IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A MODULAR MULTI-OMICS MACHINE LEARNING FRAMEWORK FOR ROBUST TRIPLE-NEGATIVE BREAST CANCER CLASSIFICATION UNDER PARTIAL DATA AVAILABILITY

[1]Kavya Gopal Bhat, [2]Kesiya Joy

[1]M. Sc. Bioinformatics, [2]Assistant Professor

[1]Department of Life Sciences, School of Sciences, Garden City University, Bengaluru, Karnataka, India.

Abstract: Triple-negative breast cancer (TNBC) is an aggressive subtype lacking established therapeutic targets, making an accurate molecular classification a key research interest. Although computational methods of classification using molecular data are increasingly adopted, limited multi-omics availability remains a major analytical challenge. This study aims for a flexible machine-learning framework to classify TNBC using genomic, transcriptomic, and multi-omics data, enabling independent analysis of incomplete data.

Public TCGA-BRCA resources were utilized, including genomic, transcriptomic, and clinical metadata for TNBC status annotation and model development. Individual genomic and transcriptomic data were independently processed and subsequently integrated using matched sample identifiers. ML models were trained on genomic, transcriptomic, and integrated datasets using consistent train-test splits. Logistic regression (LR), random forest (RF), support vector machine (SVM), and XGBoost were evaluated using ROC-AUC, recall, F1-score, accuracy, and runtime, with recall emphasized to improve TNBC detection sensitivity in this imbalanced dataset.

Multi-omics integration provided the most desirable discriminative ability, the ROC-AUC of which was 0.95. In all modalities of data, the logistic regression provided the most reliable balance between TNBC recall, discrimination, computational efficiency, and model stability. Optimization of the probability decision threshold further improved TNBC sensitivity with insignificant effects on overall classification.

This study indicates that the TNBC classification may be effectively aided by machine-learning models capable of operating under partial or complete multi-omics availability. The framework was deployed through a Streamlit-based web application for demonstrating research purposes. Overall, this study introduces scalable and interpretable TNBC classification and strong translational relevance using multi-omics machine learning.

Index Terms - Triple-negative breast cancer, multi-omics, machine learning, genomics, transcriptomics.

I. INTRODUCTION

Triple-negative breast cancer (TNBC) is defined by the simultaneous lack of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor (HER2) expression and is generally considered a clinically aggressive form of breast carcinoma. TNBC comprises about 10-15% of all breast cancer, and has a poorer prognosis as compared to other types (Zagami and Carey, 2022). TNBC is mostly found in women of younger age and pre-menopausal period, people of African origin, and carriers of

genetic BRCA1 mutations (Dietze et al., 2015). Treatment option for TNBC remains limited compared with hormone-receptor positive or HER2-positive disease. At present systemic chemotherapy remains the primary treatment, and targeted therapies are restricted only for specific patient subgroups (Obidiro et al., 2023).

Genomic and transcriptomic biomarkers reflect different but related biological processes involved in tumor development. However, each type of data provides an incomplete picture of tumor heterogeneity (Itai et al., 2023). This limitation has been overcome by integrating multiple data at different biological levels, resulting in improved disease characterization and classification performance compared with single-omics analysis (Cai et al., 2022). Machine-learning methods are useful, particularly to process high-dimensional heterogeneous data and have the necessary analytic capacity to identify complex patterns linked to molecular subtypes of cancer. As a result, multi-omics data integration with a machine learning framework improves the predictive accuracy in cancer classification (Arjmand et al., 2022; Zhang et al., 2025).

Although significant progress has been made in multi-omics studies using machine-learning methods, several important limitations still remain (Mohr et al., 2024). Many current studies often focus on single machine-learning models or customized integration approaches, without systematically providing comparisons of performance across model families and omics levels (Cai et al., 2022). This causes reproducibility and generalizability may be affected due to inconsistent experimental design, dataset curation practices, and evaluation procedures (Abbasi et al., 2025). Moreover, common prediction measures are more likely to focus on the overall accuracy, which does not consider the issues of class imbalance. This drawback is especially relevant in TNBC classification, where the minority class often represents clinically important cases (Yang and Mirzaei, 2024) These methodological gaps highlight the need for an effective evaluation framework that compares machine-learning models across genomics, transcriptomics, and integrated multi-omics datasets with balanced performance measures (Wekesa and Kimwele, 2023).

This study presents a modular machine-learning framework for TNBC classification that supports genomics-based, transcriptomics-based, and integrated multi-omics analysis (Hasin et al., 2017). This design choice directly addresses a major limitation of existing TNBC multi-omics studies by enabling robust prediction even when one or more molecular layers are missing (Song et al., 2020). Various machine-learning models were compared systematically on the three omics levels using consistent train-test splits to ensure fair and reproducible evaluation at each level (Horlacher et al., 2023). Due to inherent class imbalance in TNBC datasets, model performance was assessed on recall-oriented measures and ROC-AUC instead of on accuracy, thereby providing a more informative measure of minority-class detection without distorting the original data distribution (Saito and Rehmsmeier, 2015). Across all omics levels, logistic regression showed consistent performance, good discriminative ability, and computational efficiency, supporting its applicability to high-dimensional molecular data analysis (Hu et al., 2025). Finally, the selected models were used in a web-based platform to demonstrate deployment, reproducibility, and accessibility to research-oriented TNBC classification, providing a flexible and interpretable framework that accommodates varying data availability options and supports the development of translational bioinformatics research (Buga et al., 2025).

Past studies have supported the assumption that multi-omics integration enhances predictive performance for TNBC classification rather than single-omics approaches (Lehmann et al., 2021). The effectiveness of the multi-omics models is consistent with earlier studies shows that the combination of the genomic, transcriptomic, and other layers of molecular information better represents disease heterogeneity compared to single-omics analyses (Hasin et al., 2017). Simple models across omics layers, like logistic regression, often provide stable predictions with lower computational costs, which is consistent with results reported in multi-omics cancer biology (Couronné et al., 2018). By adjusting the probability thresholds, sensitivity to the minority class can be significantly enhanced, highlighting the importance of customized metrics in cancer cohorts with uneven class distributions (Saito and Rehmsmeier, 2015). Despite of these developments, challenges remain related to the data heterogeneity, complex feature landscapes, and the scarcity of fully harmonized multi-omics records (Hernández-Lemus and Ochoa, 2024). Future work should focus on the development of datasets, improving model interpretability, and validation of generalizability on independent populations to advance clinical applications.

II. RESEARCH METHODOLOGY

2.1 Data Collection

Breast cancer molecular and clinical data for this study were obtained from The Cancer Genome Atlas (TCGA) BRCA cohort, a large, publicly available multi-omics resource widely used in cancer research (Weinstein et al., 2013). Genomic mutation profiles, transcriptomic expression measurements, and corresponding clinical annotations were collected and harmonized for subtype definition and predictive modeling (Curtis et al., 2012). Triple-negative breast cancer (TNBC) samples were defined by the absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 receptor expression in accordance with established clinical guidelines for TNBC classification (Obidiro et al., 2023). After quality assessment and preprocessing, 1,094 transcriptomic samples were retained for gene expression-based model development and validation (Yu et al., 2020). Genomic mutation data were available for 826 samples, which were utilized for mutation-driven and integrative multi-omics analyses (Shah et al., 2012). Samples with matched genomics and transcriptomics data were selected for combined multi-omics analysis, ensuring consistent representation across molecular layers (Arjmand et al., 2022). All datasets analyzed in this work are fully de-identified and publicly accessible; therefore, no institutional ethical approval and individual consent were required (So and Knoppers, 2017).

2.2 Genomics Data Processing

Somatic mutation data in Mutation Annotation Format (MAF) were filtered to retain high-confidence variants for downstream analysis (Koboldt, 2020). Gene-level mutation matrices were constructed using binary encoding, assigning a value of 1 for mutated genes and 0 otherwise, facilitating machine learning classification tasks (Jiang and Jin, 2021). Only cancer-relevant genes with adequate mutation frequency were preserved to minimize noise and improve interpretability in genomics-based prediction (Colaprico et al., 2016). TNBC labels were incorporated into the mutation feature matrix for supervised learning, after which label columns were removed prior to model fitting (Picard et al., 2021). Following filtering and encoding, the final genomic feature set consisted of approximately 28 genes, representing the most informative mutational features.

2.3 Transcriptomics Data Processing

RNA-seq expression profiles from the TCGA-BRCA cohort were processed for predictive modeling (Yu et al., 2020). Raw expression values were log-transformed to stabilize variance and enhance comparability across samples (Shah et al., 2012). Genes with extensive missing values were excluded to reduce noise and improve downstream stability (Picard et al., 2021). Feature selection retained the most informative expression patterns, yielding a reduced set of 5,000 genes based on variability (Jiang and Jin, 2021). TNBC status derived from clinical metadata was merged with the processed expression matrix for supervised training. Before model training, label columns were removed from the input features to prevent data leakage. After quality control and preprocessing, 1,094 transcriptomic samples were retained for machine learning model development and evaluation (Liñares-Blanco et al., 2021).

2.4 Multi-omics Data Integration

Common TCGA sample barcodes were identified between the genomics and transcriptomics datasets to ensure accurate sample matching across molecular layers. Mutation features and expression profiles were combined at the sample level to generate an integrated multi-omics feature matrix representing aggregated molecular information per patient (Song et al., 2020). Missing entries introduced during the integration process were addressed using zero imputation to maintain sample completeness. The resulting multi-omics matrix comprised approximately 60,000 features and served as input for integrated TNBC classification (Lee et al., 2019). This strategy facilitates joint representation of complementary molecular signals from genomics and transcriptomics, supporting improved disease characterization relative to single-omics approaches.

2.5 Machine Learning Models

Supervised model classifiers were trained separately on the genomic, transcriptomic, and integrated multi-omics inputs. The evaluated algorithms included logistic regression (LR), random forest (RF), support vector machine (SVM), and XGBoost (Yang et al., 2025). Each dataset was divided into non-overlapping training and test sets to avoid data leakage and to provide an unbiased performance assessment. Models were implemented using standard machine learning libraries with default or recommended hyperparameters, and identical training-test splits were applied across all models within each dataset to support fair performance comparison (Abbasi et al., 2025).

2.6 Model Evaluation

Predictive performance was assessed using multiple complementary measures to capture both overall discrimination and minority-class detection. Specifically, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) quantified threshold-independently discriminative ability. Recall (sensitivity) was calculated to quantify the ability of each model to correctly identify TNBC cases, which is clinically crucial for imbalanced datasets. The F1-score, representing the harmonic mean of precision and recall, evaluated balanced performance especially for the minority class. Accuracy was also reported for general comparison (Yang et al., 2024). Confusion matrices were computed to derive true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each model. Additionally, probability threshold analysis was performed to examine the influence of decision threshold selection on recall and F1-score, thereby improving sensitivity in imbalanced classification. All metrics were computed on held-out test sets following standard evaluation practices (Lipton et al., 2014).

2.7 Model Selection

After comparative evaluation across genomics, transcriptomics, and integrated multi-omics datasets, logistic regression was selected as the final model for deployment. Although ensemble and kernel-based methods achieved competitive ROC-AUC values, logistic regression consistently demonstrated a favorable balance of TNBC recall, overall discriminative performance, computational efficiency, and model stability across omics layers. Prior studies have shown that logistic regression often performs comparably to more complex model approaches in high-dimensional biomedical data while offering improved interpretability and reduced risk of overfitting (Lynam et al., 2020). In addition, its lightweight computational requirements make it well-suited for scalable and web-based deployment. Accordingly, logistic regression was chosen to support robust TNBC prediction and practical implementation within the proposed framework.

2.8 Web Application Implementation

To demonstrate practical usability and reproducibility, the finalized logistic regression models were deployed via a Streamlit web application. The interface allows users to perform TNBC prediction independently using genomics, transcriptomics, or integrated multi-omics inputs, returning probability scores and binary classification outputs. This deployment strategy supports rapid model inference, facilitates flexible single-omics analysis when integrated data are unavailable, and promotes accessibility for research-oriented use. This application design enhances the transparency, reproducibility, and usability of machine learning models in biomedical research environments. The application architecture was designed to support modular analysis workflows while maintaining consistency across prediction pipelines.

2.9 Ethics Statement

All data used in this study were obtained from publicly available, de-identified TCGA resources. No protected health information was accessed, and no human participants were directly involved. Consequently, institutional ethical approval and informed consent were not required for this research.

III. RESULTS

3.1 Dataset

After quality control and data preparation, 1094 transcriptomic samples were retained, of which 875 were allocated for model training, and 219 were held out for evaluation. For genomics and integrated multi-omics analysis, 826 samples with matched molecular profiles were included, comprising 660 training instances and 166 test instances. Identical train-test splits were applied across transcriptomic, genomic, and multi-omics datasets to ensure consistent evaluation and fair comparison of predictive performance across all modeling scenarios

3.2 Overall Model Performance

Table 1A presents the transcriptomic model performance.

Transcriptomics	Random Forest	XG Boost	SVM	Logistic Regression
Training time	1.31 s	24.21s	0.003s	0.01 s
Accuracy	0.9041	0.8904	0.8735	0.7651
F1-score	0.4000	0.3333	0.2222	0.3810
Recall	0.3043	0.2609	0.1765	0.7059
ROC-AUC	0.9224	0.9257	0.7892	0.8030

Table 1B presents the genomic model performance.

Genomics	Random Forest	XG Boost	SVM	Logistic Regression
Training time	0.43 s	2.46 s	0.03 s	0.01 s
Accuracy	0.9096	0.8855	0.8735	0.7651
F1-score	0.3738	0.2400	0.2222	0.3810
Recall	0.2353	0.1765	0.1765	0.7059
ROC-AUC	0.7866	0.8484	0.7892	0.8030

Table 1C presents the integrated multi-omics model performance.

Multi-omics	Random Forest	XG Boost	SVM	Logistic Regression
Training time	3.48 s	279.88s	25.77 s	10.83 s
Accuracy	0.8916	0.9157	0.8916	0.8976
F1-score	0.1818	0.5000	0.4000	0.5641
Recall	0.1176	0.4118	0.3529	0.6471
ROC-AUC	0.9396	0.9503	0.9206	0.9309

Tables 1A – 1C summarize the performance of machine learning classifiers across transcriptomics, genomics, and integrated multi-omics datasets. Transcriptomics-based models demonstrated strong predictive capability. Random forest achieved the highest accuracy (0.9041), and logistic regression produced higher recall (0.7059), indicating improved sensitivity for TNBC detection. Genomics-only approaches resulted in comparatively lower F1-scores and recall, suggesting limited discriminative capacity when mutation features were analyzed independently. Integration of genomic and transcriptomic data further enhanced classification outcome, with XGBoost achieving the highest accuracy (0.9157), while logistic regression offered the most balanced performance, with recall (0.6471), F1-score (0.5641), and computational efficiency. Overall, transcriptomic features contributed substantially to model performance, while multi-omics integration provided additional improvement. Logistic regression showed stable and consistent behavior across all omics layers, with favorable sensitivity and short training time, supporting its selection for downstream deployment.

3.3 ROC Curve

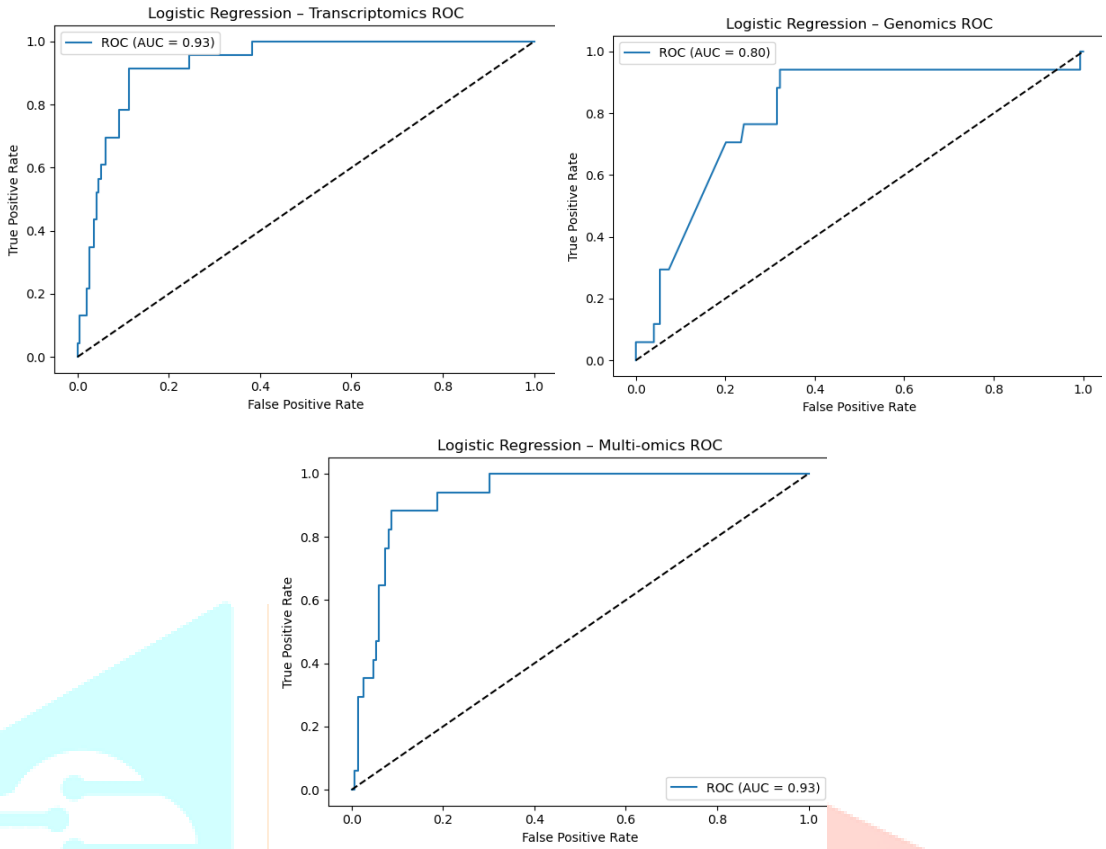

Figure 1. ROC curves of Logistic Regression models across transcriptomic, genomic, and multi-omics datasets.

Across all omics layers, logistic regression demonstrated strong discriminative ability, achieving ROC-AUC values of 0.93 for transcriptomics, 0.80 for genomics, and 0.93 for integrated multi-omics data. The highest performance was observed for transcriptomic and multi-omics models, highlighting the contribution of gene expression features to TNBC prediction. In contrast, mutation-based models showed only moderate predictive strength when used in isolation.

3.4 Confusion Matrix Analysis

Table 2A. Confusion matrix for transcriptomic analysis.

Model	TN	FP	FN	TP	Total
Random Forest	191	5	16	7	219
XGBoost	189	7	17	6	219
SVM	188	8	14	9	219
Logistic Regression	187	5	15	12	219

Table 2B. Confusion matrix for genomic analysis.

Model	TN	FP	FN	TP	Total
Random Forest	147	2	13	4	166
XGBoost	144	5	14	3	166
SVM	116	33	5	12	166
Logistic Regression	115	34	5	12	166

Table 2C. Confusion matrix for integrated multi-omics analysis.

Model	TN	FP	FN	TP	Total
Random Forest	146	3	15	2	166
XGBoost	145	4	10	7	166
SVM	142	7	11	6	166
Logistic Regression	138	11	6	11	166

To further examine classification behavior across different modelling strategies, confusion matrix statistics were computed for transcriptomic, genomic, and integrated multi-omics datasets (TN: true negatives, FP: false positives, FN: false negatives, TP: true positives).

Within transcriptomic-based classifiers, logistic regression achieved the highest number of true positive TNBC predictions, indicating improved sensitivity compared with random forest, XGBoost, and SVM. Genomics-based models yielded fewer true positive detections overall, aligning with their reduced recall values and moderate discriminative power when mutation features were used independently. In the multi-omics model, logistic regression again demonstrated a favorable balance between true positive detection and false classifications, reflecting its stable performance observed in ROC-AUC and recall metrics.

Collectively, these observations further highlight the importance of transcriptomic features in enhancing TNBC prediction and that integrated multi-omics modelling further improves sensitivity. Among all evaluated approaches, logistic regression consistently balanced classification behavior across omics layers, reinforcing its suitability for downstream deployment.

3.1 Threshold analysis

Supplementary Table S1. Threshold analysis for Random Forest (Transcriptomics)

Threshold	Accuracy	Precision	Recall	F1-score
0.50	0.904110	0.583333	0.304348	0.400000
0.45	0.908676	0.588235	0.434783	0.500000
0.40	0.908676	0.565217	0.565217	0.565217
0.35	0.908676	0.545455	0.782609	0.642857
0.30	0.904110	0.527778	0.826087	0.644068
0.25	0.899543	0.513514	0.826087	0.633333

Supplementary Table S2. Threshold analysis for XGBoost (Transcriptomics)

Threshold	Accuracy	Precision	Recall	F1-score
0.50	0.890411	0.461538	0.260870	0.333333
0.45	0.904110	0.562500	0.391304	0.461538
0.40	0.908676	0.578947	0.478261	0.523810
0.35	0.913242	0.600000	0.521739	0.558140
0.30	0.913242	0.600000	0.521739	0.558140
0.25	0.913242	0.590909	0.565217	0.577778
0.20	0.908676	0.565217	0.565217	0.565217

Supplementary Table S3. Threshold analysis for SVM (Transcriptomics)

Threshold	Accuracy	Precision	Recall	F1-score
0.50	0.9041	0.5833	0.3043	0.4000
0.45	0.9087	0.6154	0.3478	0.4444
0.40	0.8995	0.5333	0.3478	0.4211
0.35	0.9041	0.5556	0.4348	0.4878
0.30	0.9087	0.5714	0.5217	0.5455
0.25	0.8995	0.5161	0.6957	0.5926
0.20	0.8813	0.4571	0.6957	0.5517

Supplementary Table S4. Threshold analysis for XGBoost (Genomics)

Threshold	Accuracy	Precision	Recall	F1-score
0.50	0.890411	0.461538	0.260870	0.333333
0.45	0.904110	0.562500	0.391304	0.461538
0.40	0.908676	0.578947	0.478261	0.523810
0.35	0.913242	0.600000	0.521739	0.558140
0.30	0.913242	0.600000	0.521739	0.558140
0.25	0.913242	0.590909	0.565217	0.577778
0.20	0.908676	0.565217	0.565217	0.565217

Probability threshold adjustment applied to selected transcriptomic classifiers and genomics XGBoost model revealed that lowering the probability cutoff markedly increased TNBC recall, emphasizing the role of decision threshold selection in enhancing sensitivity under class-imbalanced classification conditions.

3.2 Web application

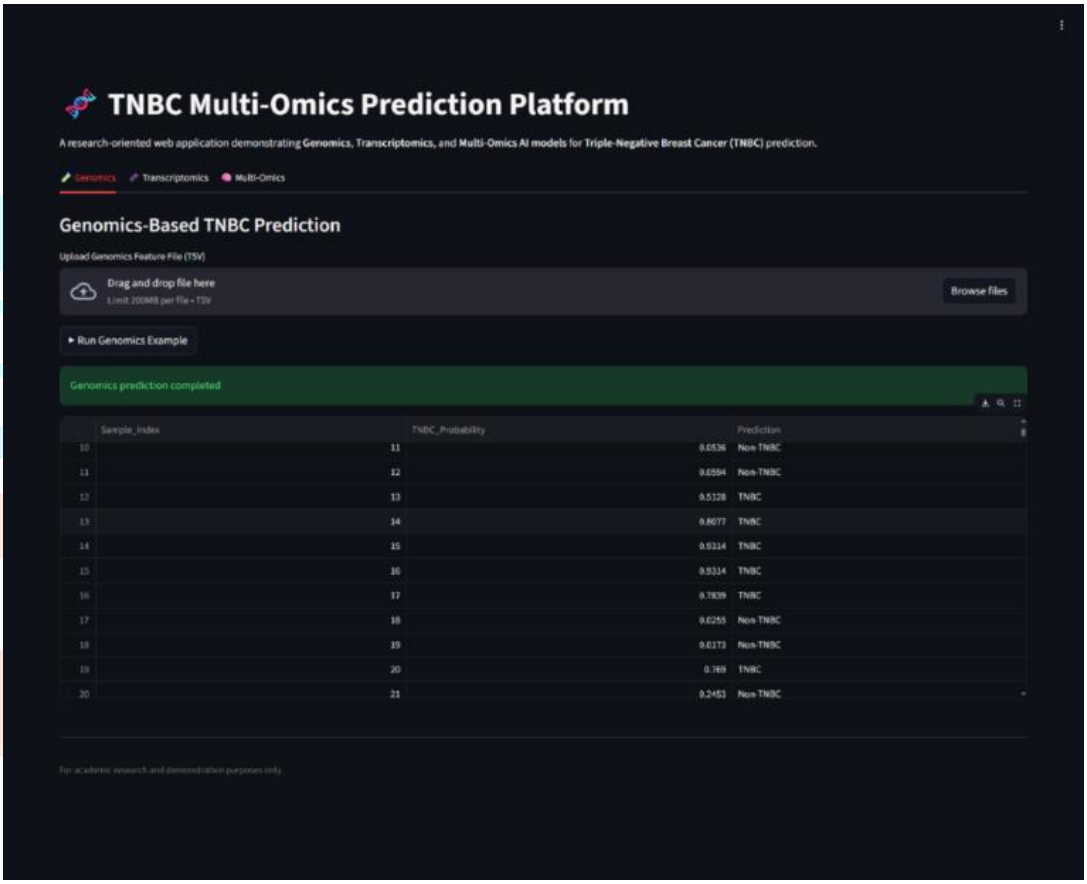

Figure 2. Streamlit-based TNBC prediction platform

A Streamlit-based web application was implemented to illustrate the practical applicability of the proposed TNBC classification framework (Figure 2). The platform offers separate interfaces for genomics, transcriptomics, and integrated multi-omics prediction. Pre-processed example datasets are included for demonstration, allowing users to run test trials directly within the interface. Upon execution, the application displays sample-wise TNBC probability scores along with predicted class labels, enabling rapid inspection of model outputs. This implementation highlights the feasibility of deploying the trained models in an interactive research-oriented environment and supports reproducible evaluation under single-omics and multi-omics scenarios.

IV. Discussion

This study assessed machine learning methods on transcriptomic, genomic, and integrated multi-omics for TCGA-BRCA data for TNBC classification (Subramanian et al., 2020; Yang et al., 2024). Integrated multi-omics models achieved the strongest overall discriminative performance, while transcriptomic models also showed high predictive capability, indicating the central role of gene expression in TNBC detection (Duan et al., 2021). Models based on mutation profiles demonstrated comparatively moderate performance (He et al., 2021; Subramanian et al., 2020). Logistic regression delivered stable behaviour

across omics layers with competitive ROC-AUC, robust recall, and minimal computational cost, justifying its selection for deployment (Bewick et al., 2005). Confusion matrix results confirmed higher true-positive detection by logistic regression (Kourou et al., 2015). Additionally, threshold adjustment markedly increased sensitivity, underscoring the value of recall-oriented evaluation in imbalanced cancer classification (Esposito et al., 2021; Subramanian et al., 2020).

These findings align with previous reports demonstrating that multi-omics integration enhances cancer classification by combining complementary molecular information (Huang et al., 2017). Prior TNBC studies have likewise reported transcriptomic features as primary drivers of subtype discrimination (Jézéquel et al., 2015). Several investigations have further shown that simpler linear models, such as logistic regression, can perform competitively with more complex classifiers in high-dimensional molecular datasets while offering greater stability and interpretability (Zhou et al., 2004). Furthermore, earlier work on imbalanced biomedical datasets has emphasized the limitations of accuracy-focused evaluation and highlighted recall as a clinically meaningful metric, particularly for aggressive cancers like TNBC, where minimizing false-negative predictions is critical (Jeni et al., 2013).

Beyond model benchmarking, this study introduces a modular prediction framework that supports transcriptomic, genomic, and integrated multi-omics analysis, addressing scenarios where complete multi-omics profiles are often unavailable. To facilitate accessibility and reproducibility, logistic regression models were deployed via Streamlit-based web application, enabling TNBC prediction using example or uploaded datasets. The platform reports probability scores and class labels across all omics settings, serves as a research-oriented demonstration tool. This implementation illustrates how lightweight machine learning workflows can be translated into interactive applications for exploratory molecular analysis.

Several limitations should be acknowledged. The study relies on TCGA data, and lacks validation on independent cohorts. Feature selection and preprocessing strategies may influence model behavior, and the relatively limited number of TNBC cases restricts statistical power. Although multi-omics integration improved performance, incorporating additional molecular layers such as proteomics or epigenomics may further strengthen predictive resolution. Future studies should prioritize external validation, broader omics integration, enhanced biological interpretability, and optimized decision thresholds to improve clinical sensitivity. These directions are important for advancing multi-omics machine learning toward translational TNBC research. The proposed framework may assist researchers in identifying TNBC cases and could support in future precision oncology.

V. Conclusion

This study introduces a modular machine learning framework for TNBC classification using transcriptomic, genomic, and integrated multi-omics TCGA-BRCA data. Transcriptomic features demonstrated strong predictive capability, while multi-omics integration further enhanced discrimination. Although ensemble approaches achieved high ROC-AUC in the integrated setting, logistic regression offered the most consistent balance of recall, stability, and computational efficiency across omics layers, supporting its selection for deployment. Threshold adjustment emphasized the importance of recall-oriented evaluation in this imbalanced dataset. The Streamlit-based application illustrates practical implementation for single-omics and multi-omics prediction. Overall, this work provides a scalable, interpretable, and reproducible TNBC classification framework even when the complete multi-omics data are unavailable. This highlights the potential of adaptable multi-omics machine learning pipelines for research-oriented translational bioinformatics research.

VI. Acknowledgment

The author would like to thank Kesiya Joy, assistant professor, Garden City University, for academic support and mentorship throughout the research period.

VII. Conflict of Interest

The author declares that there is no conflict of interest.

VIII. Data Availability

The datasets analyzed in this study are publicly available from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) portal. All analyses were performed using de-identified data. No new dataset was generated during this study.

IX. References

[1] Abbasi, A. F., Sajjad, M., Asim, M. N., Vollmer, S. and Dengel, A. (2025). "Multi-omics driven computational framework for cancer molecular subtype classification". Sci. Rep. https://doi.org/10.1038/s41598-025-32051-5

[2] Arjmand, B. et al. (2022). "Machine learning: a new prospect in multi-omics data analysis of cancer". Front. Genet., vol. 13, p. 824451. https://doi.org/10.3389/fgene.2022.824451

[3] Bewick, V., Cheek, L. and Ball, J. (2005). "Statistics review 14: Logistic regression". Crit. Care, vol. 9, no. 1, p. 112. https://doi.org/10.1186/cc3045

[4] Buga, R. et al. (2025). "Streamlit Application and Deep Learning Model for Brain Metastasis Monitoring After Gamma Knife Treatment". Biomedicines, vol. 13, no. 2, p. 423. https://doi.org/10.3390/biomedicines13020423

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

[6] Colaprico, A. et al. (2016). "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". Nucleic Acids Res., vol. 44, no. 8, pp. e71–e71. https://doi.org/10.1093/nar/gkv1507

[7] Couronné, R., Probst, P. and Boulesteix, A. L. (2018). "Random forest versus logistic regression: a large-scale benchmark experiment". BMC Bioinformatics, vol. 19, no. 1, p. 270. https://doi.org/10.1186/s12859-018-2264-5

[8] Curtis, C. et al. (2012). "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". Nature, vol. 486, no. 7403, pp. 346–352. https://doi.org/10.1038/nature10983

[9] Dietze, E. C. et al. (2015). "Triple-negative breast cancer in African-American women: disparities versus biology." Nat. Rev. Cancer, vol. 15, no. 4, pp. 248–254, 2015. https://doi.org/10.1038/nrc3896

[10] Duan, R. et al. (2021). "Evaluation and comparison of multi-omics data integration methods for cancer subtyping". PLoS Comput. Biol., vol. 17, no. 8, p. e1009224, 2021. https://doi.org/10.1371/journal.pcbi.1009224

[11] Esposito, C. et al. (2021). "GHOST: adjusting the decision threshold to handle imbalanced data in machine learning". J. Chem. Inf. Model., vol. 61, no. 6, pp. 2623–2640. https://pubs.acs.org/doi/10.1021/acs.jcim.1c00160

[12] Hasin, Y., Seldin, M. and Lusis, A. (2017). "Multi-omics approaches to disease". Genome Biol., vol. 18, no. 1, p. 83. https://doi.org/10.1186/s13059-017-1215-1

[13] He, Z., Zhang, J., Yuan, X. and Zhang, Y. (2021). "Integrating somatic mutations for breast cancer survival prediction using machine learning methods". Front. Genet., vol. 11, p. 632901. https://doi.org/10.3389/fgene.2020.632901

[14] Hernández-Lemus, E. and Ochoa, S. (2024) "Methods for multi-omic data integration in cancer research". Front. Genet., vol. 15, p. 1425456. https://doi.org/10.3389/fgene.2024.1425456

[15] Horlacher, M. et al. (2023). "A systematic benchmark of machine learning methods for protein–RNA interaction prediction". Brief. Bioinform., vol. 24, no. 5, p. bbad307.

[16] https://doi.org/10.1093/bib/bbad307

[17] Hu, Y. et al. (2025). "Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality". J. Med. Internet Res., vol. 27, p. e77721. https://doi.org/10.2196/77721

[18] Huang, S., Chaudhary, K., and Garmire, L. X. (2017). "More is better: recent progress in multi-omics data integration methods". Front. Genet., vol. 8, p. 84. https://doi.org/10.3389/fgene.2017.00084

[19] Itai, Y., Rappoport, N. and Shamir, R. (2023). "Integration of gene expression and DNA methylation data across different experiments", Nucleic Acids Res., vol. 51, no. 15, pp. 7762–7776, 2023. https://doi.org/10.1093/nar/gkad566

[20] Jeni, L. A. et al. (2013). "Facing imbalanced data--recommendations for the use of performance metrics". Humaine association conference on affective computing and intelligent interaction, IEEE, 2013, pp. 245–251. https://doi.org/10.1109/ACII.2013.47

[21] Jézéquel, P. et al. (2015) "Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response". Breast Cancer Research, vol. 17, no. 1, p. 43. https://doi.org/10.1186/s13058-015-0550-y

[22] Jiang, Q. and Jin, M. (2021). "Feature selection for breast cancer classification by integrating somatic mutation and gene expression". Front. Genet., vol. 12, p. 629946. https://doi.org/10.3389/fgene.2021.629946

[23] Koboldt, D. C. (2020). "Best practices for variant calling in clinical sequencing". Genome Med., vol. 12, no. 1, p. 91. https://doi.org/10.1186/s13073-020-00791-w

[24] Kourou, K. et al. (2015). "Machine learning applications in cancer prognosis and prediction". Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[25] Lee, E., Yoo, S., Wang, W., Tu, Z. and Zhu, J. (2013). "A probabilistic multi-omics data matching method for detecting sample errors in integrative analysis". Gigascience, vol. 8, no. 7, p. giz080. https://doi.org/10.1093/gigascience/giz080

[26] Lehmann, B. D. et al. (2021). "Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes". Nat. Commun., vol. 12, no. 1, p. 6276. https://doi.org/10.1038/s41467-021-26502-6

[27] Liñares-Blanco, J. et al. (2021). "Machine learning analysis of TCGA cancer data". PeerJ Comput. Sci., vol. 7, p. e584. https://doi.org/10.7717/peerj-cs.584

[28] Lipton, Z. C., Elkan, C. and Naryanaswamy, B. (2014). "Optimal thresholding of classifiers to maximize F1 measure", in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 225–239. https://doi.org/10.1007/978-3-662-44851-9_15

[29] Lynam, A. L. et al. (2020). "Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults". Res., vol. 4, no. 1, p.6. https://doi.org/10.1186/s41512-020-00075-2

[30] Mohr, A. E., Ortega-Santos, C. P., Whisner, C. M., Klein-Seetharaman, J. and Jasbi, P (2024). "Navigating challenges and opportunities in multi-omics integration for personalized healthcare". Biomedicines, vol. 12, no. 7, p. 1496. https://doi.org/10.3390/biomedicines12071496

[31] Obidiro, O., Battogtokh, G., and Akala, E. O. (2023). "Triple negative breast cancer treatment options and limitations: future outlook". Pharmaceutics, vol. 15, no. 7, p. 1796. https://doi.org/10.3390/pharmaceutics15071796

[32] Picard, M. et al. (2021). "Integration strategies of multi-omics data for machine learning analysis". Comput. Struct. Biotechnol. J., vol. 19, pp. 3735–3746. https://doi.org/10.1016/j.csbj.2021.06.030

[33] Saito, T. and Rehmsmeier, M. (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". PLoS One, vol. 10, no. 3, p. e0118432. https://doi.org/10.1371/journal.pone.0118432

[34] Shah, S. P. et al. (2012). "The clonal and mutational evolution spectrum of primary triple-negative breast cancers". Nature, vol. 486, no. 7403, pp. 395–399. https://doi.org/10.1038/nature10933

[35] So, D. and Knoppers, B. M. (2017). "Ethics approval in applications for open-access clinical trial data: an analysis of researcher statements to clinicalstudydatarequest.com". PLoS One, vol. 12, no. 9, p. e0184491. https://doi.org/10.1371/journal.pone.0184491

[36] Song, M. et al. (2020). "A review of integrative imputation for multi-omics datasets". Front. Genet., vol. 11, p. 570255. https://doi.org/10.3389/fgene.2020.570255

[37] Subramanian, I. et al. (2020). "Multi-omics data integration, interpretation, and its application". Bioinform. Biol. Insights, vol. 14, p. 1177932219899051. https://doi.org/10.1177/1177932219899051

[38] Weinstein, J. N. et al. (2013). "The cancer genome atlas pan-cancer analysis project". Nat. Genet., vol. 45, no. 10, pp. 1113–1120. https://doi.org/10.1016/j.jbi.2004.07.009

[39] Wekesa, J. S. and Kimwele, M. (2023). "A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment". Front. Genet., vol. 14, p. 1199087. https://doi.org/10.3389/fgene.2023.1199087

[40] Yang, S., Wang, Z., Wang, C., Li, C. and Wang, B. (2024). "Comparative evaluation of machine learning models for subtyping triple-negative breast cancer: a deep learning-based multi-omics data integration approach". J. Cancer, vol. 15, no. 12, p. 3943. https://doi.org/10.7150/jca.93215

[41] Yang, Y. and Mirzaei, G. (2024). "Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification". PLoS One, vol. 19, no. 2, p. e0293607. https://doi.org/10.1371/journal.pone.0293607

[42] Yang, Z. et al. (2025). "MLOmics: Cancer Multi-Omics Database for Machine Learning," Sci. Data, vol. 12, no. 1, p. 913. https://doi.org/10.1038/s41597-025-05235-x

[43] Yu, Z. Wang, X. Yu, and Z. Zhang, (2020). "RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches". Comput. Intell. Neurosci., vol. 2020, no. 1, p. 4737969. https://doi.org/10.1155/2020/4737969Digital Object Identifier (DOI)

X. Zagami, P. and Carey, L. A. (2022) "Triple negative breast cancer: Pitfalls and progress". NPJ Breast Cancer, vol. 8, no. 1, p. 95. https://doi.org/10.1038/s41523-022-00468-0

XI. Zhang, J. et al. (2025). "Deep learning–driven multi-omics analysis: Enhancing cancer diagnostics and therapeutics". Brief. Bioinform., vol. 26, no. 4, p. bbaf440. https://doi.org/10.1093/bib/bbaf440

XII. Zhou, X., Liu, K.Y., and Wong, S. T. C. (2004). "Cancer classification and prediction using logistic regression with Bayesian gene selection". J. Biomed. Inform., vol. 37, no. 4, pp. 249–259. https://doi.org/10.1016/j.jbi.2004.07.009