Review Article

A Review of Artificial Intelligence Techniques for Biomarker Discovery

Surbhi Rani

Department of Computer Science, University of Delhi, Delhi, India
Correspondence: srani@cs.du.ac.in
ORCiD: https://orcid.org/0009-0004-3237-4688

Received:*2nd October, 2025;* ***Accepted***: *29th October, 2025;* ***Published***: *31st October, 2025*

Vantage: Journal of Thematic Analysis

A Peer Reviewed Multidisciplinary Publication of Centre for Research, Maitreyi College, University of Delhi. Volume 6, Issue 2, October 2025. *https://vantagejournal.com* , ISSN(E): 2582-7391

How to cite:

Surbhi Rani. (2025). A Review of Artificial Intelligence Techniques for Biomarker Discovery. *Vantage: Journal of Thematic Analysis*, 6(2), 83-92. https://doi.org/10.52253/vjta.2025.v06i02.83

ABSTRACT

Biomarkers play a crucial role in medical research, helping guide diagnosis, prognosis, and treatment decisions. Traditionally, discovering these biomarkers has been a labor-intensive process, relying on experiments and statistical analysis that often capture only a small part of biological complexity. The introduction of Artificial Intelligence has brought a significant shift in this field. Machine learning approaches first allowed researchers to uncover patterns in genomic and clinical data that were previously difficult to recognize, and later, deep learning expanded these possibilities to include multi-omics data and medical imaging. Recently, explainable Artificial Intelligence techniques have addressed the challenge of trust and interpretability, enabling clinicians to understand and validate Artificial Intelligence-derived biomarkers. This review presents the journey of biomarker discovery with Artificial Intelligence, highlighting key developments, current trends, and challenges, while discussing the potential for Artificial Intelligence to enhance the clinical impact of biomarker research.

Keywords: Biomarker Discovery, Artificial Intelligence, Machine learning, Deep Learning, Explainable Artificial Intelligence

1. INTRODUCTION

Biomarkers are measurable indicators of biological processes, diseases, or responses to treatment, playing a crucial role in modern medicine. They assist clinicians in predicting disease risk, monitoring treatment efficacy, and personalizing patient care (Strimbu & Tavel, 2010). Traditionally biomarker discovery depended on clinical research, laboratory experiments, and statistical methods. These approaches offered important insights, but they often required substantial time, resources, and expertise while missing subtle patterns in complex biological data. The increasing complexity of biological systems together with the rapid growth of biomedical data ranging from genomic sequences to high-resolution imaging has made conventional methods less effective (Rifai et al., 2006). These limitations necessitate computational approaches capable of analyzing large-scale, multi-dimensional datasets to uncover hidden relationships.

Artificial Intelligence (AI) has emerged as a powerful tool in this domain. Early applications of Machine Learning (ML), such as Support Vector Machines (SVM), decision trees, and logistic regression, enabled the analysis of larger datasets,

identified patterns, and ranked potential biomarkers with greater efficiency compared to traditional approaches (Libbrecht & Noble, 2015). These methods often depended on careful feature selection and domain knowledge, and the results were at times difficult to interpret in a biological context. The development of Deep Learning (DL) has further advanced biomarker discovery. Models such as Convolutional Neural Networks (CNN) and transformers can automatically extract hierarchical features from complex data, including multi-omics datasets and medical images, enabling the identification of new and clinically meaningful biomarkers (Angermueller et al., 2016; Min et al., 2017). The integration of genomics, proteomics, transcriptomics, and metabolomics has contributed to a more holistic and deep understanding of disease mechanisms.

Despite these advances, the black box nature of DL models has limited their clinical adoption. Explainable AI (XAI) methods such as SHAP, LIME, and Grad CAM help overcome this limitation by offering transparent and interpretable insights (Selvaraju et al., 2017). These tools help identify genes, proteins, or pathways critical for distinguishing disease states, improving both the reliability and clinical relevance of AI-discovered biomarkers (Holzinger et al., 2019). Today, the combination of AI and XAI is driving a new phase of biomarker research, supporting faster and more reliable discoveries for personalized medicine.

2. RELATED STUDY

2.1 Early machine learning-based approaches in biomarker discovery

Before the advent of DL and multi-omics integration, biomarker discovery largely relied on classical ML techniques. These methods provided the first computational means to analyze high-dimensional biological datasets, such as gene expression profiles, proteomics, and clinical data, enabling the identification of potential disease-associated biomarkers. Cox (1972) developed the proportional hazards model, a classical survival analysis tool. Later, ML-based adaptations integrated Cox regression with feature selection to identify prognostic biomarkers from high-dimensional omics data. These models became standard for linking molecular features to patient survival outcomes. Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (LASSO), which became a popular method for biomarker selection in high-dimensional omics data. By incorporating regularization, LASSO allowed both prediction and feature selection at the same time, making it well suited for identifying sparse biomarker signatures from thousands of candidate features. Its impact persists in current multi-omics studies and survival analysis workflows.

Cover & Hart, (1967) proposed the k-NN algorithm, which later found application in biomarker discovery through gene expression data analysis. Early cancer studies demonstrated that k-NN could classify tumor subtypes by identifying patterns of similarity in gene expression profiles. Freund & Schapire (1997) represented the AdaBoost algorithm, which later proved useful for biomarker classification tasks. In cancer genomics, boosting ensembles improved accuracy over single learners by combining weak classifiers. This research work highlighted the significance of ensemble approaches for improving reliability in biomarker discovery. Breiman et al., (2017) employed classification and regression trees, which later became important in biomarker studies. Decision tree allowed straightforward interpretation of gene expression data, helping identify biomarkers with clear threshold-based decision rules. Guyon et al., (2002) introduced SVM-RFE method for iterative feature elimination. Applied to cancer datasets, it provided ranked gene lists and improved biomarker selection stability. Díaz-Uriarte & Alvarez de Andrés, (2006) introduced RF as a robust method for biomarker identification and gene expression-based classification. They demonstrated its ability to manage high-dimensional data with noise and generate contribution scores that assist the selection process of biomarkers. The summarization of ML-based research works is tabulated in Table 1.

2.2 Deep learning based approaches in biomarker discovery

DL has emerged as a transformative approach for biomarker discovery with the rise of large-scale biomedical datasets. In contrast to traditional ML methods, DL models are capable of automatically learning hierarchical representations and capturing complex non-linear relationships within data (Bedi et al., 2025). These capabilities have enabled the identification of more robust and clinically meaningful biomarkers, enhancing applications in cancer diagnosis, prognosis and prediction of therapeutic responses.

Chaudhary et al., (2018) integrated multi-omics data (RNA-seq, miRNA, and methylation) from The Cancer Genome Atlas hepatocellular cancer cohort using a deep autoencoder framework. Their method revealed prognostic subgroups and uncovered novel

Table 1: A summary of ML-based research works for biomarker discovery of breast cancer

Author (Year)	Method	Data Type	Key Contribution	Outcome
Cover & Hart, (1967)	k-Nearest Neighbours	Gene expression	Early proximity-based classification of tumour subtypes	Showed simple similarity learning works for biomarker tasks
Cox, (1972)	Cox Proportional Hazards Model	Clinical + omics survival data	Linked biomarkers to patient outcomes	Standard for prognostic biomarker identification
Freund & Schapire, (1997)	AdaBoost (Ensemble)	Cancer genomics	Boosting improved weak learner performance	Higher accuracy and robustness in biomarker discovery
Tibshirani, (1996)	LASSO Regression	Omics datasets	Regularization for simultaneous prediction and feature selection	Sparse biomarker signatures, widely adopted in omics
Hosmer & Lemeshow, (2000)	Logistic Regression	Genomic & clinical data	Widely applied to binary classification (disease vs. control)	Interpretable links between biomarkers and disease risk
Friedman et al., (2000)	Bayesian Networks	Gene expression	Modelled probabilistic gene dependencies	Revealed biomarker networks and pathway interactions
Guyon et al., (2002)	SVM	Gene expression (microarray)	Demonstrated SVM for cancer classification and biomarker selection	Identified informative gene subsets for cancer subtypes
Breiman et al., (2017)	Decision tree	Gene expression	Threshold-based rules for classification	Transparent biomarker selection and subgroup classification

biomarkers that were significantly associated with overall survival, highlighting the effectiveness of unsupervised feature learning in heterogeneous datasets. Hao et al., (2019) carried out a pan-cancer study using deep neural networks on transcriptomic datasets spanning multiple tumor types. Their model identified cross cancer biomarkers genes consistently dysregulated across cancers providing insights into shared molecular mechanisms and suggesting candidates for broad-spectrum cancer diagnostics. Verma et al., (2021) developed a two-stage DL pipeline that segments high-cellularity tumor regions from H&E-stained slides and predicts overall survival in glioblastoma patients. Using multi-cohort datasets, the model achieved high accuracy in tumour region identification and robust prognostic performance across independent validation cohorts. Qiu et al., (2022) employed a CNN to analyze lung adenocarcinoma gene expression data. The model not only achieved high accuracy in patient classification but also pinpointed genes such as MMP11 and COL1A1 as biomarkers with potential prognostic relevance, highlighting the interpretability of CNN filters for biomarker discovery. Zhang & Wang, (2019) explored multimodal integration by combining histopathology images with transcriptomic data using CNNs. Applied to glioblastoma multiforme, their framework revealed biomarkers that connected morphological tissue features with gene expression signatures,

Table 2: A summary of DL-based research work for biomarker discovery of breast cancer

Author (Year)	Method	Data Type	Key Contribution	Outcome
Alipanahi et al., (2015)	Convolutional deep networks (Deep Bind)	DNA/RNA binding assays (sequence)	Learned sequence motifs / predicted protein nucleic acid binding specificities useful for regulatory biomarker discovery.	Outperformed traditional motif discovery methods in predicting binding specificity.
Tan et al., (2015)	Autoencoder + reverse-learning feature extraction	Gene expression (RNA-Seq)	Autoencoder-based pipeline to extract and rank gene biomarkers combined with feature elimination for compact signatures.	Produced compact, high-performing biomarker signatures validated across datasets.
Miotto et al., (2016)	Unsupervised deep representation (stacked autoencoders) (DeepPatient)	Electronic Health Records (EHR)	Learned patient representations that improved prediction of future disease a foundation for EHR-derived biomarkers.	Generated robust features predictive of multiple disease onsets.
Angermueller et al., (2017)	CNN + recurrent modules (DeepCpG)	Single-cell DNA methylation	Predicted single-cell CpG methylation states; enabled methylation-based biomarker inference.	Improved methylation state prediction at single-cell resolution.
Esteva et al., (2017)	CNN (transfer learning)	Dermatology images (clinical & dermoscopic)	Achieved dermatologist-level skin-lesion classification image features usable as diagnostic biomarkers.	Matched dermatologist performance in skin cancer detection.
De Fauw et al., (2018)	CNN (transfer learning)	Retinal OCT images	Automated diagnosis of treatable retinal diseases image-derived biomarkers for screening.	Enabled rapid and accurate screening of retinal diseases in clinical practice.
Coudray et al., (2018)	Inception-v3 CNN	Histopathology whole-slide images (lung)	Predicted cancer subtype and common driver mutations from H&E images morphological biomarkers linked to genetics.	Identified both histology subtypes and genetic mutations with high accuracy.
Chaudhary et al., (2018)	Fully connected deep networks for multi-omics integration	Multi-omics (mRNA, miRNA, methylation) HCC (liver)	Integrated omics to predict survival and subtypes; identified multi-omic signatures prognostic for HCC.	Stratified patients into survival groups with significant clinical relevance.

Katzman et al., (2018)	DeepSurv (deep Cox network)	Clinical + molecular covariates (survival data)	DL survival model that identifies prognostic features/risk biomarkers with better discrimination.	Improved survival prediction compared to standard Cox models.
Zhang et al., (2022)	Graph-based autoencoder (multi-modal)	Spatial transcriptomics + chromatin images	Integrated spatial multi-modal data to identify joint spatial molecular biomarkers.	Revealed spatially resolved biomarkers relevant to Alzheimer's pathology.

demonstrating the promise of integrative DL approaches.

Lin & Wang (2020) focused on breast cancer and proposed omics-specific encoders within a DL framework to separately process gene expression and methylation data before integration. This modular design enabled the discovery of subtype-specific biomarkers, such as methylation-driven CpG sites, with high clinical significance in differentiating Luminal and HER2-positive subtypes. Lee et al. (2020) developed a multi-omics DL model for colorectal cancer, integrating DNA methylation, mRNA, and miRNA profiles. Their system achieved improved survival prediction and highlighted novel biomarkers, particularly in Wnt signaling pathways, reinforcing the biological interpretability of deep networks. Guo et al. (2020) applied denoising autoencoders combined with L1 logistic regression on lung cancer datasets. Their approach enhanced noise robustness in high-dimensional gene expression data and uncovered survival-associated biomarkers, particularly in immune-related pathways, validating their relevance through biological literature. Bote-Curiel et al. (2021) designed an approach combining autoencoder embeddings with multivariate feature selection for ovarian cancer. By compressing high-dimensional transcriptomic data into informative latent features, the method improved patient stratification and uncovered gene signatures associated with chemotherapy response. Chen et al., (2022) introduced transformer-based architectures to gastric cancer biomarker discovery. Unlike traditional DL models, transformers captured long-range dependencies in omics data, leading to the identification of genes such as CLDN18 and MUC6 as prognostic markers. This study marked one of the first uses of attention-based models in this domain. The summarization of DL-based research works is tabulated in Table 2.

2.3 XAI-based approaches in biomarker discovery

In recent years, the predictions generated by AI-based models have been increasingly supported by explainable artificial intelligence (XAI) techniques. XAI provides interpretability by elucidating the reasoning behind model predictions, thereby enhancing the confidence of clinical practitioners in AI-driven decisions. Several researchers have utilized XAI frameworks for biomarker discovery across different cancer types. For instance, Dwivedi et al. (2023) developed a deep learning framework comprising an autoencoder and a feed-forward neural network to distinguish between lung adenocarcinoma and squamous-cell carcinoma subtypes. They identified 52 potential biomarkers for lung cancer, including seven novel candidates that had not been previously reported, and achieved an accuracy of 95.7% with their proposed model. Another study by Colak et al., 2025 introduced a methodology combining random forest and Light GBM models, achieving an AUC of 98%, followed by SHAP-based interpretation to highlight the most discriminatory metabolites. This work reflects the growing emphasis on integrating high-dimensional omics data with interpretable machine learning approaches to identify meaningful biomarkers rather than solely optimizing predictive performance. Furthermore, Withnell et al. (2021) proposed a variational autoencoder (VAE)-based deep learning architecture to uncover the contributions of individual genes and latent dimensions to classification outcomes. The authors demonstrated that the model's predictions were consistent with existing biological evidence, thus supporting the validity of their findings. In addition, Li et al. (2025) introduced a multi-view graph-level representation learning (MGRL) framework that integrates large-scale single-cell transcriptomic and epigenetic (DNA methylation) data with built-in explainability through XAI methods, providing deeper insights into underlying molecular mechanisms.

3. COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING APPROACHES

The comparison between traditional ML and DL highlights the distinct strengths and limitations of each paradigm in biomarker discovery. ML methods such as SVM, RF, and logistic regression have long been preferred due to their interpretability, lower data requirements, and computational efficiency (Breiman et al., 2017; Guyon et al., 2002). These approaches are especially effective when applied to structured data like gene expression or clinical variables, where transparent feature selection and ranking are essential. On the other hand, DL approaches such as CNNs, autoencoders, and graph-based architectures excel in handling large-scale, heterogeneous datasets (Alipanahi et al., 2015; Chaudhary et al., 2018). By learning complex non-linear representations directly from raw data, DL methods often achieve superior predictive accuracy, particularly in multi-omics integration and imaging. However, this performance comes at the cost of increased complexity, high data requirements, and limited interpretability compared to traditional ML.

When examining specific studies, the strengths of ML are evident in early gene expression-based biomarker research. For instance, SVM was successfully applied to cancer classification tasks, identifying informative gene subsets for distinguishing subtypes (Guyon et al., 2002). RF was developed to manage noisy, high-dimensional data and has proven useful for ranking variable importance in biomarker identification (Breiman et al., 2017). Similarly, LASSO regression provided sparse and interpretable biomarker signatures in omics datasets (Tibshirani, 1996). While SVM-RFE delivered stable ranked gene lists for cancer biomarkers (Guyon et al., 2002), logistic regression offered clear associations between genetic or clinical biomarkers and disease risk (Hosmer & Lemeshow, 2000), and Cox proportional hazards models linked biomarker signatures with patient outcomes in survival analysis (Cox, 1972). Other approaches such as decision tree and Bayesian networks, modeled subgroup classifications, and probabilistic gene dependencies, respectively (Breiman et al., 2017). These studies collectively demonstrated that ML could provide reliable and interpretable biomarker signatures, particularly in situations with limited sample sizes and where computational efficiency was important.

In contrast, DL studies have demonstrated remarkable progress in expanding the scope and power of biomarker discovery. Alipanahi et al. (2015) used CNNs in DeepBind to predict protein nucleic acid binding specificities, outperforming traditional motif discovery methods. Angermueller et al. (2016) developed DeepCpG, which improved the prediction of single-cell methylation states. In clinical imaging, Esteva et al. (2017) achieved dermatologist-level skin cancer detection using CNNs, while Coudray et al. (2018) identified both histological subtypes and genetic mutations directly from histopathology slides. DL has also shown unique strength in multi-omics integration. Chaudhary et al., (2018) applied deep networks to combine mRNA, miRNA, and methylation data for liver cancer prognosis, significantly stratifying patient survival groups. Even in survival analysis, Katzman et al., (2018) demonstrated that DeepSurv outperformed the standard Cox model. More recent advances, such as Al Abir Fuad et al. (2022) and Chen et al. (2022) further highlight DL's potential in extracting compact gene signatures and integrating spatial multi-modal data, respectively.

Taken together, ML studies emphasize interpretability, feature stability, and robustness under limited data conditions, while DL studies highlight the capacity for automated representation learning, higher predictive accuracy, and integration of diverse biomedical modalities. For example, whereas ML approaches such as SVM-RFE produced stable ranked gene lists (Guyon et al., 2002), DL-based autoencoders extracted compact but highly predictive signatures (Al Abir Fuad et al., 2022). Similarly, while Cox proportional hazards models established the benchmark for linking biomarkers to outcomes (Cox, 1972) DL-based survival models like DeepSurv provided more nuanced stratification (Katzman et al., 2018). The transition from manually designed, interpretable ML features to end-to-end DL models reflects the evolution of biomarker discovery, with each approach providing complementary advantages.

4. CHALLENGES AND OPPORTUNITIES IN BIOMARKER DISCOVERY

Despite the remarkable progress of ML and DL in biomarker discovery, several challenges remain that limit their clinical translation (Javaid et al., 2025). One of the primary issues is data scarcity and heterogeneity. High-throughput omics technologies produce large-scale datasets, but labeled clinical data with long-term follow-up remain limited (Ahmed et al., 2024). In addition, multi-center studies often face batch effects and inconsistent preprocessing, which reduce reproducibility across datasets. Another key

challenge is the complexity and interpretability of models. Although DL architectures can achieve high predictive accuracy, their black box nature makes it difficult to understand how individual biomarkers influence predictions (Mohammed et al., 2023). This lack of transparency limits their adoption in clinical settings, where trust and explainability are essential. ML approaches, while more interpretable, often encounter difficulties in scaling to the high-dimensional nature of omics data.

Reproducibility remains a major challenge in biomarker research. Many studies report promising findings on small cohorts but fail to generalize to independent datasets. Issues such as overfitting, inadequate validation strategies, and the absence of standardized pipelines further compound this problem (Kalesinskas et al., 2022). Translating computational biomarkers into clinical practice also requires extensive validation, regulatory approval and seamless integration into clinical workflows, which are often neglected during model development. At the same time, these challenges present new opportunities. Multi-omics integration has emerged as an effective approach to capture complementary biological signals that may be overlooked by single-omics analyses. Likewise, transfer learning and federated learning provide solutions for data scarcity by leveraging information from related domains and enabling collaborative model training without sharing sensitive data. XAI methods including SHAP, LIME and attention-based visualization are increasingly used to improve interpretability and foster clinical trust. Furthermore, combining ML and DL based biomarkers with precision medicine frameworks offers the potential for more accurate diagnosis, prognosis and treatment stratification.

5. FUTURE DIRECTION

The field of biomarker discovery is poised for a transformative shift driven by advancements in AI, data integration, and clinical translation. A central priority is the creation of standardized, large-scale multi-omics datasets with consistent pre-processing and annotation. Such datasets will improve reproducibility and enable robust benchmarking of ML and DL models across studies and disease cohorts. Another promising avenue is combining explainability with performance. Future DL models will need to balance predictive accuracy with interpretable outputs that clinicians can rely on. The increasing adoption of explainable DL frameworks and visualization techniques will be crucial in connecting algorithmic predictions with biological understanding.

To advance the research of biomarkers requires a collaboration among computational scientists, biologists, and clinicians. Building cohesive pipelines that connect computational predictions with experimental testing and clinical evaluation can greatly speed up the transition of discoveries from the lab to medical application. New technologies such as self-supervised learning, graph neural networks, and foundation models for omics data offer fresh possibilities for uncovering intricate biomarker relationships that traditional approaches often miss due to human error. At the same time, federated learning and privacy-preserving AI provide ways to collaborate across institutions while protecting sensitive patient data. The overarching aim is to develop biomarkers that are clinically meaningful, reproducible, and transparent, supporting precision medicine and improving outcomes for patients worldwide.

For discovering the biomarkers, AI becomes an influential tool that can process vast and complex biological datasets to reveal the molecular patterns that can link it to conditions like cancer. Though ML and DL have the potential to pinpoint the biomarkers (i.e., diagnostic and prognostic) with significant accuracy. However, it is not sufficient to validate the biomarkers; therefore, Polymerase Chain Reaction (PCR), Western blotting, or immunohistochemistry are crucial to verify the relevance of the discovered biomarkers. In conclusion, these biological validation tests and AI collaboration become a powerful, complementary approach. This approach helps in discovering the reliable biomarkers for speedy recovery and more effective treatment.

6. CONCLUSION

In the research of biomarker discovery, AI plays a vital role which offers immerse amount of advantages. The expertise of AI is highly relevant due to its interpretability, lower computational demands, and effectiveness with moderate sized datasets. AI provides clear insights into the importance of features, which is significant for clinical validation. AI excels in capturing the complex pattern and non-linear relationships in large scale multi-omics data. This enables end-to-end feature learning and the identification of hidden biomarker signatures that can be neglected by traditional approaches. Despite these strengths, there are various limitations, such as limited data, known as data scarcity, limited interpretability, and reproducibility concerns, which continue to hinder

the AI-based clinical applications. Therefore, to mitigate these issues, hybrid approaches that combine representation of AI alongside XAI techniques have been utilized. Rigorous validation across independent cohorts, transparent reporting, and the integration of multi-omics datasets will also be essential for translating computational biomarkers into clinical practice. In conclusion, AI should be seen as complementary rather than competing paradigms, and their combined evolution is expected to accelerate biomarker discovery and advance precision medicine.

Conflict of Interest

The author has no conflict of interest to declare that is relevant to the content of this article.

Funding

Not Applicable.

REFERENCES

Ahmed, Z., Wan, S., Zhang, F., & Zhong, W. (2024). Artificial intelligence for omics data analysis. *BMC Methods 2024 1:1*, *1*(1), 1–4. https://doi.org/10.1186/S44330-024-00004-5

Al Abir Fuad, Shovan S. M., Hasan Md. Al Mehedi, Sayeed Abu, & Shin Jungpil. (2022). Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination - Molecular Omics (RSC Publishing). *Molecular Omics*, *18*, 652–661. https://doi.org/DOI:10.1039/D1MO00467K

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Angermueller, C., Lee, H. J., Reik, W., & Stegle, O. (2017). DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, *18*(1), 1–13. https://doi.org/10.1186/S13059-017-1189-Z/FIGURES/5

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7). https://doi.org/10.15252/MSB.20156651

Bedi, P., Rani, S., Gupta, B., Bhasin, V., & Gole, P. (2025). EpiBrCan-Lite: A lightweight deep learning model for breast cancer subtype classification using epigenomic data. *Computer Methods and Programs in Biomedicine*, *260*, 108553. https://doi.org/10.1016/J.CMPB.2024.108553

Bote-Curiel, L., García, M., & Pérez, J. (2021). Ovarian cancer data analysis using deep learning. *ScienceDirect*.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. *Classification and Regression Trees*, 1–358. https://doi.org/10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-OLSHEN-CHARLES-STONE/RIGHTS-AND-PERMISSIONS

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *24*(6), 1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853

Chen, X., Zhang, Y., & Li, J. (2022). Transformer models for biomarker discovery in gastric cancer. *PMC Central*.

Colak, C., Yagin, F. H., Algarni, A., Algarni, A., Al-Hashem, F., & Ardigò, L. P. (2025). Proposed Comprehensive Methodology Integrated with Explainable Artificial Intelligence for Prediction of Possible Biomarkers in Metabolomics Panel of Plasma Samples for Breast Cancer Detection. *Medicina*, *61*(4), 581. https://doi.org/10.3390/MEDICINA61040581

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, *24*(10), 1559–1567. https://doi.org/10.1038/S41591-018-0177-5

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cox, D. R. (1972). Regression Models and Life-Tables in the Chair]. *Research Section, on Wednesday*. https://academic.oup.com/jrsssb/article/34/2/187/7027194

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350. https://doi.org/10.1038/S41591-018-0107-6

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 1–13. https://doi.org/10.1186/1471-2105-7-3/FIGURES/1

Dwivedi, K., Rajpal, A., Rajpal, S., Agarwal, M., Kumar, V., & Kumar, N. (2023). An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification. *Computers in Biology and Medicine*, *153*, 106544. https://doi.org/10.1016/J.COMPBIOMED.2023.106544

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/NATURE21056;SUBJMETA

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *7*(3–4), 601–620. https://doi.org/10.1089/106652700750050961

Guo, Y., Wang, L., & Zhang, H. (2020). Biomarker identification in lung cancer using autoencoders. *BMC Bioinformatics*, *21*, 1–9.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797/METRICS

Hao, J., Masum, M., Oh, J. H., & Kang, M. (2019). Gene- and Pathway-Based Deep Neural Network for Multi-omics Data Integration to Predict Cancer Survival Outcomes. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11490 LNBI*, 113–124. https://doi.org/10.1007/978-3-030-20242-2_10/TABLES/4

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312. https://doi.org/10.1002/WIDM.1312

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. *Applied Logistic Regression*. https://doi.org/10.1002/0471722146

Javaid, H., Petrescu, C. C., Schmunk, L. J., Monahan, J. M., O'Reilly, P., Garg, M., McGirr, L., Khasawneh, M. T., Al Lail, M., Ganta, D., Stubbs, T. M., Sun, B. B., Vitsios, D., & Martin-Herranz, D. E. (2025). The impact of artificial intelligence on biomarker discovery. *Emerging Topics in Life Sciences*, *8*(2), 89–105. https://doi.org/10.1042/ETLS20243003

Kalesinskas, L., Gupta, S., & Khatri, P. (2022). Increasing reproducibility, robustness, and generalizability of biomarker selection from meta-analysis using Bayesian methodology. *PLOS Computational Biology*, *18*(6), e1010260. https://doi.org/10.1371/JOURNAL.PCBI.1010260

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1), 1–12. https://doi.org/10.1186/S12874-018-0482-1/FIGURES/6

Lee, S., Kim, H., & Lee, J. (2020). Multi-omics deep learning for colorectal cancer biomarkers. *PMC Central*.

Li, Z. P., Du, Z., Huang, D. S., & Teschendorff, A. E. (2025). Interpretable deep learning of single-cell and epigenetic data reveals novel molecular insights in aging. *Scientific Reports*, *15*(1). https://doi.org/10.1038/S41598-025-89646-1

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews. Genetics*, *16*(6), 321–332. https://doi.org/10.1038/NRG3920

Lin, Y., & Wang, J. (2020). Classifying breast cancer subtypes using deep neural networks. *PMC Central*.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, *18*(5), 851–869. https://doi.org/10.1093/BIB/BBW068

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, *6*(1), 1–10. https://doi.org/10.1038/SREP26094;TECHMETA

Mohammed, M. A., Abdulkareem, K. H., Dinar, A. M., & Zapirain, B. G. (2023). Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review. *Diagnostics (Basel, Switzerland)*, *13*(4). https://doi.org/10.3390/DIAGNOSTICS13040664

Qiu, W. R., Qi, B. B., Lin, W. Z., Zhang, S. H., Yu, W. K., & Huang, S. F. (2022). Predicting the Lung Adenocarcinoma and Its Biomarkers by Integrating Gene Expression and DNA Methylation Data. *Frontiers in Genetics*, *13*, 926927. https://doi.org/10.3389/FGENE.2022.926927/BIBTEX

Rifai, N., Gillette, M. A., & Carr, S. A. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnology*, *24*(8), 971–983. https://doi.org/10.1038/NBT1235

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-October*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463–466. https://doi.org/10.1097/COH.0B013E32833ED177

Tan, J., Ung, M., Cheng, C., & Greene, C. S. (2015). Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pacific Symposium on Biocomputing*, 132–143. https://doi.org/10.1142/9789814644730_0014

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X

Verma, R., Cohen, M., Toro, P., Mokhtari, M., & Tiwari, P. (2021). NIMG-60. PREDICTING OVERALL SURVIVAL IN GLIOBLASTOMA USING HISTOPATHOLOGY VIA AN END-TO-END DEEP LEARNING PIPELINE: A LARGE MULTI-COHORT STUDY. *Neuro-Oncology*, *23*(Supplement_6), vi143–vi143. https://doi.org/10.1093/NEUONC/NOAB196.558

Withnell, E., Zhang, X., Sun, K., & Guo, Y. (2021). XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab315

Zhang, X., Wang, X., Shivashankar, G. V., & Uhler, C. (2022). Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. *Nature Communications*, *13*(1), 1–17. https://doi.org/10.1038/S41467-022-35233-1;TECHMETA

Zhang, X., & Wang, Z. (2019). Multi-modal deep learning identifies glioblastoma biomarkers. *PMC Central*.