IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.006, LIIR412
DOI: https://doi.org/10.52731/liir.v006.412

Predicting Neoadjuvant Therapy Response in Breast Cancer Patients: A Multi-Omics and Machine Learning Perspective

Lina AlRifai [*], Mostafa Z. Ali [*], Qasem Abu Al-Haija [‡], Talal Z. Ali [§], Mera Ababneh [¶]

Abstract

Breast Cancer (BC) treatment response varies due to underlying heterogeneity. Personalized therapy based on multi-omics profiling enhances efficacy by identifying patient-specific biomarkers and optimizing strategies. Multiomics integrates diverse biological data to understand mechanisms and enable customized treatments. Advances in ML and DL revolutionize BC therapy response prediction, leveraging multi-omics to improve precision, identify biomarkers, and refine strategies, reducing morbidity and mortality. This study presents a comparative analysis of multi-omics-dependent models for predicting neoadjuvant therapy response, highlighting techniques like DeepSurv, Gradient Boosting Machine (GBM), and Weighted MultiSource Canonical Correlation Analysis (WMSCCA). These models use data sets such as TCGA, METABRIC, and ICGC to boost predictive power. DL enables automated feature extraction, while ML offers interpretability for balanced predictive analytics. Despite progress, challenges remain, including data limitations, lack of external validation, and interpretability issues.

*Keywords:*Breast Cancer, Machine Learning, Multi-Omics, Neoadjuvant Therapy.

1 Introduction

Breast cancer (BC) arises from genetic and molecular abnormalities that cause uncontrolled cell division in breast tissue. Early detection and appropriate treatment increase survival

[*] Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan
[‡] Cybersecurity Department, Jordan University of Science and Technology, Irbid, Jordan
[§] Nursing Department, Wayne State University, Detroit, Michigan, United States
[¶] Clinical Pharmacy Department, Jordan University of Science and Technology, Irbid, Jordan

rates and improve quality of life.[1] BC is the second most common cause of death in women. It primarily affects women but can also affect men in some cases.Approximately one in eight women in the United States will be diagnosed with BC in their lifetime. Cancer can either be benign, meaning it does not spread to other tissues, or malignant, meaning it can invade nearby tissues.[2]

Treatment is classified into neoadjuvant and adjuvant therapies. Neoadjuvant therapy is administered before surgery to help reduce the tumor burden and includes chemotherapy, hormonal therapy, and other treatments.[3] To make surgery less invasive, the goal is to shrink a tumor or stop the spread of malignancy. In addition, adjuvant therapy aims to reduce the risk of recurrence and eradicate any remaining cancer cells.[4]

A critical challenge in BC treatment is the heterogeneity, which depends on the type of neoplasm, histological grade, and genetic constitution.[5] This variation highlights the need to identify diverse biomarkers to predict treatment responses based on individual patients. Consequently, this has led to the use of multi-omics data integration analysis.[6]

To improve personalized BC therapy, we need methods to determine the best therapy for the right patients based on genetic information and to handle complex molecular data. Researchers have been studying genetic predispositions to cancer for decades, identifying key genes that increase the risk, such as BRCA1, BRCA2, PALB2, CHEK2, ATM, BARD1, RAD51C, and RAD51D.[7]

Multi-omics, a relatively new concept, is gaining significant attention from researchers. Multi-omics can be achieved by augmenting ”omes” data: genome, proteome, transcriptome, epigenome, metabolome, and microbiome. This augmentation allows us to explore diseases further, understand their etiologies, and tailor treatments.[8] Machine learning can help predict drug responses in BC patients through multi-omics, improve accuracy, and identify robust biomarkers across various drug responses.[6]

Machine learning and deep learning have revolutionized the prediction of BC therapy responses. They can assist in cases where patients do not respond to treatment and may require combination therapies.[6] These advanced computational techniques have enabled us to analyze and interpret very complex multi-omics data due to their ability to perform dimensionality reduction by using several machine learning algorithms and models, such as autoencoders, Cox PH, Principal Component Analysis (PCA), K-means, and Support Vector Machines (SVMs).[9]

In this paper, we review the available literature on the use of machine learning for predicting BC drug response through multi-omics data, as well as the importance of handling multi-omics data and selecting features. We highlight models well-suited for predicting personalized therapies, emphasizing the potential of multi-omics approaches in tailoring treatment to individual patients. About our contribution:

- Review models that predict therapy response in multi-omics breast cancer.
- Review multi-omics-dependent models for breast cancer-related tasks.
- Discuss model use, strengths, and limitations.
- Discuss challenges, future research directions, and practical implications.

The paper is divided into five main parts. In the introduction, we define BC and multi-omics, explain treatment approaches, and discuss the challenges in predicting treatment responses due to heterogeneity. The literature review investigates available studies on BC utilizing multi-omics, explores therapy response prediction leveraging multi-omics, compares

models and performance metrics. In the discussion, we perform a comparative analysis of impactful models, highlight their strengths and limitations, and propose future research directions.

2 Review of Literature

In this section, we will review models related to neoadjuvant therapy response to multi-omics for breast cancer, and multi-omics-dependent models for breast cancer-related tasks.

2.1 Multi-omics Dependent Models for Breast Cancer Neoadjuvant Therapy Response

Table I highlights the variety of architectures and methodologies by comparing several models utilized for therapy response. The significant aspects compared include the year, architecture, datasets, input data, preprocessing steps, performance metrics, main strengths, and limitations.

The Neural Network(NN) Model [10] has two hidden layers used for survival and drug response prediction. Neighbourhood Component Analysis (NCA) was employed as a feature selection algorithm to identify high-weighted features. FireBrowse and CpG islands were used to preprocess the TCGA and GDSC datasets, respectively. Copy number values were normalized within the range of -1 to 1, where 0 indicates a normal copy number, -1 indicates a loss of copy genes, and +1 indicates a gain of copy genes. Bayesian Hyperparameter Optimization (BayesOpt) was used to optimize hyperparameters, and losses were propagated using scaled conjugate gradient backpropagation. It is used to enable effective identification with fewer evaluations in accordance with other optimizing techniques. The network's output for the risk group was determined using a categorical variable ranging from zero to one in survival prediction.

$$CE = \begin{cases} -\log(f(s_1)) & \text{if } t_1 = 1, \\ -\log(1 - f(s_1)) & \text{if } t_1 = 0. \end{cases} \tag{1}$$

Here, t_1=1 indicates the assignment of C_1= C_i or the sample. Grid Search and BayesOpt were utilized to optimize the entire network and its parameters. To forecast IC50 values for predicting drug therapy response, the NN was constructed as a regression problem. K-means clustering was used to divide IC50 values and binarize the prediction. An accuracy of 94% was reported for evaluating the model as a performance indicator.

Random Forest, Logistic Regression, and SVM [11] were used to predict the response to neoadjuvant therapy. Z-scores were calculated to determine feature importance. Five-fold cross-validation [12] was used to optimize model parameters. A 1000-step five-fold cross-validation was employed for randomization in hyperparameter optimization. Logistic regression was applied with elastic net regularization, using $L1$ ratio in the range 0.1 to 1, $C = 10^{-3}$ to 10^{3}. For SVM, radial basis function, sigmoid, or linear kernels were used, with gamma parameters ranging from *gamma*= 10^{-9} to 10^{-2} and C and C parameters similar to those in logistic regression. For Random Forest, the maximum number of features ranged between 5% and 70%, and the minimum sample split ranged between 2 and 15. The area under the curve (AUC) was used to evaluate the model's performance. The ARTemis dataset was used with fully trained models to validate and assess generalizability.

Two predictive models were proposed to support personalized treatment [13]: one to predict the BC subtype of a patient and another, called DCNN-DR, for drug response prediction. The first model was trained using BC patient omics and clinical data, while BC cell line omics data was used to train the second model. The omics data and associated features of 42 cell lines are summarized in Figure 1. The four omics, including methylation, CNV, mRNA, and mutation, were used to train the models and predict therapy responses. The DCNN-DR model was utilized to predict possible drugs for BLBC subtype patients. Convolutional NNs (CNNs) were the primary model used because they recognize hierarchical patterns and capture intricate relationships across multiple omics layers. The architecture consisted of multiple convolutional layers with ReLU activation functions, max-pooling layers for dimensionality reduction, and fully connected layers for final prediction generation. Dropout regularization was applied to prevent overfitting, and the Adam optimizer was used to train the model with a cross-validation-tuned learning rate. Principal Component Analysis (PCA) was employed to reduce dimensionality while retaining the most informative features, imputation was used to address missing values, and standardization was applied to ensure comparability among omics datasets. The CNN model was compared with single-omics and traditional machine-learning models, such as Random Forests and SVMs. With an AUC of 91%, accuracy of 85%, precision of 84%, recall of 82%, and an F1-score of 83%, CNN outperformed these alternative approaches.

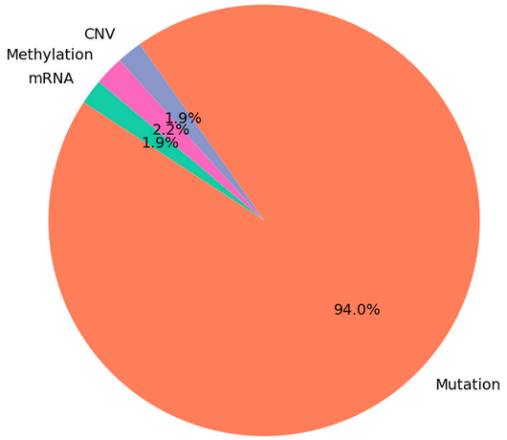

Figure 1: Number of features in each omics dataset for 42 cell-line samples used in various predictive models for BC therapy response.

The ComBat [14] classification algorithm was utilized. Six classification algorithms were examined, including K-Nearest Neighbors (K-NN), Random Forest, SVMs (SVM), NNs, linear models (Glmnet), and Kernel SVM. Two-fold cross-validation with Super-Learner was used to evaluate BEAUTY features, with AUC employed as the performance evaluation metric. It was observed that the kernel SVMs, kNN, and random forest models achieved the best median classification AUC.

In this preprint paper [15], the importance of developing causal discovery algorithms

was highlighted. These algorithms aim to uncover relationships by analyzing observational data. Explainable methodologies were utilized to forecast disease prognosis and provide appropriate treatments. A suitable causal discovery approach was employed to investigate how various genomic changes affect the prognosis of BC (BC). The methods used included PC, a Generalized Precision Matrix-based approach, and Greedy Equivalence Search (GES). To validate the results, BlueBERT [16] was utilized.

The Weighted Multi-Class SCCA (WMSCCA) model [6] was employed for feature selection. The main inspiration behind WMSCCA is its ability to interactively identify drug response class-specific multimodal biomarkers to enhance drug response prediction. Logistic regression with PCA was used to predict drug response categories. Five-fold cross-validation was employed to evaluate the prediction performance for treatment response objectively.

Table 1: Summary of Models and Performance

Ref.	Year	Architecture	Datasets	Input Data	Preprocessing Steps	Performance Metrics	Main Strengths	Limitations
[10]	2021	A NN, Regression, and K-means	The Cancer Genome Atlas (TCGA), and The Genomics of Drug Sensitivity in Cancer (GDSC)	Genomic data	**About TCGA dataset:** z-scaled RSEM of RNA and miRNA log2-RNA were used. Methylation and protein expression were already scaled. The Synthetic Minority Oversampling Technique (SMOTE) was used to balance the data. **About GDSC dataset:** RNA normalized basal expression levels were prevalent. CNVs were scaled between -1 and 1.	**Accuracy:** 94%, **AUC:** 98%	Survival Prediction, Drug Response Prediction	Small sample size,Limit the number of drugs.
[11]	2022	Logistic Regression, Random Forest, and SVMs (SVM)	The European Genome-phenome Archive (EGA)	Genomic data	Imputation to handle missing values, recursive feature elimination to identify the most relevant biomarkers, and standardization and normalization of data across omics layers.	**AUC:** 87%	Drug response prediction.	—
[13]	2022	SVMs	The Cancer Genome Atlas (TCGA), The Genomics of Drug Sensitivity in Cancer (GDSC), and The Cancer Cell Line Encyclopedia (CCLE)	Genomic data	Adam, SGD, and RMSprop were tested for optimizers. To prevent overfitting, dropout was used. A mean-squared error was used as the loss function.	**Accuracy:** 0.75, **Sensitivity:** 0.7, **Specificity:** 0.8	Drug response prediction.	Small datasets.
[14]	2023	Six algorithms	BEAUTY	Genomic data	Pre-NAC biopsies showed minimal gene expression differences between recurrent and non-recurrent cases.	AUC.	Drug response prediction	Small sample size.
[15]	2023	Discovery algorithms and language models	The Cancer Genome Atlas (TCGA)	Genomic data	Max-Min Markov Blanket (MMMB) and Mutual Information (MI) were used for feature selection, chosen depending on the data type. Causal discovery methods were applied to learn causal graphs, including PC, GES, FGES, and GPM.	NaN	Model's ability in feature selection	Incomplete work / pre-print.
[6]	2024	The Weighted Multi-Class SCCA	Own dataset, including 147 patients' BC data (clinical, mutation, molecular pathways, gene expression, tumor microenvironment cells)	Genomic data	Log2 transformation and Frobenius normalization were applied. PCA and five-fold cross-validation were used for drug response prediction.	**AUC:** 98%	Enhanced drug response	MOMLIN relies on correlation-based algorithms for data integration methods, Current state-of-the-art methods struggle a lot with causal inference.

2.2 Multi-omics Dependent Models for Breast Cancer Related Tasks

Table II compares different models for multi-omics in BC-related tasks, focusing on the year, architecture, datasets, input data, preprocessing steps, performance metrics, main strengths, and limitations.

The Gradient Boosting Machine (GBM) [17] is a supervised machine learning classifier and an ensemble learning method based on decision trees. It combines multi-omics data to predict the risk of developing BC (BC) in premenopausal women. Clustering was performed using Non-Negative Matrix Factorization (NMF) to divide the data into two groups using k-means clustering. The evaluation metric used was AUC, which achieved a score of 91%.

A Concatenation Autoencoder (ConcatAE) [18] incorporates the hidden attributes learned from each modality for data integration. It aims to enhance overall survival prediction by utilizing images. It uses multi-omics data, including gene expression, DNA methylation, miRNA expression, and Copy Number Variations (CNV). To maximize agreement between modalities and produce modality-invariant representations, the Cross-Modality Autoencoder (CrossAE) was used. T-distributed stochastic Neighbor Embedding (t-SNE) was employed to understand the similarity between paired hidden features further. The MNIST dataset was used to validate the effectiveness of the proposed models, which were then applied to the TCCA for overall survival prediction. In the first step, each data modality had its reconstruction loss in ConcatAE. The reconstruction loss was the summation of the separate reconstruction losses, as clarified in the equation when two modalities were integrated; the new reconstruction loss was:

$$L'_{\text{recon}} = \frac{1}{N}\sum_{n=1}^{N}\left((x_{1,n}-\hat{x}_{1,n})^2+(x_{2,n}-\hat{x}_{2,n})^2\right) \tag{2}$$

In the second step, the encoder and decoder were trained again using CrossAE, as shown in Figure 2. This process involved reconstructing modality 1 input data from modality 2 hidden features.

In the third step, the hidden features from each modality were combined, and the encoders, along with the task-specific network, were trained using task-specific loss functions such as cross-entropy or log-likelihood loss.

DSCCN [19] was used to classify BC subtypes. DSCCN performs distinctive analyses to highlight correlated features among multi-omics-expressed genes. The methods compared include Ensemble RF, Ensemble EN, Concate RF, Concate EN, DIABLO, SMPSL, DeepMo, and DSCCN. The FGL-SCCA model was utilized to predict BC subtypes and extract linear structured features from mRNA and DNA data. In the first step, the module encoder includes a fully connected layer that links features in the omics data. The module vectors Mj represent the j-th omics data, Wj denotes the weight of the fully connected layer, and Fj represents the module encoder. Given (xj,y)(xj,y), xj denotes a sample from the j-th omics data, while y is the classification label represented in this equation:

$$M^j(x^j) = \mathscr{F}^j_{\text{module}}(x^j; W^j_{\text{module}}) \tag{3}$$

The second attention mechanism focuses on models with high similarity between omics data. Cosine similarity was used to assess the degree of correlation. A cross-entropy error was applied in the third stage between the true and predicted labels.

The iSOM-GSN model [20] predicts using three CNNs with an integration layer, where the model's output depends on the majority vote from the predictions. The chi-square test is

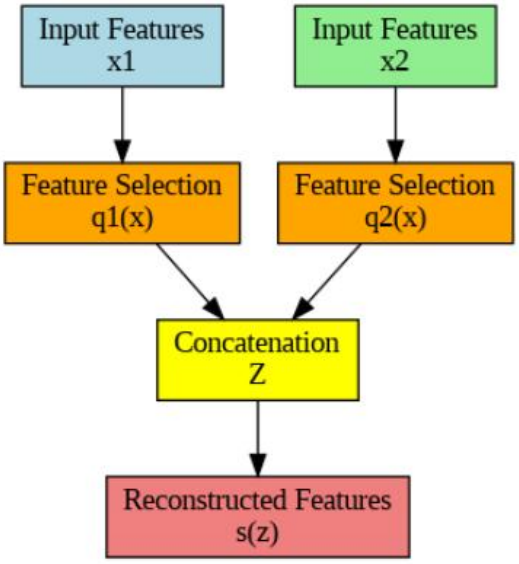

Figure 2: Multi-omics Data Augmentation with ConcatAE.

used to find the most relevant features. An SVM with the Radial Basis Function (RBF) Kernel is used to find the best subset of features: SOM reduces the dimensionality of datasets. iSOM-GSN uses the Euclidean distance to find the similarity between feature vectors. It updates the network neurons based on the Euclidean distance between the sample gene gij and the feature center cij as this equation:

$$d_j = \left[\sum_{i=1}^{n} (g_{ij} - c_{ij}) \right]^{1/2} \tag{4}$$

Where n (The number of samples). j (The current gene feature in the feature vector $\mathbf{v} = \{v_1, v_2, \ldots, v_m\}$. m (The number of features). The neurons with the lowest d_j in somNet are taken as the competition's winner to represent the data. This neuron is additionally called as the best matching unit(BMU) The ADAM optimizer is used for regularization and optimization during training to avoid overfitting.[21]

The Random Forest Model (RFM) and Generalized Linear Regression Model (GLRM) [22] were used to establish an Axillary Lymph Node (ALN) prediction model in the training queue. RFM is based on the Gini impurity formula, given by:

$$G = \sum_{i} p(i) \cdot (1 - p(i)) \tag{5}$$

$p(i)$ represents the probability that the random sample belongs to category i. To convert variables into a nomogram, the Feature Mapping Algorithm (FMA) was used.

The Multi-omics Stacked Fusion Network (MSFN) [23] has three components. First, a Residual Graph NN (ResGCN) is used to gain correlative prognostic information. Second, Convolutional NNs (CNNs) are employed to obtain specific prognostic information from multi-omics data. Third, AdaboostRF is used for survival prediction.

Table 2: Summary of Models and Performance

Ref.	Year	Architecture	Datasets	Input Data	Preprocessing Steps	Performance Metrics	Main Strengths	Limitations
[17]	2018	The Gradient Boosting Machine (GBM)	Local dataset from a radiological clinic	Genomic data	The number k of clusters in the analysis ranged from 2 to 10, and the silhouette index and cophenetic correlation were investigated as classical methods to evaluate clustering solutions.	**AUC:** 91%	Specific to premenopausal BC women	Covers a small scope of patients (only premenopausal).
[18]	2020	Concatenation Autoencoder, Cross-Modality Autoencoder	The Modified National Institute of Standards and Technology (MNIST), The Cancer Genome Atlas (TCGA)	Image	Imputation for missing values was applied using a log2 transform. Min-max normalization (0–1) and four-fold validation were used.	**ConcatAE:** 0.641±0.031, **CrossAE:** 0.63±0.081	Feature selection from large multi-omics data	The sample size was small, The validation dataset lacked genomic data.
[20]	2020	iSOM-GSN	METABRIC	Image	It used Chi-Square for feature selection. SVM was applied to find the best subset of features. ADAM was used for optimization and regularization.	**Accuracy:**94.32%, **Precision:**93.34%, **Recall:**97.7%, **F1-Score:**95.49%	Feature selection from large multi-omics data	Limited features in the multi-omics dataset.
[19]	2024	DSCCN	The Cancer Genome Atlas (TCGA)	Genomic data	Identifying correlated genes using SCCA with a fused pairwise group lasso penalty. The graph-guided pairwise group lasso (GGL) penalty was used to model bi-multivariate associations of mRNA and DNAm.	**Accuracy:** 0.926, **AUC:** 0.982	Classifying breast cancer subtypes	Limited number of features in the dataset, Lack of noncoding gene analysis, and Data imbalance.
[22]	2024	Random Forest Model (RFM), Generalized Linear Regression Model (GLRM)	Own datasets collected retrospectively	Image	Segmentation and capture in digital images, grayscale value capture of digital multi-omics variables.	AUC GLRM: 81.8%, AUC RFM: 89.3%	Early warning of axillary lymph node metastasis in BC patients	Retrospective and single-center design limited the generalizability, Machine learning algorithms were limited.
[23]	2024	AdaboostRF	The Cancer Genome Atlas (TCGA)	Genomic data	Feature extraction to obtain specific features for each omics data using CNN. Dropout and L2 regularization were applied to prevent overfitting.	**Accuracy:** 0.978, **AUC:** 0.991, **Precision:** 0.932, **Recall:** 0.964, **F1-Score:** 0.944	Survival Prediction	Lack of exploration into the interpretability of the survival prediction model.

3 Discussion and Future Direction

This section will discuss the models and algorithms applied, the strengths and limitations of existing work, the challenges and research gaps, future research directions, and practical implications.

3.1 Key Models and Applied Algorithms

When the ensemble method, Gradient Boosting Machine (GBM) [17], was used to predict the risk of developing BC in premenopausal women by combining multi-omics, it achieved a high AUC of 91%. To identify drug response class-specific multimodal biomarkers and enhance drug response prediction, WMSCCA [6] was used. The small size of datasets is a common limitation in drug response studies. Deep learning requires large databases and automatically extracts features, while traditional machine learning depends on structured data and manual feature engineering. Heuristic methods rely on trial and error for complex problems, whereas classical optimization provides mathematically optimal solutions for structured issues. Heuristic and deep learning methods are effective for unstructured problems but are computationally expensive. Classical optimization and machine learning are more interpretable and effective.

3.2 Strengths and Limitations of Existing Work

This section provides a detailed analysis of the strengths and limitations of existing work in the field. By examining various models and algorithms, we aim to highlight their notable achievements and the areas where they fall short. The discussion will cover multiple aspects, offering a comprehensive overview to guide further investigations and improvements in this domain.

3.2.1 Discuss the advantages of specific models and frameworks

In drug response prediction, NN, regression, k-means [10] and WMSCCA [6] achieved high accuracy. SVMs were used in small datasets and achieved high performance.[13] iSOM-GSN [20], DSCCN [19] were used in feature selection from large multi-omics data. Conventional ultrasound examination has advantages, including non-invasive, radiation-free, low-cost, and fast imaging speed.[22]. The reverse phase protein array (RPPA) is a high-throughput sequencing platform for protein detection that depends on antibodies to measure a target protein's expression. It is cost-effective and extremely sensitive to the target protein.[18] MOMLIN techniques reduce cost while maintaining predictive accuracy.[6]

3.2.2 Address limitations

Some limitations include limited generalizability due to the lack of external validation sets.[22] Some models lack exploration into the interpretability of predictions.[23] Biased limitations are faced in some models when using FPKM and RPM to normalize gene or miRNA expression. [18]

3.3 Open Challenges and Research Gaps

Despite remarkable advancements, several open challenges and data gaps persist, including the need for further real-world validation of these models. Many models perform excep-

tionally well in controlled experimental settings but fail in dynamic environments. This gap underscores the necessity of more extensive deployment and real-world stress testing. Additionally, insufficient focus on explainability makes it challenging to interpret model decisions. The absence of standardized benchmark datasets is another important limitation, making it challenging to evaluate generalization across diverse scenarios and fairly compare various approaches. Furthermore, hybrid models that combine data-driven approaches with rule-based methods offer favorable pathways toward greater robustness and generalization.

3.4 Future Research Directions

Future research should integrate technologies like federated learning and explainable AI to improve interpretability and clinical use. LLMs enhance NLP, reduce bias, and may combine with federated learning for private, decentralized training. Quantum computing could advance processing for complex models. As these tools evolve, ethical and societal issues must be addressed. Researchers should refine guidelines for privacy and fairness, with real-world validation needed.

3.5 Practical Implications and Industry Relevance

The reviewed techniques show strong potential for healthcare application. Effective adaptation requires addressing standardization in healthcare systems. Strategic planning must tackle deployment bottlenecks, including staff training and integration into hospital infrastructure. Our research supports confidence in these solutions, showing they are robust, improve patient outcomes, and gain stakeholder support by demonstrating benefits like enhanced safety, better outcomes, and greater efficiency.

4 Conclusion

This comparative study examined multi-omics-based models for predicting breast cancer neoadjuvant therapy response, reviewing architectures, datasets, preprocessing, and performance. It highlighted strengths and limitations of current methods, focusing on accuracy and interpretability. Existing models improve accuracy, enable biomarker discovery, and guide treatment decisions. However, challenges like data heterogeneity, limited validation, and clinical translation persist. Machine and deep learning automate feature extraction, while interdisciplinary collaboration fuels progress. These advances support better outcomes and personalized care. Key findings stress the need for robust validation, better generalization, and transparent models. Future work should prioritize external validation, standardized benchmarks, and novel methods like federated learning and explainable AI. Collaboration is crucial for real-world impact.

References

[1] Q. A. Al-Haija and A. Adebanjo, “Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network,” *Proc. 2020 IEEE Int. IOT, Electronics and Mechatronics Conf. (IEMTRONICS)*, Oct. 2020, Vancouver, BC, Canada, pp. xxx–yyy.

[2] National Breast Cancer Foundation, "About breast cancer," 2024. [Online]. Available: https://www.nationalbreastcancer.org/about-breast-cancer/

[3] A.S. Qari, A.H. Mowais, S.M. Alharbi, M.J. Almuayrifi, A.A. Al Asiri, S.A. Alwatid, A.A. Aljohani, R.M. Alanazi, and F. Al Thoubaity, "Adjuvant and Neoadjuvant Therapy for Breast Cancer: A Systematic Review," *Eur. J. Breast Health*, vol. 20, no. 3, pp. 156–166, May 2024.

[4] M.-E. Akbari, M. Ghelichi-Ghojogh, Z. Nikeghbalian, M. Karami, A. Akbari, M. Hashemi, S. Nooraei, M. Ghiasi, M. Fararouei, and F. Moradian, "Neoadjuvant vs adjuvant chemotherapy in patients with locally advanced breast cancer: A retrospective cohort study," *Ann. Med. Surg. (Lond)*, vol. 84, Nov. 2022, Art. no. 104921.

[5] Z.-Y. Zhang, Q.-L. Wang, J.-Y. Zhang, Y.-Y. Duan, J.-X. Liu, Z.-S. Liu, and C.-Y. Li, "Machine learning applications in breast cancer survival and therapeutic outcome prediction based on multi-omic analysis," *Yi Chuan*, vol. 46, no. 10, Oct. 2024, pp. 820–832.

[6] M. M. Rashid and K. Selvarajoo, "Advancing drug-response prediction using multimodal and -omics machine learning integration (MOMLIN): a case study on breast cancer clinical data," *Briefings in Bioinformatics*, vol. 25, no. 4, Jul. 2024, Article bbae300. [Online]. Available: https://doi.org/10.1093/bib/bbae300

[7] A. Yoshimura, I. Imoto, and H. Iwata, "Functions of breast cancer predisposition genes: Implications for clinical management," *International Journal of Molecular Sciences*, vol. 23, no. 13, Jul. 2022, Art. no. 7481. [Online]. Available: https://doi.org/10.3390/ijms23137481

[8] L. Marshall, B. N. Peshkin, T. Yoshino, J. Vowinckel, H. E. Danielsen, G. Melino, I. Tsamardinos, C. Haudenschild, D. J. Kerr, C. Sampaio, S. Y. Rha, K. T. FitzGerald, E. C. Holland, D. Gallagher, J. Garcia-Foncillas, and H. Juhl, "The essentials of multiomics," *The Oncologist*, 2022.

[9] D. Feldner-Busztin, P. F. Nisantzis, S. J. Edmunds, G. Boza, F. Racimo, S. Gopalakrishnan, M. T. Limborg, L. Lahti, and G. G. de Polavieja, "Dealing with dimensionality: the application of machine learning to multi-omics data," *Bioinformatics*, vol. 39, no. 2, Feb. 2023, Art. no. btad021.

[10] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomics*, vol. 22, Art. no. 214, Mar. 2021.

[11] S.-J. Sammut *et al.*, "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, pp. 623–629, Jan. 2022, doi: 10.1038/s41586-021-04278-5.

[12] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of machine learning algorithms with different K values in K-fold cross-validation," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 6, pp. 61–71, Dec. 2021.

[13] D. Khan and S. Shedole, "Leveraging deep learning techniques and integrated omics data for tailored treatment of breast cancer," *J. Pers. Med.*, vol. 12, no. 5, p. 674, Apr. 2022.

[14] X. Tang, K. J. Thompson, K. R. Kalari, J. P. Sinnwell, V. J. Suman, P. T. Vedell, S. A. McLaughlin, D. W. Northfelt, A. M. Aspitia, R. J. Gray, J. M. Carter, R. Weinshilboum, L. Wang, J. C. Boughey, and M. P. Goetz, "Integration of multiomics data shows down regulation of mismatch repair and tubulin pathways in triple-negative chemotherapy-resistant breast tumors," *Breast Cancer Res.*, vol. 25, no. 1, p. 57, May 2023.

[15] M. Farooq, S. Hardan, A. Zhumbhayeva, Y. Zheng, P. Nakov, and K. Zhang, "Understanding breast cancer survival: Using causality and language models on multi-omics data," *arXiv preprint arXiv:2305.18410*, May 2023.

[16] H. R. Seireg, Y. M. K. Omar, F. E. Abd El-Samie, A. S. El-Fishawy, and A. Elmahalawy, "Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction," *IEEE Access*, vol. 10, pp. [page range], 2022.

[17] H. Fröhlich, S. Patjoshi, K. Yeghiazaryan, C. Kehrer, W. Kuhn, and O. Golubnitschaja, "Premenopausal breast cancer: potential clinical utility of a multi-omics based machine learning approach for patient stratification," *EPMA Journal*, vol. 9, no. 2, pp. 175–186, Apr. 2018.

[18] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," BMC Medical Informatics and Decision Making, vol. 20, Art. no. 225, Sep. 2020.

[19] Y. Huang, P. Zeng, and C. Zhong, "Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning," BMC Bioinformatics, vol. 25, Art. no. 132, Mar. 2024.

[20] A. Alkhateeb, L. Zhou, A. Abou Tabl, and L. Rueda, "Deep learning approach for breast cancer InClust 5 prediction based on multiomics data integration," in Proceedings of the 11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB), Virtual Event, Oct. 2020.

[21] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Computing and Applications, vol. 35, no. 23, pp. 1–18, Apr. 2023.

[22] K. Ke, L. Shen, and J. Shao, "Early warning of axillary lymph node metastasis in breast cancer patients using multi-omics signature: A machine learning-based retrospective study," International Journal of General Medicine, vol. 2024.

[23] G. Zhang, C. Ma, C. Yan, H. Luo, J. Wang, W. Liang, and J. Luo, "MSFN: a multi-omics stacked fusion network for breast cancer survival prediction," Frontiers in Genetics, vol. 15, Aug. 2024.