRNA

3.4. Identification of Prognostic Pathway-Derived Features

A total of 79 pathway-derived candidate features were evaluated in the discovery cohort comprising 427 patients. Univariate Cox proportional hazards regression followed by Benjamini–Hochberg correction identified five features significantly associated with overall survival (FDR < 0.05) refer Table 2.

Table 2. Univariate Cox regression of pathway-derived features in the discovery cohort.

Gene	HR	95% CI	p-value	FDR
CYP17A1	0.053	0.015–0.184	4.0×10^{-6}	3.16×10^{-4}
VANGL1	6.668	2.286–19.450	5.13×10^{-4}	0.020
NRAS	5.350	1.934–14.801	0.00124	0.033
PRICKLE2	0.202	0.073–0.555	0.00195	0.039
RAC1	4.243	1.636–11.002	0.00295	0.047

3.5. Construction of prognostic signature

The five FDR-significant features were incorporated into a multivariable Cox proportional hazards model. The final model demonstrated strong prognostic performance (log-likelihood ratio test $p < 0.001$) with a concordance index of 0.651 referring Table 3.

Table 3. Multivariable Cox regression model for the five-gene prognostic signature.

Gene	Coefficient	HR	95% CI	p-value
CYP17A1	−2.307	0.100	0.028–0.357	<0.0005
VANGL1	1.050	2.858	0.814–10.044	0.101
NRAS	0.241	1.273	0.349–4.646	0.715
PRICKLE2	−1.044	0.352	0.125–0.990	0.048
RAC1	0.757	2.132	0.794–5.722	0.133

Model statistics:

- Concordance index: 0.651
- Log-likelihood ratio test: $p < 0.001$
- Partial AIC: 1539.48

3.6. Prognostic Performance of the Risk Score in the Discovery Cohort

A prognostic risk score was computed as a weighted linear combination of gene expression values using the multivariable Cox coefficients. Univariate Cox analysis of the risk score demonstrated a strong association with overall survival (HR = 2.105, 95% CI: 1.693–2.616, $p < 0.0005$) refer Table 4.

Table 4. Univariate Cox analysis of the risk score (training cohort).

Variable	HR	95% CI	p-value	C-index
Risk Score	2.105	1.693–2.616	<0.0005	0.639

Patients were divided into two risk groups high and low based on risk scores median value. Kaplan–Meier survival analysis displayed a clear partition in overall survival between these two groups (log-rank $p < 0.001$, Fig. 1). The prognostic performance of the signature was further assessed using receiver operating characteristic (ROC) analysis, yielding a C-index of 0.639, indicating good discriminative ability (Fig. 2).

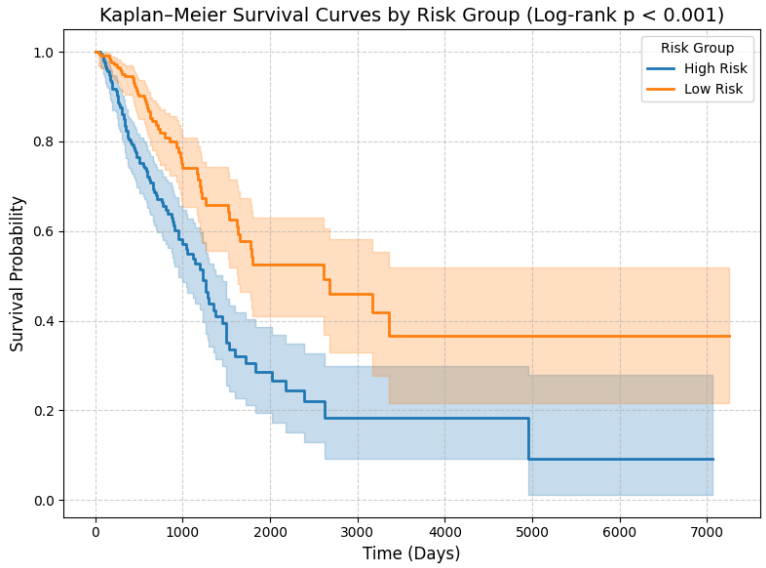

Fig. 1. Kaplan–Meier survival curves for high- and low-risk groups in the training cohort.

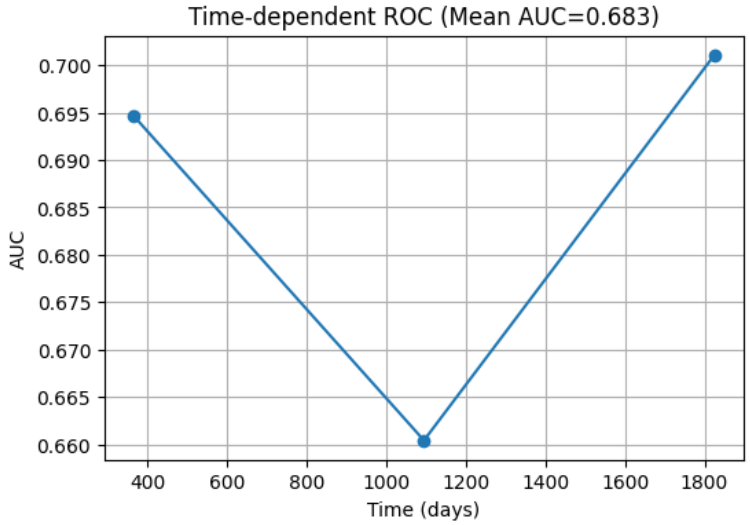

Fig. 2. Time-dependent ROC curve for the prognostic risk score in the training cohort.

3.7. External Validation in GSE42127

External validation was performed in the GSE42127 cohort (n = 176; 64 events). All five signature genes were present. The risk score remained significantly associated with survival (HR = 1.209, p = 0.014), with a C-index of 0.614 refer Table 5. Patients were divided into high- and low-risk groups according to the median risk score, and Kaplan–Meier survival analysis revealed a significant difference in overall survival between the two groups. (Fig. 3).

Table 5. External validation of the prognostic risk score in GSE42127.

Variable	HR	95% CI	p-value	C-index
Risk Score	1.209	1.039–1.406	0.014	0.614

Univariate Cox analysis of individual genes showed significant associations for NRAS and PRICKLE2. (Table 6)

Table 6. Univariate Cox analysis of individual genes in GSE42127.

Gene	HR	p-value
CYP17A1	0.978	0.942
VANGL1	1.614	0.052
NRAS	1.741	0.011
PRICKLE2	0.656	0.0017
RAC1	0.926	0.783

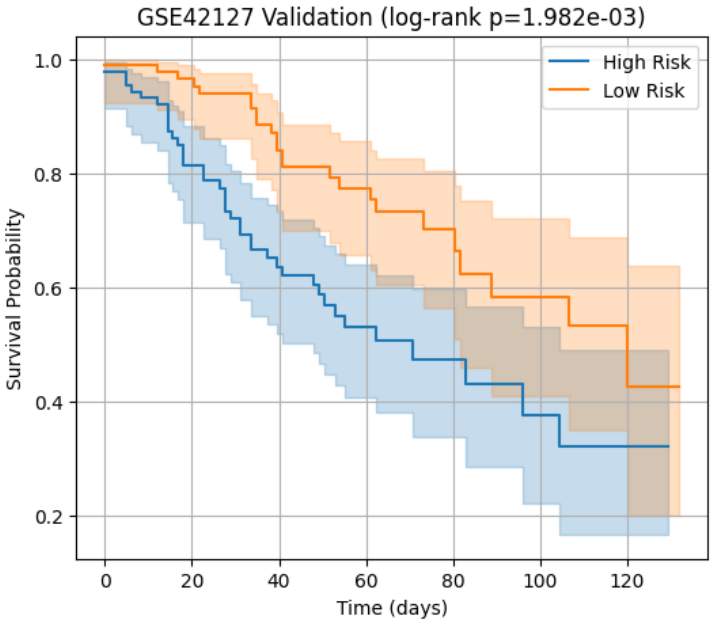

Fig. 3. Kaplan–Meier survival curves for high- and low-risk groups in GSE42127.

3.8. External Validation in GSE72094

The prognostic signature was further evaluated in GSE72094 (n = 442; 122 events). The risk score showed borderline statistical significance (HR = 1.104, p = 0.050) with a C-index of 0.562 (Table 7). Based on the median risk score, patients were categorized into high- and low-risk groups, with Kaplan–Meier analysis showing a clear survival separation. (Fig. 4).

Table 7. External validation of the prognostic risk score in GSE72094.

Variable	HR	95% CI	p-value	C-index
Risk Score	1.104	1.000–1.218	0.050	0.562

Among individual genes, VANGL1 remained significantly associated with survival (Table 8).

Table 8. Univariate Cox analysis of individual genes in GSE72094.

Gene	HR	p-value
CYP17A1	0.870	0.307
VANGL1	1.728	0.0049
NRAS	1.285	0.220
PRICKLE2	0.876	0.357
RAC1	1.048	0.912

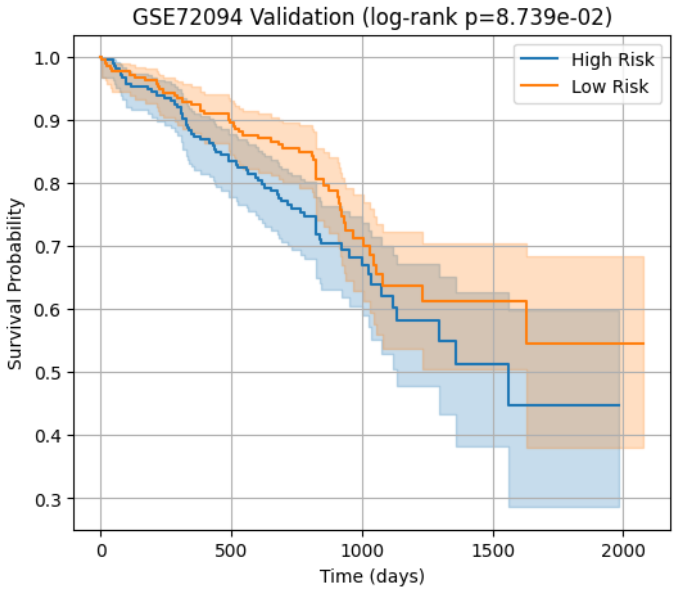

Fig. 4. Kaplan–Meier survival curves for high- and low-risk groups in GSE72094.

3.9. PubMed Literature Validation

To determine the extent of existing research support, each of the six biomarkers was queried in PubMed. The number of matching publications ("PubMed hits") served as a metric of prior biological evidence. NRAS and PIK3R1 showed high literature support, whereas SOCS6, MAPK13, CCND2, and CYP17A1 had fewer citations, indicating potential novelty. PubMed hit counts are summarized in Table 9.

Table 9. Number of PubMed articles associated with each biomarker gene.

Gene	PubMed Articles Found
CYP17A1	9
VANGL1	1
NRAS	85
PRICKLE2	3
RAC1	55

4. DISCUSSION

Conventional multiomics integration approach relies on same patient data across different omic profiles. Our approach integrates different patient samples across omics, as a result it improves the generalizability.

Key strength of our approach is the pathway level integration of distinct omics data. RNA and miRNA features were analyzed separately. Biological convergence was achieved by finding shared KEGG pathways and intersecting genes across omic layers. This approach avoids requirement of patient sample pairing, a key limitation of commonly used multiomics integration methods such as MOFA, iCluster. By using pathway guided integration, it finds biologically relevant signals, in an unpaired data.

The obtained prognostic signatures (CYP17A1, VANGL1, NRAS, PRICKLE2, RAC1) demonstrate robust prognostic values in training dataset and keeps predictive ability in two GEO cohorts used for external validation. The risk score derived from multivariate cox models effectively classified patients in low-risk and high-risk groups, as evidenced by Kaplan Meier survival differences and consistent with concordance index between different datasets. External validation was performed using coefficients derived from training cohort.

Prognostic performance slightly decreases in external cohorts specifically in GSE72094. It is expected because there is difference in patients' population, platform used and methods used for preprocessing. Though these variations are present, signature maintained statistically or borderline significant association with survival and concordance indices. These findings support the pathway guided signature identification across different cohorts.

The biological relevance of the genes included in the prognostic signature further supports the robustness of the proposed pathway-guided multi-omics framework. NRAS and RAC1 are key regulators of oncogenic signaling cascades, including MAPK and PI3K–AKT pathways, and are known to influence tumor cell proliferation, migration, and survival in lung adenocarcinoma and other solid tumors [27, 28]. VANGL1 and PRICKLE2 are core components of the planar cell polarity signaling pathway, which has been implicated in regulating cytoskeletal organization, directional cell movement, and invasive behavior in cancer progression [29, 30]. CYP17A1, a critical enzyme involved in steroid hormone biosynthesis, has been associated with tumor growth dynamics and adverse survival outcomes across multiple cancer types, highlighting its potential prognostic relevance [31]. Notably, the convergence of these genes through both RNA- and miRNA-driven pathway enrichment analyses suggests that they represent biologically coherent prognostic markers rather than isolated statistical associations, reinforcing the value of pathway-level integration in unpaired multi-omics data.

The use of GAIN imputation effectively without reducing size of data. LASSOCV and FDR controlled survival ensured robust feature selection. RNA pathway enrichment was treated as exploratory due to small size of genes, whereas miRNA derived pathways use stringent multiple testing correction (p-adjusted <0.05).

This study has several limitations such as clinical covariates (age, smoking status, tumor stage) were not included in survival models. Although prognostic signatures were validated with two GEO datasets, entire analysis were retrospective. Experimental validation is necessary to confirm biological mechanism of identified biomarkers.

Overall, this study shows that multiomics integration is possible even if paired samples are absent, it offers alternative omic integration approach over paired data integration.

5. CONCLUSIONS

We propose a pathway level multiomics integration framework designed for unpaired RNA and miRNA data. It addresses limitations of existing multiomics integration approach.

By using pathway guided integration, it preserves sample size and avoids need of matched patient data.

The identified five genes showed consistent survival classification in training and external validation cohorts. It highlights the effectiveness of pathway level integration for unpaired omic data.

This approach provides a generalizable approach for multiomics biomarker identification, where entire multiomics profile is not available. Future study will focus on using additional omic layers and clinical covariates. Experimental validation will enhance the clinical impact of proposed framework.

Data Availability Statement: All datasets analyzed in this study were obtained from publicly accessible repositories, as detailed in the manuscript. To support transparency and reproducibility, the full analysis workflow—including preprocessing, feature selection, multi-omics integration, and survival modeling—has been made openly available at: https://github.com/Multiomics-dot/LUAD-biomarker-identification.

Acknowledgement: The author(s) would like to thank MKSSS Cummins College of Engineering for Women, Pune for providing the necessary facilities to carry out this work.

REFERENCES

[1] S. Zhao, S. Huang, L. Yang, W. Zhou, K. Li, S. Wang, "Detection of LUAD-associated genes using Wasserstein distance in multi-omics feature selection," *Bioengineering*, vol. 12, no. 7, p. 694, 2025, doi: 10.3390/bioengineering12070694.

[2] W. Zhang, L. Zhao, T. Zheng, L. Fan, K. Wang, G. Li, "Comprehensive multi-omics integration uncovers mitochondrial gene signatures for prognosis and personalized therapy in lung adenocarcinoma," *Journal of Translational Medicine*, vol. 22, no. 1, p. 952, 2024, doi: 10.1186/s12967-024-05754-y.

[3] T. Han, Y. Bai, Y. Liu, Y. Dong, C. Liang, L. Gao, J. Zhou, J. Guo, J. Wu, D. Hu, "Integrated multi-omics analysis and machine learning to refine molecular subtypes, prognosis, and immunotherapy in lung adenocarcinoma," *Functional and Integrative Genomics*, vol. 24, no. 4, p. 118, 2024, doi: 10.1007/s10142-024-01388-x.

[4] Q. Luo, X. Li, Z. Meng, H. Rong, Y. Li, G. Zhao, H. Zhu, L. Cen, Q. Liao, "Identification of hypoxia-related gene signatures based on multi-omics analysis in lung adenocarcinoma," *Journal of Cellular and Molecular Medicine*, vol. 28, no. 2, p. e18032, 2024, doi: 10.1111/jcmm.18032.

[5] S. Wu, J. Pan, Q. Pan, L. Zeng, R. Liang, Y. Li, "Multi-omics profiling and experimental verification of tertiary lymphoid structure-related genes: molecular subgroups, immune infiltration, and prognostic implications in lung adenocarcinoma," *Frontiers in Immunology*, vol. 15, p. 1453220, 2024, doi: 10.3389/fimmu.2024.1453220.

[6] S. Srivastava, N. Jayaswal, S. Kumar, P. Sharma, T. Behl, A. Khalid, S. Mohan, A. Najmi, K. Zoghebi, H. Alhazmi, "Unveiling the potential of proteomic and genetic signatures for precision therapeutics in lung cancer management," *Cellular Signalling*, vol. 113, p. 110932, 2024, doi: 10.1016/j.cellsig.2023.110932.

[7] S. Xu, X. Chen, H. Ying, J. Chen, M. Ye, Z. Lin, X. Zhang, T. Shen, Z. Li, Y. Zheng, D. Zhang, "Multi-omics identification of a signature based on malignant cell-associated ligand–receptor genes for lung adenocarcinoma," *BMC Cancer*, vol. 24, no. 1, p. 1138, 2024, doi: 10.1186/s12885-024-12911-5.

[8] V. Bourbonne, M. Geier, U. Schick, F. Lucia, "Multi-omics approaches for the prediction of clinical endpoints after immunotherapy in non-small cell lung cancer: a comprehensive review," *Biomedicines*, vol. 10, no. 6, p. 1237, 2022, doi: 10.3390/biomedicines10061237.

[9] X. Yang, M. Li, Z. Chen, X. Fan, L. Guo, B. Jin, Y. Huang, Q. Wang, L. Wu, C. Zhan, "Multi-omics analysis identifies distinct subtypes with clinical relevance in lung adenocarcinoma harboring KEAP1/NFE2L2," *Journal of Cancer*, vol. 13, no. 5, pp. 1512–1523, 2022, doi: 10.7150/jca.66241.

[10] X. Ruan, Y. Ye, W. Cheng, L. Xu, M. Huang, Y. Chen, J. Zhu, X. Lu, F. Yan, "Multi-omics integrative analysis of lung adenocarcinoma: an in-silico profiling for precise medicine," *Frontiers in Medicine*, vol. 9, p. 894338, 2022, doi: 10.3389/fmed.2022.894338.

[11] D. Kong, K. Wang, Q. N. Zhang, Z. T. Bing, "Systematic analysis reveals key microRNAs as diagnostic and prognostic factors in progressive stages of lung cancer," *arXiv preprint*, arXiv:2201.05408, 2022.

[12] H. Li, L. Tong, H. Tao, Z. Liu, "Genome-wide analysis of the hypoxia-related DNA methylation-driven genes in lung adenocarcinoma progression," *Bioscience Reports*, vol. 40, no. 2, p. BSR20194200, 2020, doi: 10.1042/BSR20194200.

[13] A. Namani, Z. Zheng, X. J. Wang, X. Tang, "Systematic identification of multi-omics-based biomarkers in KEAP1-mutated TCGA lung adenocarcinoma," *Journal of Cancer*, vol. 10, no. 27, pp. 6813–6824, 2019, doi: 10.7150/jca.35489.

[14] Y. Tang, Z. Li, L. Lazar, Z. Fang, C. Tang, J. Zhao, "Metabolomics workflow for lung cancer: Discovery of biomarkers," *Clinica Chimica Acta*, vol. 495, pp. 436–445, 2019, doi: 10.1016/j.cca.2019.05.012.

[15] C. Kikutake, K. Yahara, "Identification of epigenetic biomarkers of lung adenocarcinoma through multi-omics data analysis," *PLOS ONE*, vol. 11, no. 4, p. e0152918, 2016, doi: 10.1371/journal.pone.0152918.

[16] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, O. Stegle, "Multi-omics factor analysis—A framework for unsupervised integration of multi-omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, p. e8124, 2018, doi: 10.15252/msb.20178124.

[17] Q. Mo, R. Shen, "iClusterPlus: Integrative clustering of multiple genomic data sets," *R package version 1.19*, 2018, https://bioconductor.statistik.uni-dortmund.de/packages/3.10/bioc/vignettes/iClusterPlus/inst/doc/iManual.pdf.

[18] Q. Liu, K. Song, "ProgCAE: a deep learning-based method that integrates multi-omics data to predict cancer subtypes," *Briefings in Bioinformatics*, vol. 24, no. 4, p. bbad196, 2023, doi: 10.1093/bib/bbad196.

[19] J. Wu, Z. Chen, S. Xiao, G. Liu, W. Wu, S. Wang, "DeepMoIC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification," *BMC Genomics*, vol. 25, no. 1, p. 1209, 2024, doi: 10.1186/s12864-024-11112-5.

[20] J. L. Ballard, Z. Wang, W. Li, L. Shen, Q. Long, "Deep learning-based approaches for multi-omics data integration and analysis," BioData Mining, vol. 17, no. 1, p. 38, 2024, doi: 10.1186/s13040-024-00391-z.

[21] A. Abu-Doleh, A. Al Fahoum, "XgCPred: Cell type classification using XGBoost-CNN integration and exploiting gene expression imaging in single-cell RNA-seq data," *Computers in Biology and Medicine*, vol. 181, p. 109066, 2024, doi: 10.1016/j.compbiomed.2024.109066.

[22] LinkedOmics, "TCGA-LUAD multi-omics dataset," LinkedOmics Data Download Portal, Lung Adenocarcinoma (TCGA-LUAD), 2025, https://www.linkedomics.org /data_download/TCGA-LUAD.

[23] W. Dong, D. Fong, J. Yoon, E. Wan, L. Bedford, E. Tang, C. Lam, "Generative adversarial networks for imputing missing data for big data clinical research," *BMC Medical Research Methodology*, vol. 21, no. 1, p. 78, 2021, doi: 10.1186/s12874-021-01272-3.

[24] J. Yoon, J. Jordon, M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," International Conference on Machine Learning, 2018, doi: 10.48550/arXiv.1806.02920.

[25] R. Edgar, M. Domrachev, A. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002, doi: 10.1093/nar/30.1.207.

[26] miRTarBase, "Experimentally validated microRNA–target interactions," miRTarBase Database, 2023, http://mirtarbase.cuhk.edu.cn/.

[27] National Center for Biotechnology Information, 2025, https://pubmed.ncbi.nlm.nih.gov.

[28] Y. Pylayeva-Gupta, E. Grabocka, D. Bar-Sagi, "RAS oncogenes: Weaving a tumorigenic web," *Nature Reviews Cancer*, vol. 11, no. 11, pp. 761–774, 2011, doi: 10.1038/nrc3106.

[29] H. Bid, R. Roberts, P. Manchanda, P. Houghton, "RAC1: an emerging therapeutic option for targeting cancer angiogenesis and metastasis," *Molecular Cancer Therapeutics*, vol. 12, no. 10, pp. 1925–1934, 2013, doi: 10.1158/1535-7163.MCT-13-0164.

[30] A. Humphries, M. Mlodzik, "From instruction to output: Wnt/PCP signalling in development and cancer," *Current Opinion in Cell Biology*, vol. 51, pp. 110–116, 2018, doi: 10.1016/j.ceb.2017.12.005.

[31] J. Hatakeyama, J. H. Wald, I. Printsev, H. Y. Ho, K. L. Carraway, "Vangl1 and Vangl2: Planar cell polarity components with a developing role in cancer," *Endocrine-Related Cancer*, vol. 21, no. 5, pp. R345–R356, 2014, doi: 10.1530/ERC-14-0141.