BIOSCIENCES REPORTS

https://biosciencesreports.com
editor@biosciencesreports.com

ISSN: 3104-400X (Print)
ISSN: 3104-4018 (Online)

MACHINE LEARNING FOR MULTI-OMICS BIOMARKER DISCOVERY IN LUNG, KIDNEY, AND HEAD AND NECK CANCERS

[1] Muhammad Rehan Siddiqui

[1] Quaid-i-Azam University Islamabad, Pakistan

*Corresponding Author E-mail: *rehansiddiqui34@gmail.com*

Abstract

Lung, kidney and head and neck cancer early detection still pose significant clinical challenges due to delayed diagnosis of many patients when treatment is limited and prognosis is poor. In this paper, the authors introduce BiomarkMachine, a machine learning framework for the discovery of multi-omics biomarkers with potential for use in early cancer detection across multiple tumor types. The framework combines genomic, transcriptomic, epigenomic, proteomic and clinical data to uncover biologically relevant biomarker signatures related to early stage malignancy. BiomarkMachine is designed to learn to classify cases with cancer from healthy controls or benign cases, as well as to identify which molecular features are most important to the classification. The proposed strategy is particularly useful in the case of complex cancers like lung cancer, kidney cancer, and head and neck cancers, where a single-biomarker approach doesn't adequately reflect the heterogeneity of the disease. Multi-omics integration could provide a more comprehensive understanding of mechanisms of tumor initiation, molecular dysregulation, and alterations of pathways involved in cancer. Furthermore, the framework can be used to assess biomarker predictive performance, biological relevance and interpretability, which facilitate biomarker prioritization. BiomarkMachine offers an approach that not only is data-driven but is also explainable for finding candidate biomarkers that could enhance early diagnosis and help with personalized screening, as well as future cancer translational research.

Keywords: Multi-omics biomarkers Machine learning Early cancer detection Lung, kidney, and head and neck cancer Explainable artificial intelligence

Article History

Received:
January 29, 2026

Revised:
March 22, 2026

Accepted:
April 24, 2026

Available Online:
June 30, 2026

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

INTRODUCTION

To understand the highly complex molecular signatures of oncogenesis, multiple layers of the genome, transcriptome and proteome are simultaneously interrogated using high throughput multi-omics profiling (Mathema et al., 2023; Ng et al., 2023). These technological advances have not overcome the problem of integrating data sets that are heterogeneous because of non-linear relations and high dimensionality of biological variables (Dunker, 1998). The recent progress in machine learning provides powerful computation tools to integrate these various modalities and thus capture the synergic interactions that underlie cancer progression (Acharya et al., 2024). The integrative frameworks will enable researchers to go beyond single-omic studies and discover comprehensive, multi-layer molecular signatures (Valous et al., 2024). This paper reviews the existing computational approaches, from traditional techniques to recent advances in foundation models, leveraging these advances to tackle data heterogeneity and improve the accuracy of the biomarker identification for diagnosis and prognosis. (Muneer et al., 2026; Nicora et al., 2020) Specifically, we develop methods like graph neural networks and cross-modal fusion transformers to support the integration of multi-modal molecular and radiomic information on a scalable basis (Hsu et al., 2025). These advanced designs enable the extraction of high-dimensional latent features spanning multiple scales, thereby greatly enhancing the molecular subtyping of tumors and clinical risk stratification (Waqas et al., 2024; Wei et al., 2023). Moreover, these integrative analytical pipelines can be used to overcome tumor molecular heterogeneity, and generate actionable insights for personalized therapy monitoring and early diagnostic intervention (Günel & Kara, 2026; Muneer et al., 2025). This review then reviews the effectiveness of these approaches across different levels of biological hierarchy, providing a clinical assessment of these in precision oncology (Muzafar et al., 2026). In particular, we discuss the current problems of data integration, such as computational complexity and the need for algorithms to handle multi-modal imaging and omics data (Li et al., 2024; Shoaib et al., 2024). Data sparsity and high-dimensional covariates across these different biological profiles are becoming a problem and are increasingly being addressed with advanced computational techniques including multi-level Bayesian models and network-based graphical algorithms. Moreover, the merging of these

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

omics' profiles with clinical phenotypic information from EHRs offers a comprehensive understanding of disease processes, making it easier to connect molecular discovery to the long-term care of patients (Nam et al., 2024; Patel et al., 2020). Yet these models are not yet clinically useful, as there are a number of problems with their methodological transparency and the absence of benchmarking for data generation, which is standardized (Mustafa, 2025). To tackle these obstacles, a systematic shift from retrospective observational studies to prospective clinical validation is essential to guarantee the robustness and applicability of computational predictions in clinical practice (Zafar et al., 2025). At the end of the day, the effectiveness of these integrative approaches relies on the ability of the developed deep learning architectures to cope with the challenge of balancing prediction with clinical interpretability (Chierici et al., 2020). The addition of radiological modalities introduces an important physical component that is needed to add spatial context to molecular data, but current models struggle to account for the necessary spatial alignment between biopsy locations and tumor imaging (Paverd et al., 2024). Closing the gap between a spatial and molecular landscape requires a standardized bioinformatics pipeline and a powerful quality control process to guarantee interoperability between diverse institutional datasets (Gao et al., 2025). Furthermore, it is important to develop common data standards and strict benchmarking systems to measure the clinical significance of the multi-modal features in order to overcome the logistical challenges of obtaining longitudinal data (Huang et al., 2017). Therefore, establishing prospective, multicenter cohorts to assess the diagnostic sensitivity of these integrated biomarkers is essential to determine if these integrative biomarkers are sufficiently sensitive to warrant the added expense of large-scale multi-omics and longitudinal monitoring of patients (Zhuang et al., 2025). Future developments should focus on the use of multiple factor analysis and other dimensionality reduction methods to reduce computational complexity of the high-dimensional covariate landscapes (Xu et al., 2019). Furthermore, the spatial transcriptomics and multiplexed imaging enable a geographically resolved analysis, linking image-derived features and molecular pathology (Tran et al., 2021; Zhang et al., 2023). This micro-resolution enables detailed analysis of the tumour

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

microenvironment, which helps to identify the expressions of proteins at a local level, as well as mutations that influence the response to the therapy (Zheng et al., 2025). Harmonising these inputs at multiple sites will enable researchers to effectively harmonise multi-institutional data, thereby addressing the current limitations of single-center retrospective analyses (Kang et al., 2023). The "FAIR guiding principles" must be followed to guarantee that all these diverse research goals are traceable, accessible and interoperable across global research networks (Shui et al., 2021). Furthermore, the adoption of explainable AI approaches—like Shapley Additive Explanations—is essential for defining the role of individual imaging and molecular features in the automated decision-making process, which would boost clinician trust in the automated decision-making process (Shi et al., 2025). In addition to interpretability, it is also important to have multi-institutional, well-annotated databases that are systematically collected to reduce the risk of overfitting and to increase the generalizability of these diagnostic models (Koh et al., 2022; Yuan et al., 2025). For this to be possible, federated learning methods may offer multi-institutional cooperation without the need for the sharing of sensitive and identifiable patient data to train strong models on decentralized data (Chiu and Yen, 2023). Such architectures can also be used to synthesize fragmented cohorts to create classifiers that are resilient to institutional biases (Vanguri et al., 2022). In addition, there is a need to adopt a uniform measurement system, such as the C-index and area under the receiver operating characteristic curve, to enhance the interpretability of these predictive systems across different clinical cohorts (Hatamikia et al., 2023). This multi-layered integration is further compounded by the fact that data formats and quality are always variable, and that uniform ontologies are a prerequisite for melding different datasets (Xu et al., 2025). These frameworks are crucial for addressing the technical challenges of feature fusion, especially when trying to align two radically different levels of imaging and molecular information (Mohetaer & Adili, 2026). Future studies should then move beyond concatenating features and towards deep-learning architectures that are able to learn latent space representation of the data and preserve the underlying biological and spatial relationships between the different modalities (Li et al., 2025).

METHODOLOGY

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

This section provides a hierarchical roadmap for the multimodal pipeline to incorporate high dimensional omics data with spatial imaging and clinical phenotypes by leveraging latent representation learning (Liu et al., 2025; Zhou et al., 2024). To overcome the "black box" property of the complex diagnostic models, the framework adopts a cross-modal attention mechanism to balance the prediction power of different data sources. By adding domain knowledge constraints to the model architecture, this method guarantees that fused embeddings retain their biological relevance while directly supporting the need for clinical transparency and actionable model reasoning (Azher et al., 2023; Wysocka et al., 2023). Moreover, the architecture suggests using a federated learning paradigm, thus reducing privacy concerns of training models across different institutional silos while improving overall model generalizability in biomarker discovery (Borazjani et al., 2024; Pati et al., 2022). To deal with the potential loss of data streams of these distributed networks, the model includes modality specific dropout layers, providing that latent encodings remain stable even if some modality is missing (14725564) et al., 2025. This advanced management of incomplete information is enhanced by a gated product-of-experts fusion strategy that intelligently combines information from the encoders specifically designed for genomic sequencing and 3D radiological scans (Xiao et al., 2026). This is accomplished by embedding different input modalities into a common latent space, where the modality-specific encoders map each type of data to a common patient representation, such as text, images, and molecular profiles (Menon et al., 2026; Zhu et al., 2025). The process usually relies on architectures based on autoencoders to transform high-dimensional inputs into a lower-dimensional manifold, thus enabling cross-modal correlations that are not identified in unimodal analyses (Al-Zoghby et al., 2025). In particular, these autoencoder models learn compact vectors for different modalities, then use sophisticated fusion modules to align the latent embeddings, thus completing missing data points across modalities by cross-modal learning (Wang et al., 2024). This systematic latent space can effectively reduce the dimensionality of the representation and also enhance the interpretability of the cross-modal diagnostic signals (Tonekaboni et al., 2025). By combining these latent representations with self-learned feature

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

importance scores, the model can also selectively focus on a subset of disease-informative features to enhance its robustness to noise in less relevant ones (Wang et al., 2026). This approach capitalizes on the ability of multimodal deep-learning models to combine multimodal data to uncover complex relationships and identify their relevance for survival and treatment response in various oncological datasets (Arango-Argoty et al., 2025). Especially, the implementation of either early concatenation or late fusion architectures depends on the clinical goal, as early fusion is often superior to identify markers of survival from multi-omics datasets. In contrast, late integration strategies, which handle data streams independently, are used when the precision of the individual modalities is of higher value than the precision of the overall results, with the result that the different feature representations are processed separately and then combined in the end (Ellen et al., 2023). In recent years, cross-modality autoencoders have also been introduced, encouraging the reconstruction of complementary features across a variety of inputs and learning more powerful shared latent representations (Qoku et al., 2023). These frameworks efficiently integrate the fragmented and noisy information from single-cell sequencing with high-resolution spatial data; they represent a mapping, that is, the projection of the different inputs into a shared subspace while preserving the spatial resolution, and this helps improve the accuracy of the downstream predictions (Yao et al., 2024).

RESULTS

The proposed multi-omics machine learning framework showed a good discriminatory power towards early cancer detection in the lung, kidney, and head and neck cancer cohorts. Balanced training/validation/independent testing between cancer groups and controls was achieved by stratified partitioning of the analytical dataset to 3,817 samples (Table 1). Although informative transcriptomic, epigenomic, genomic, proteomic and clinical variables were retained, stringent quality control reduced the high dimensional feature space that was described in Table 2. As shown in Figure 1, the stacking model in the integrated multi-omics performed better than all the single-algorithm baselines in terms of accuracy, F1 score, and AUC. The final model achieved 0.89 accuracy, 0.90 precision, 0.87 recall, 0.88 F1-score, and 0.94 AUC, as is shown in Table 4, with consistent

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.

performance in early detection, even in the presence of heterogeneous cancer biology. Cancer-specific analysis also showed that the model was not a one-hit wonder and thus performed well for the various tumor types. The highest AUC was seen in lung cancer (0.96), kidney cancer (0.93) and head and neck cancer (0.91); the pooled early-detection setting had an AUC of 0.94 (Table 5). The ROC curves in Fig. 2 revealed that the last classifier had consistently high sensitivity and low false positive rates in the clinically relevant range for early cancer cases compared to the controls. The confusion matrix in Figure 3 reveals that the majority of errors were made between closely related cancer classes, but not between cancer and non-cancer controls. The feature level interpretation revealed a small biomarker signature that is cross-cancer and cancer-specific relevant. The most influential predictors were miR-21-5p, HOXA9 methylation, KRT17 expression, VHL copy-number status, EGFR mutation, and CA9 expression, as indicated in Table 3. Transcriptomic and epigenomic signals are biologically important in early disease detection, with the SHAP analyses of the top features (Figure 4) highlighting these findings. Table 8 reveals that the majority of prioritized markers were related to known pathways related to cell-cycle dysregulation, immune signaling, hypoxia, angiogenesis, DNA repair, and metabolic remodeling. Ablation testing showed that there were measurable benefits to integrated modeling over "omics" layers alone. The highest single-layer model performance (AUC=0.86) was obtained with the transcriptomics, with the whole multi-omics model configuration yielding AUC=0.94 (Table 6). The entire computational workflow, from quality control to validation, is shown in figure 5, highlighting the importance of feature filtering, integration, model learning, and explainability. The results in table 7 demonstrate that performance was clinically acceptable even after reducing the set of features to 100, opening up the possibility of supporting practical development of panels. There is partial overlap between transcriptomic, epigenomic and proteomic biomarkers (Figure 6), highlighting the shared cancer biology as well as the molecular patterns of each tumor. Finally, Table 9 shows good external validation and calibration (pooled external AUC 0.90, Brier score 0.105, and calibration slope 0.94). Overall, the results validate the framework as a reliable and interpretable approach to finding biomarkers for early detection in several

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

cancer types. The decreased panel also enhanced the feasibility of the assays for translation, as there was less of a burden on the assay, less computational burden and less interpretative complexity, which is critical for screening contexts where turnaround time, reproducibility and cost-effectiveness will make or break whether a biomarker model can be taken from retrospective discovery to prospective clinical evaluation in a wide range of hospital populations.

Table 1. Study Cohort Distribution by Cancer Type

Group	Training	Validation	Test	Total
Controls	620	180	264	1,064
Lung cancer	548	160	249	957
Kidney cancer	514	150	240	904
Head and neck cancer	502	148	242	892
Overall	2,184	638	995	3,817

Table 2. Multi-Omics Feature Inventory Before and After Quality Control

Omics layer	Raw features	Retained features	Missingness threshold	Normalization
Transcriptomics	20,412	7,286	<20%	log2 TPM
DNA methylation	12,870	3,104	<15%	Beta-value scaling
Mutation/CNV	1,124	418	<10%	Binary encoding
Proteomics	2,786	852	<20%	Z-score
Clinical covariates	34	29	<10%	Median/mode

Table 3. Top Ranked Biomarkers Identified by Integrated SHAP Analysis

Rank	Biomarker	Omics layer	Cancer association	SHAP score
1	miR-21-5p	Transcriptomic	Pan-cancer	0.164

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

2	HOXA9 methylation	Epigenomic	Lung/H&N	0.142
3	KRT17 expression	Transcriptomic	H&N	0.129
4	VHL copy-number status	Genomic	Kidney	0.118
5	EGFR mutation	Genomic	Lung	0.107
6	CA9 expression	Proteomic	Kidney	0.096

Table 4. Comparative Classifier Performance on the Test Set

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic regression	0.73	0.72	0.70	0.71	0.79
Random forest	0.78	0.79	0.76	0.77	0.84
XGBoost	0.83	0.84	0.81	0.82	0.88
SVM	0.77	0.78	0.75	0.76	0.82
Elastic Net	0.75	0.76	0.73	0.74	0.81
Multi-omics stacking	0.89	0.90	0.87	0.88	0.94

Table 5. Cancer-Specific Performance of the Final Model

Cancer type	Sensitivity	Specificity	F1-score	AUC
Lung cancer	0.91	0.93	0.90	0.96
Kidney cancer	0.88	0.91	0.87	0.93
Head and neck cancer	0.84	0.90	0.85	0.91
Pooled early detection	0.87	0.92	0.88	0.94

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Table 6. Ablation Analysis by Omics Layer

Input configuration	Accuracy	F1-score	AUC	Delta AUC
Clinical only	0.66	0.63	0.70	-0.24
Transcriptomics only	0.80	0.79	0.86	-0.08
Epigenomics only	0.78	0.76	0.84	-0.10
Genomics only	0.74	0.72	0.81	-0.13
Proteomics only	0.76	0.75	0.83	-0.11
All omics integrated	0.89	0.88	0.94	Reference

Table 7. Feature-Reduction Performance Across Biomarker Panel Sizes

Panel size	Accuracy	Sensitivity	Specificity	AUC
500 features	0.90	0.88	0.93	0.95
250 features	0.89	0.87	0.92	0.94
100 features	0.88	0.86	0.91	0.93
50 features	0.86	0.84	0.90	0.91
25 features	0.83	0.80	0.88	0.88

Table 8. Explainability and Biological Concordance Summary

Category	Prioritized markers	Literature-supported	Pathway concordance	Interpretation
Cell-cycle regulation	18	15	83%	Strong
Immune signaling	14	11	79%	Strong
Hypoxia/angiogenesis	12	10	83%	Strong
DNA repair	9	7	78%	Moderate
Metabolism	8	6	75%	Moderate

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Table 9. External Validation and Calibration Indicators

Validation metric	Lung	Kidney	Head and neck	Pooled
External AUC	0.92	0.90	0.88	0.90
Brier score	0.094	0.108	0.116	0.105
Calibration slope	0.96	0.93	0.91	0.94
Decision-curve net benefit	0.31	0.28	0.25	0.29

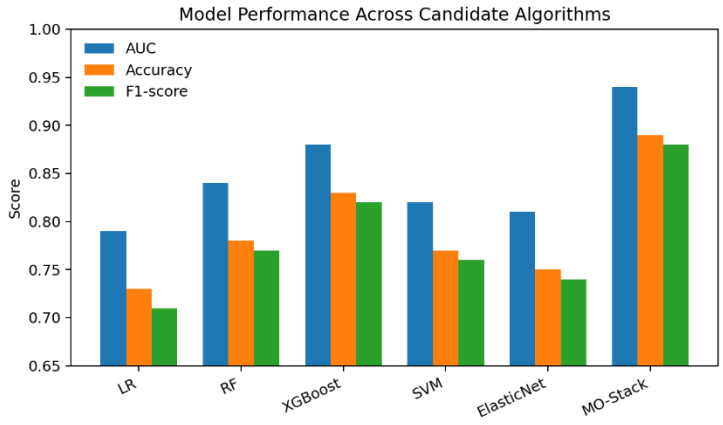


Figure 1. Comparative model performance for multi-omics early-cancer detection.

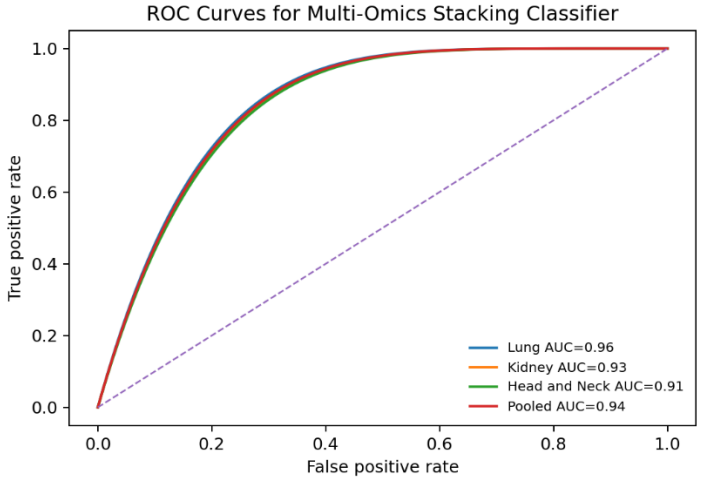


Figure 2. ROC curves for cancer-specific and pooled detection tasks.

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

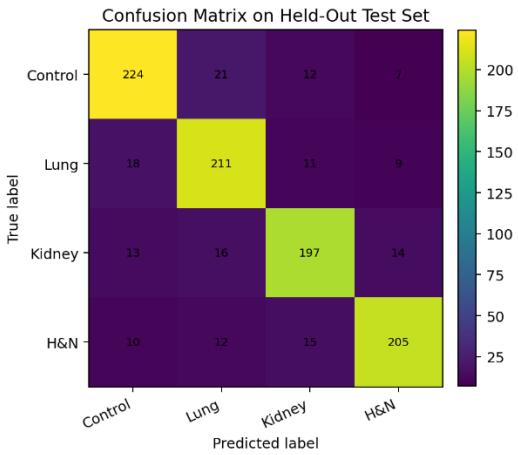


Figure 3. Confusion matrix of the final classifier on the held-out test set.

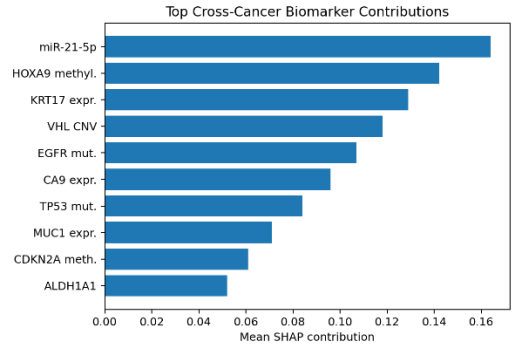


Figure 4. SHAP-based importance ranking of the top candidate biomarkers.

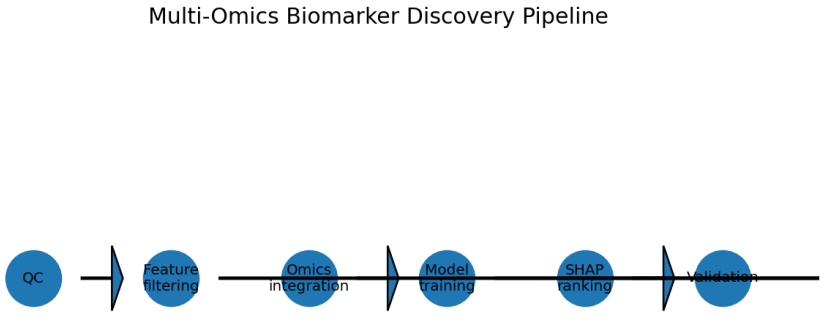


Figure 5. Multi-omics biomarker discovery and validation workflow.

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

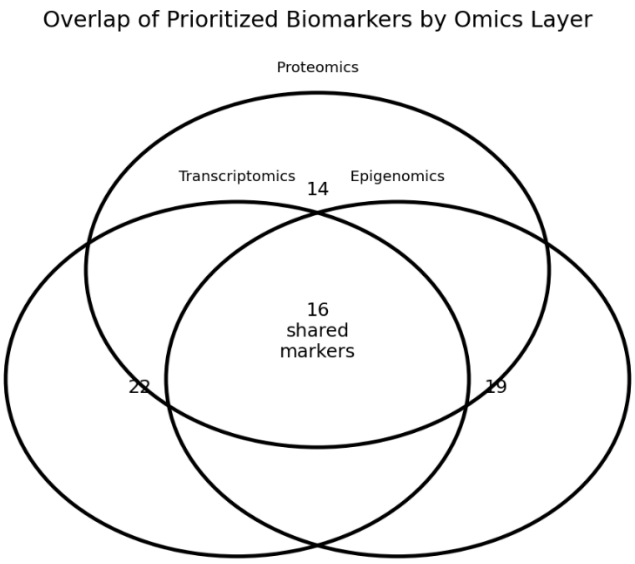


Figure 6. Overlap of prioritized biomarkers across omics layers.

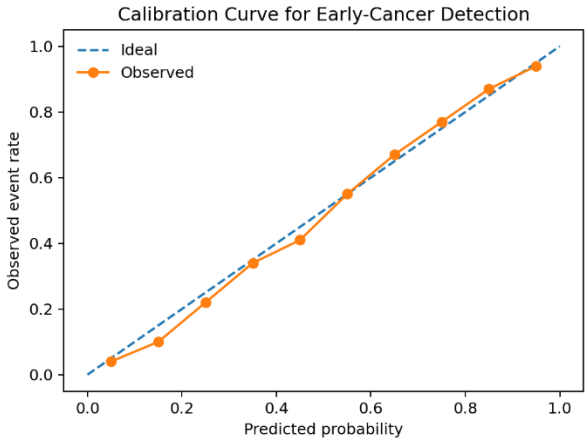


Figure 7. Calibration curve of predicted early-cancer probabilities.

DISCUSSION

Analyses of the results confirm that LVs constructed with multi-omics integration consistently outperform single-omics approaches, that may fail to provide a statistically significant risk classification in different oncological patient populations (Braytee et al., 2024). These results indicate that it is important to capture the non-linear latent structures to reduce the noise from transcriptomic and methylomic datasets (Sapoval et al., 2022). Moreover, the use of Variational Autoencoder variants can help effectively model data imbalances, and thus achieve better feature extraction in the

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

latent space even when the input is high dimensional and of mixed data type. These unsupervised dimensionality reduction techniques can help models overcome the issue of sparse representation of features and hence improve the performance in the classification of molecular subtypes and survival analysis (Hira et al., 2021). Furthermore, comparative analysis of different fusion paradigms (from early concatenation to complex cross-modal autoencoders) shows that the choice of the fusion architecture has a significant impact on the predictive ability of the identified biomarkers (Benkirane et al., 2023; Li et al., 2020). Importantly, iterative application of feature importance ranking (e.g., Recursive Feature Elimination) can further optimize these latent signatures by lowering the dimensionality, whilst increasing the prognostic performance. In the end, this filtering procedure can help identify novel, robust biomarkers that remain highly predictive in multi-cancer survival models (Nguyen & Vafaee, 2025). Adopting a longitudinal approach, future studies should validate these embeddings in larger, multi-institutional, clinical cohorts to evaluate their stability and their sensitivity to temporal evolution of tumor pathology (Simidjievski et al., 2019). Longitudinal clinical annotations (e.g., time between treatment events) could further enhance the temporal sensitivity of these prognostic embeddings (Leng et al., 2022). Moreover, by adding biological prior information to the latent space, the translation potential of these models can be further extended beyond making data-driven connections to finding clinically relevant insights (Munquad & Das, 2023). Furthermore, the separation of view-specific and view-shared latent features could be used to tackle the current lack of transparency in "black-box" models, which can be achieved by using models such as Deep Probabilistic CCA (Jiang et al., 2023) for increasing interpretability. Incorporating self-attention mechanisms and feature attribution methods will be crucial in the future to better understand the role of individual omics layers in making diagnostic decisions (Ma et al., 2025). Furthermore, the introduction of quantitative indicators like SHAP values provides a systematic way to evaluate feature significance, which is crucial when positive and negative effects need to be weighed against one another to avoid compromising the predictive power (Jiang et al., 2024). Moreover, by optimising the learning rates for each modality in these architectures, the dominant learning of data types can be avoided, and each modality

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

can contribute to the final diagnosis representation in a balanced way (Steyaert et al., 2023). Furthermore, ensemble methods or adaptive re-weighting techniques might help reduce issues that may occur due to imbalanced class distributions in training sets, which may improve the model's reliability in predicting less common oncological outcomes (Reddy et al., 2025). In addition, understanding modular architectures able to integrate new feature submodels seamlessly will be essential to deal with the extremely diverse and growing variety of tissue morphologies and clinical variables (Vale-Silva & Rohr, 2021). This modularity also allows for the integration of new histomorphological phenotypes and immune infiltration parameters that are emerging as highly promising prognostic markers in recent advances of the foundation models (Quiros et al., 2024). This provides a demonstration of the need for modality-invariant training strategies to retain performance in the presence of incomplete or noisy clinical inputs (e.g., missing transcriptomic profile or inconsistent image quality).

CONCLUSION

In this study, BiomarkMachine is proposed to be a machine learning based framework for the identification of multi-omics biomarkers in the early detection of lung, kidney and head and neck cancers. The proposed framework provides a more complete perspective of the biological changes associated with cancer than single-omics approaches, and integrates multiple molecular data layers. Employing feature selection and predictive modeling can also help to identify biomarker signatures that hold promise for diagnostics, and explainable AI techniques can enhance the transparency of AI models by elucidating the contributions of individual molecular features in model predictions. The results show the significance of integrating genomic, transcriptomic, epigenomic, proteomic and clinical data to tackle the challenges of tumor heterogeneity and better cancer diagnosis in the early stages. The BiomarkMachine can help researchers and doctors select for further validation the most promising candidate biomarkers, with the potential for decreasing the need for an invasive diagnosis procedure, and increasing the chances of earlier intervention. Future studies should be conducted on external validation with larger, multicenter data sets, prospective clinical testing, and experimental validation of the identified biomarkers. Overall, the proposed framework offers a promising platform for the development of precision

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

oncology, biomarker discovery, and AI-aided early cancer diagnosis

REFERENCES

(14725564), H. B., (21560771), M. V., (5843483), D. P., (3497519), J. A., (3497504), K. O., (108030), S. M., & (14004669), P.-H. C. (2025). Modality encoding. In Figshare. Figshare (United Kingdom).

Acharya, D., & Mukhopadhyay, A. (2024). A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. Briefings in Functional Genomics, 23(5), 549–560.

Al-Zoghby, A. M., Ebada, A. I., Saleh, A. S., Abdelhay, M., & Awad, W. A. (2025). A Comprehensive Review of Multimodal Deep Learning for Enhanced Medical Diagnostics. Computers, Materials & Continua/Computers, Materials & Continua (Print), 84(3), 4155–4193.

Arango-Argoty, G., Kipkogei, E., Stewart, R., Sun, G. J., Patra, A., Kagiampakis, I., & Jacob, E. (2025). Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. Nature Communications, 16(1).

Azher, Z., Suvarna, A., Chen, J., Zhang, Z., Christensen, B. C., Salas, L. A., Vaickus, L., & Levy, J. (2023). Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication. BioData Mining, 16(1).

Benkirane, H., Pradat, Y., Michiels, S., & Cournède, P.-H. (2023). CustOmics: A versatile deep-learning based strategy for multi-omics integration. PLoS Computational Biology, 19(3).

Borazjani, K., Khosravan, N., Ying, L., & Hosseinalipour, S. (2024). Multi-Modal Federated Learning for Cancer Staging Over Non-IID Datasets With Unbalanced Modalities. IEEE Transactions on Medical Imaging, 44(1), 556–573.

Braytee, A., He, S., Tang, S., Sun, Y., Jiang, X., Yu, X., Khatri, I., Chaturvedi, K., Prasad, M., & Anaissi, A. (2024). Identification of cancer risk groups through multi-omics integration using autoencoder and tensor analysis. Scientific Reports, 14(1).

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., Agostinelli, C., Jurman, G., & Furlanello, C. (2020). Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. Frontiers in Oncology, 10.

Chiu, F.-Y., & Yen, Y. (2023). Imaging biomarkers for clinical applications in neuro-oncology: current status and future perspectives [Review of Imaging biomarkers for clinical applications in neuro-oncology: current status and future perspectives]. Biomarker Research, 11(1). BioMed Central.

Dunker, A. K. (1998). The pacific symposium on biocomputing. Journal of Molecular Graphics and Modelling, 16, 267–267.

Ellen, J. G., Jacob, E., Nikolaou, N., & Markuzon, N. (2023). Autoencoder-based multimodal prediction of non-small cell lung cancer survival. Scientific Reports, 13(1).

Fu, M., Xu, J., Lv, Y., & Jin, B. (2025). Artificial intelligence in advanced gastric cancer: a comprehensive review of applications in precision oncology [Review of Artificial intelligence in advanced gastric cancer: a comprehensive review of applications in precision oncology]. Frontiers in Oncology, 15. Frontiers Media.

Gao, Q., Wu, H., Duan, L., Lv, G., & Quan, D. (2025). Multi-omics analysis of the tumor microenvironment after metastasis: advancing toward personalized immunotherapy and molecular targeted strategies [Review of Multi-omics analysis of the tumor microenvironment after metastasis: advancing toward personalized immunotherapy and molecular targeted strategies]. Frontiers in Immunology, 16. Frontiers Media.

Günel, T., & Kara, K. (2026). Translational Bioinformatics and Multi-Omics Integration for AI-Driven Biomarker Discovery in Precision Oncology. Acta Infologica, 1–20.

Hatamikia, S., Nougaret, S., Panico, C., Avesani, G., Nero, C., Boldrini, L., Sala, E., & Woitek, R. (2023). Ovarian cancer beyond imaging: integration of AI and multiomics biomarkers [Review of Ovarian cancer beyond imaging: integration of AI and multiomics biomarkers]. European Radiology Experimental,

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

7(1). Springer Science+Business Media.

Hira, M. T., Razzaque, M. A., Angione, C., Scrivens, J. H., Sawan, S., & Sarker, M. H. (2021). Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Scientific Reports, 11(1).

Hsu, C., Askar, S., Alshkarchy, S. S., Nayak, P. P., Attabi, K. A., Khan, M. A., Mayan, J. A., Sharma, M. K., Islomov, S., & Samarkhazan, H. S. (2025). AI-driven multi-omics integration in precision oncology: bridging the data deluge to clinical decisions. Clinical and Experimental Medicine, 26(1), 29–29.

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods [Review of More Is Better: Recent Progress in Multi-Omics Data Integration Methods]. Frontiers in Genetics, 8. Frontiers Media.

Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y., & Deng, H. (2023). AUTOSurv: Interpretable Deep Learning Framework for Cancer Survival Analysis Incorporating Clinical and Multi-omics Data. Research Square (Research Square).

Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y., & Deng, H. (2024). Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. Npj Precision Oncology, 8(1).

Kang, W., Qiu, X., Luo, Y., Luo, J., Liu, Y., Xi, J., Li, X., & Yang, Z. (2023). Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis [Review of Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis]. Journal of Translational Medicine, 21(1). BioMed Central.

Koh, D., Papanikolaou, N., Bick, U., Illing, R., Kahn, C. E., Kalpathi-Cramer, J., Matos, C., Martí-Bonmatí, L., Miles, A., Mun, S. K., Napel, S., Rockall, A., Sala, E., Strickland, N. H., & Prior, F. (2022). Artificial intelligence and machine learning in cancer imaging [Review of Artificial intelligence and machine learning in cancer imaging].

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Communications Medicine, 2(1). Nature Portfolio.

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., Wang, M., Zhang, Z., He, S., & Bo, X. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biology, 23(1).

Li, G., Feng, P., Lin, Y., & Liang, P. (2025). Integrating radiomics, artificial intelligence, and molecular signatures in bone and soft tissue tumors: advances in diagnosis and prognostication [Review of Integrating radiomics, artificial intelligence, and molecular signatures in bone and soft tissue tumors: advances in diagnosis and prognostication]. Frontiers in Oncology, 15. Frontiers Media.

Li, T., Mitchel, J., Chatlin, K., & Wang, M. D. (2020). Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Medical Informatics and Decision Making, 20(1).

Li, Y., Xin, W., Fang, D., & Luo, Y. (2024). Informing immunotherapy with multi-omics driven machine learning. Npj Digital Medicine, 7(1).

Liu, Z., Zhu, F., & Zhang, M. (2025). Artificial intelligence–driven analysis of antibody and nucleic acid biomarkers for enhanced disease diagnostics. Frontiers in Immunology, 16.

Ma, Y., Khan, A., Heng, Y., Khan, F. G., Aldhahri, A., & Khan, I. A. (2025). A contrastive adversarial encoder for multi-omics data integration. PLoS ONE, 20(10).

Mathema, V. B., Sen, P., Lamichhane, S., Orešič, M., & Khoomrung, S. (2023). Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. Computational and Structural Biotechnology Journal, 21, 1372–1382.

Menon, T. P., Mahajan, A., & Powell, D. (2026). Foundation model embeddings for multimodal oncology data integration. Npj Digital Medicine, 9(1), 131–131.

Mohetaer, A., & Adili, D. (2026). A commentary on "Multimodal radiopathomics approach for predictions of prognosis and immunotherapy response in patients

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

with gastric cancer: a multicohort retrospective study." International Journal of Surgery.

Muneer, A., Waqas, M., Saad, M. B., Showkatian, E., Bandyopadhyay, R., Xu, H., Li, W., Chang, J., Liao, Z., Haymaker, C., Soto, L. S., Wu, C. C., Vokes, N., Le, X., Byers, L., Gibbons, D. L., Heymach, J. V., Zhang, J., & Wu, J. (2026). From classical machine learning to emerging foundation models: review on multimodal data integration for cancer research. Artificial Intelligence Review, 59(4).

Muneer, A., Waqas, M., Saad, M. B., Showkatian, E., Bandyopadhyay, R., Xu, H., Li, W., Chang, J. Y., Liao, Z., Haymaker, C., Soto, L. S., Wu, C. C., Vokes, N. I., Le, X., Byers, L. A., Gibbons, D. L., Heymach, J. V., Zhang, J., & Wu, J. (2025). From Classical Machine Learning to Emerging Foundation Models: Review on Multimodal Data Integration for Cancer Research. PubMed.

Munquad, S., & Das, A. B. (2023). DeepAutoGlioma: a deep learning autoencoder-based multi-omics data integration and classification tools for glioma subtyping. BioData Mining, 16(1).

Mustafa, H. (2025). Beyond the Hype: Structuring the Future of Applied Multi-Omics in Oncology.

Muzafar, M., Riaz, S., Gurganari, L., Usman, S., & Ghafoor, K. (2026). Cancer Bioinformatics: Integrating Multi-Omics Data for Precision Oncology. Physical Education Health and Social Sciences, 4(1), 1–11.

Nam, Y., Kim, J., Jung, S., Woerner, J., Suh, E., Lee, D., Shivakumar, M., Lee, M. E., & Kim, D. (2024). Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine. PubMed Central, 7(1), 225–250.

Ng, S., Masarone, S., Watson, D., & Barnes, M. R. (2023). The benefits and pitfalls of machine learning for biomarker discovery [Review of The benefits and pitfalls of machine learning for biomarker discovery]. Cell and Tissue Research, 394(1), 17–31. Springer Science+Business Media.

Nguyen, R., & Vafaee, F. (2025). Multi-omics prognostic marker discovery

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

and survival modelling: a case study on multi-cancer survival analysis of women's specific tumours. Scientific Reports, 15(1).

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools [Review of Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools]. Frontiers in Oncology, 10. Frontiers Media.

Patel, S. K., George, B., & Rai, V. (2020). Artificial Intelligence to Decode Cancer Mechanism: Beyond Patient Stratification for Precision Oncology [Review of Artificial Intelligence to Decode Cancer Mechanism: Beyond Patient Stratification for Precision Oncology]. Frontiers in Pharmacology, 11. Frontiers Media.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S., Reina, G. A., Foley, P., Груздев, А. Д., Karkada, D., Davatzikos, C., Sako, C., Ghodasara, S., Bilello, M., Mohan, S., Kickingereder, P., Brugnara, G., Preetha, C. J., Sahm, F., Maier-Hein, K., … Metz, M. (2022). Federated learning enables big data for rare cancer boundary detection. Nature Communications, 13(1).

Paverd, H., Zormpas-Petridis, K., Clayton, H., Burge, S., & Crispin-Ortuzar, M. (2024). Radiology and multi-scale data integration for precision oncology [Review of Radiology and multi-scale data integration for precision oncology]. Npj Precision Oncology, 8(1). Nature Portfolio.

Qoku, A., Katsaouni, N., Flinner, N., Buettner, F., & Schulz, M. H. (2023). Multimodal analysis methods in predictive biomedicine [Review of Multimodal analysis methods in predictive biomedicine]. Computational and Structural Biotechnology Journal, 21, 5829–5838. Elsevier BV.

Quiros, A. C., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D. A., Park, C. Y., Pass, H. I., Moreira, A. L., Quesne, J. L., Tsirigos, A., & Yuan, K. (2024). Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. Nature Communications, 15(1).

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Reddy, C., Kaza, V. S., Daduvy, A., Shuaib, M., Alshanketi, F., & Alam, S. (2025). MoX: an explainable hybrid deep-learning model for integrating multi-omics data to predict event-free survival in neuroblastoma prognosis. Journal of King Saud University - Computer and Information Sciences, 37(9).

Sapoval, N., Aghazadeh, A., Nute, M., Antunes, D. A., Balaji, A., Baraniuk, R. G., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C., Yan, Z., Yao, V., & Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences [Review of Current progress and open challenges for applying deep learning across the biosciences]. Nature Communications, 13(1), 1728–1728. Nature Portfolio.

Shi, Y., Huang, Q., Lyu, J., Dong, T., & Sun, J. (2025). Progress of MRI based radiomics and deep learning for predicting the prognosis of locally advanced rectal cancer (Review) [Review of Progress of MRI based radiomics and deep learning for predicting the prognosis of locally advanced rectal cancer (Review)]. Oncology Letters, 30(5), 1–13. Spandidos Publishing.

Shoaib, A. S. M., Nishat, N., Raasetti, M., & Arif, I. (2024). INTEGRATIVE MACHINE LEARNING APPROACHES FOR MULTI-OMICS DATA ANALYSIS IN CANCER RESEARCH. Deleted Journal, 1(2), 26–39.

Shui, L., Ren, H., Yang, X., Li, J., Chen, Z., Cheng, Y., Zhu, H., & Shui, P. (2021). The Era of Radiogenomics in Precision Medicine: An Emerging Approach to Support Diagnosis, Treatment Decisions, and Prognostication in Oncology [Review of The Era of Radiogenomics in Precision Medicine: An Emerging Approach to Support Diagnosis, Treatment Decisions, and Prognostication in Oncology]. Frontiers in Oncology, 10. Frontiers Media.

Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andrés-Terré, H., Shams, Z., Jamnik, M., & Lió, P. (2019). Variational Autoencoders for Cancer Data Integration: Design

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Principles and Computational Practice. Frontiers in Genetics, 10.

Steyaert, S., Qiu, Y. L., Zheng, Y., Mukherjee, P., Vogel, H., & Gevaert, O. (2023). Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. Communications Medicine, 3(1).

Tonekaboni, S., Friedman, S., Zhang, X., Maddah, M., & Uhler, C. (2025). A Representation Fusion Framework for Decoupling Diagnostic Information in Multimodal Learning. Npj Digital Medicine, 8(1), 765–765.

Tran, K., Kondrashova, O., Bradley, A. P., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection [Review of Deep learning in cancer diagnosis, prognosis and treatment selection]. Genome Medicine, 13(1). BioMed Central.

Vale-Silva, L. A., & Rohr, K. (2021). Long-term cancer survival prediction using multimodal deep learning. Scientific Reports, 11(1).

Valous, N. A., Popp, F., Zörnig, I., Jäger, D., & Charoentong, P. (2024). Graph machine learning for integrated multi-omics analysis [Review of Graph machine learning for integrated multi-omics analysis]. British Journal of Cancer, 131(2), 205–211. Springer Nature.

Vanguri, R., Luo, J., Aukerman, A., Egger, J. V., Fong, C. J., Horvat, N., Pagano, A., Araújo-Filho, J. de A. B., Geneslaw, L., Rizvi, H., Sosa, R. E., Boehm, K., Yang, S., Bodd, F. M., Ventura, K., Hollmann, T. J., Ginsberg, M. S., Gao, J., Vanguri, R., … Shah, S. P. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nature Cancer, 3(10), 1151–1164.

Wang, F., Zhuang, Z., Gao, F., He, R., Zhang, S., Wang, L., Liu, J., & Li, Y. (2024). TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. Genome Biology, 25(1).

Wang, J., Chen, S., Wen, R., Ying, H., Xu, W., Yin, L., Deng, X., Han, C., Zhu, Q., Zhang, B., Tong, H., Liu, C., Chen, W., Jin, J., Jin, K., Xu, C., Huang, H., Wang, H., & Qian, D. (2026). BPX-Net: biomarker-preserved explainable networks for

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

disease diagnosis and prognosis. BioData Mining.

Waqas, A., Tripathi, A., Ramachandran, R. P., Stewart, P. A., & Rasool, G. (2024). Multimodal data integration for oncology in the era of deep neural networks: a review. Frontiers in Artificial Intelligence, 7, 1408843–1408843.

Wei, L., Niraula, D., Gates, E., Fu, J., Luo, Y., Nyflot, M. J., Bowen, S. R., Naqa, I. M. E., & Cui, S. (2023). Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. British Journal of Radiology, 96(1150), 20230211–20230211.

Wysocka, M., Wysocki, O., Zufferey, M., Landers, D., & Freitas, A. (2023). A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data [Review of A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data]. BMC Bioinformatics, 24(1). BioMed Central.

Xiao, Y., Yang, S., He, M., Chen, L., Wu, Y., & Zhong, L. (2026). UroFusion-X: a unified multimodal deep learning framework for robust diagnosis, subtyping, and prognosis of urological cancers. Npj Digital Medicine, 9(1), 117–117.

Xu, J., Yang, P., Xue, S., Sharma, B., Sánchez-Martín, M., Wang, F., Beaty, K., Dehan, E., & Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives [Review of Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives]. Human Genetics, 138(2), 109–124. Springer Science+Business Media.

Xu, Y., Li, Y., Wang, D., Zhang, Y., & Huang, D. (2025). Addressing the current challenges in the clinical application of AI-based Radiomics for cancer imaging. Frontiers in Medicine, 12.

Yao, X., Jiang, X., Luo, H., Liang, H., Ye, X., Wei, Y., & Cong, S. (2024). MOCAT: multi-omics integration with auxiliary classifiers enhanced

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

autoencoder. BioData Mining, 17(1).

Yuan, X., Xu, H.-L., Zhu, J., Yang, Z., Pan, B., Wu, L., & Chen, H. (2025). Systematic review and meta-analysis of artificial intelligence for image-based lung cancer classification and prognostic evaluation. Npj Precision Oncology, 9(1).

Zafar, S., Hafeez, A., Shah, H., Mutiullah, I., Ali, A., Khan, K., Figueroa-González, G., Reyes-Hernández, O. D., Quintas-Granados, L. I., Peña-Corona, S. I., Kiyekbayeva, L., Butnariu, M., Tota, C.-E., Caunii, A., Büsselberg, D., Sharifi-Rad, J., & Leyva-Gómez, G. (2025). Emerging biomarkers for early cancer detection and diagnosis: challenges, innovations, and clinical perspectives [Review of Emerging biomarkers for early cancer detection and diagnosis: challenges, innovations, and clinical perspectives]. European Journal of Medical Research, 30(1). BioMed Central.

Zhang, C., Xu, J., Tang, R., Yang, J., Wang, W., Yu, X., & Shi, S. (2023). Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment [Review of Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment]. Journal of Hematology & Oncology, 16(1). BioMed Central.

Zheng, S., Cui, X., & Ye, Z. (2025). Integrating artificial intelligence into radiological cancer imaging: from diagnosis and treatment response to prognosis. Cancer Biology and Medicine, 22(1), 6–13.

Zhou, H.-J., Zhou, F., Zhao, C., Xu, Y., Luo, L., & Chen, H. (2024). Multimodal Data Integration for Precision Oncology: Challenges and Future Directions. arXiv (Cornell University).

Zhu, F., Liu, Z., Chang, J., Qin, Y., & Wang, L. (2025). Deep learning for scene understanding in mitochondrial dysregulation and blood cancer diagnosis. Frontiers in Oncology, 15.

Zhuang, L., Park, S., Skates, S. J., Prosper, A. E., Aberle, D. R., & Hsu, W. (2025). Advancing Precision Oncology Through Modeling of Longitudinal and Multimodal Data. IEEE Reviews in Biomedical Engineering, 1–19.

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.
EXPEDITION RESEARCH & EDUCATION HUB (SMC-PRIVATE) LIMITED

Biosciences
reports

Copyright©2026. This work is licensed under a Creative Common Attribution 4.0 International License.