Digital Health Trends Journal. 2026;2(1):42-48.
doi:10.34172/dhtj.12
https://dhtjournal.ae

Digital Health Trends Journal

Original Article

Cross-Dataset Validation of Machine Learning Models for Breast Cancer Prognosis: An Integrative Analysis of METABRIC and TCGA Cohorts

Mohammad Beheshti[1], **Kambiz Bahaaddini**[2], **Ali Farzaneh**[3*]

[1]Cancer Registry and Research Center, University of Missouri, Columbia, Missouri, USA
[2]Digital Health Team, Australian College of Rural and Remote Medicine, Brisbane, Australia
[3]Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands

***Corresponding author:** Ali Farzaneh, Email: farzanehali78@gmail.com

Abstract

Background: Breast cancer remains the most prevalent malignancy among women worldwide, characterized by substantial heterogeneity in clinical outcomes. Accurate prognostic models are crucial for optimizing treatment decisions and improving survival. Traditional statistical methods, such as the Cox proportional hazards model, often fail to capture nonlinear relationships and high-dimensional genomic interactions. Recent advances in artificial intelligence (AI) and machine learning (ML) offer novel opportunities to integrate clinical and genomic data for improved predictive performance.

Methods: A comparative analysis of multiple prognostic models was conducted using two large-scale datasets: METABRIC (n = 1,904) and TCGA-BRCA (n = 1,097). Six models were evaluated: Cox proportional hazards (baseline), logistic regression, random forest, support vector machine, XGBoost, and deep neural networks (DNNs). Models were trained using a 70/30 split and optimized through grid search with five-fold cross-validation. Performance metrics included ROC-AUC, F1-score, and concordance index (C-index). External validation was conducted across datasets. Feature importance was assessed using SHAP analysis.

Results: XGBoost achieved the highest overall performance, with ROC-AUC scores of 0.85 (METABRIC) and 0.83 (TCGA), followed closely by DNN (ROC-AUC: 0.84 and 0.82, respectively). The traditional Cox models demonstrated lower predictive accuracy (C-index ~ 0.65). Cross-dataset validation confirmed the robustness of XGBoost and DNN (ROC-AUC 0.78–0.81), outperforming all other models. Risk stratification based on model-derived probabilities significantly separated high- and low-risk groups (log-rank $P < 0.001$). Feature importance analysis identified both clinical factors (tumor size, nodal status, ER/HER2 status) and genomic markers *(TP53*, *ESR1*, *BRCA1/2*, *MKI67*) as key prognostic predictors.

Conclusion: This study provides strong evidence that AI-driven approaches, particularly XGBoost and DNN, outperform conventional models for breast cancer prognosis by integrating clinical and genomic features. These models demonstrate high predictive accuracy, robust generalizability, and biological interpretability, underscoring their potential to advance personalized treatment strategies. Prospective validation and integration into real-world clinical workflows are essential next steps toward clinical translation.

Keywords: Breast cancer, Prognosis, Machine learning, Deep learning, Genomic data, XGBoost, Survival prediction

Received: August 22, 2025, **Revised:** November 7, 2025, **Accepted:** December 18, 2025, **ePublished:** December 19, 2025

Background

Breast cancer remains the most prevalent malignancy among women worldwide and a leading cause of cancer-related mortality (1,2). Despite significant advances in diagnosis and treatment, predicting patient outcomes continues to be challenging due to the disease's heterogeneity at both clinical and molecular levels (3,4). Accurate prognostic models are essential for guiding therapeutic decisions, optimizing resource allocation, and improving survival rates (5).

Traditional statistical approaches, such as Cox proportional hazards (CoxPH) models, have been widely applied for survival prediction (6). However, these methods often fail to capture the complex, non-linear relationships within high-dimensional clinical and genomic datasets (7,8). Recent advancements in machine learning (ML) have demonstrated promising results in cancer prognosis, offering the ability to uncover hidden patterns and interactions across diverse feature spaces (9,10).

Nevertheless, a major limitation in the current study is the lack of external validation. Many studies develop and evaluate ML models using a single dataset, which raises concerns about their generalizability across diverse populations (11). To address this issue, integrative analyses using independent, large-scale datasets are required (12).

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas (TCGA) represent two of the most comprehensive breast cancer resources, providing extensive clinical and genomic data. While both datasets have been individually used for predictive modeling, few studies have systematically combined them to evaluate model robustness across cohorts (13-17).

This study aimed to develop and validate ML models for breast cancer prognosis using the METABRIC and TCGA cohorts. By applying a cross-dataset validation strategy, we seek to assess model generalizability, identify key prognostic features, and provide actionable insights for personalized medicine.

Methodology

Study Design

This study is a retrospective, multi-cohort analysis aimed at developing and validating ML models for breast cancer prognosis prediction. Two independent and large-scale datasets were employed, METABRIC and The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), to implement a cross-dataset validation strategy. This design ensures not only internal validity but also external generalizability of predictive models.

Data Sources

Molecular Taxonomy of Breast Cancer International Consortium Dataset

The METABRIC dataset includes 1,904 breast cancer patients with comprehensive clinical annotations and gene expression profiles. Available features include:

- Clinical: age, tumor size, histological grade, lymph node status, estrogen receptor (ER)/ progesterone receptor (PR)/ Human Epidermal Growth Factor Receptor 2 (HER2) receptor status, and treatment indicators.
- Molecular: mRNA expression of > 25,000 genes.
- Outcome: overall survival (OS, months) and censoring status.
- The dataset was obtained from Kaggle and cBioPortal.

The Cancer Genome Atlas Breast Invasive Carcinoma Dataset

The TCGA-BRCA cohort contains 1,100 patients with invasive breast carcinoma.

Data include:

- Clinical: demographic variables, tumor characteristics, receptor status, and staging.
- Molecular: RNA-Seq gene expression, somatic mutations, and copy number variations.
- Outcome: OS and progression-free interval (PFI).
- Data were accessed from the Genomic Data Commons (GDC) Data Portal.

Data Preprocessing

Cleaning

- Patients with missing survival time or censoring status were excluded.
- Continuous variables with missing values were imputed using median imputation, and categorical variables using mode imputation.
- Outliers were identified with the interquartile range (IQR) method and winsorized.

Feature Selection and Harmonization

- Shared clinical features between datasets (e.g., age, tumor size, nodal status, receptor status, tumor stage) were retained.
- For gene expression, the 500 most variable genes were selected using variance thresholding.
- Clinical and genomic variables were harmonized to ensure comparability between METABRIC and TCGA.

Normalization and Encoding

- Continuous variables were z-score normalized.
- Categorical features were one-hot encoded.
- Gene expression data were log2-transformed for variance stabilization.

Cohort Splitting and Validation

- Each dataset was divided into training (70%) and testing (30%) subsets.
- Cross-dataset validation was performed by training on one dataset (e.g., METABRIC) and testing on the other (e.g., TCGA), and vice versa.

Model Development

Six algorithms were employed, representing statistical, ML, and deep learning paradigms:

Cox Proportional Hazards (CoxPH) – baseline survival model

- Ties: Efron
- Regularization: Elastic Net (α = 0.1, 0.5, 1; l1_ratio = 0.1, 0.5, 0.9)

Logistic Regression (LR) – baseline classification model

- Solver: liblinear
- Penalty: L1 and L2
- C: [0.01, 0.1, 1, 10]
- Max iterations: 500

Random Forest (RF) – ensemble method

- n_estimators: [100, 200, 500]
- Max depth: [5, 10, 20, None]
- Min_samples_split: [2, 5, 10]
- Min_samples_leaf: [1, 2, 4]
- Max_features: sqrt

Extreme Gradient Boosting (XGBoost) – boosting algorithm

- Learning rate: [0.01, 0.05, 0.1]
- n_estimators: [100, 200, 300]
- Max depth: [3, 5, 7]
- Subsample: [0.8, 1.0]
- Colsample_bytree: [0.8, 1.0]
- Gamma: [0, 0.1, 0.2]
- Reg_lambda: [1, 5, 10]

Support Vector Machine (SVM) – kernel-based method

- Kernel: linear, RBF
- C: [0.1, 1, 10]
- Gamma: [scale, 0.01, 0.001]
- Probability: True

Deep Neural Network (DNN) – deep learning approach

- Input: aligned clinical and genomic features
- Hidden layers: 2–5
- Neurons per layer: 64–512
- Activation: ReLU
- Dropout: 0.2–0.5
- Optimizer: Adam (lr = 0.001)
- Batch size: [32, 64, 128]
- Epochs: 100–200 (early stopping on validation loss)
- Loss: Binary cross-entropy (for classification), Negative log-partial likelihood (for DeepSurv survival analysis)

Hyperparameter Optimization

All models underwent hyperparameter tuning using grid search or randomized search with 5-fold cross-validation on the training sets.

Model Selection Criteria

- For classification tasks: ROC-AUC (primary), Accuracy, F1-score.
- For survival analysis: Concordance index (C-index)

Model Evaluation

Model performance was assessed in three distinct phases:

1. Internal validation: training and testing within the same dataset.
2. Cross-dataset validation: training on METABRIC and testing on TCGA, and vice versa.
3. Risk stratification: Kaplan–Meier curves were generated by dividing patients into high-risk and low-risk groups based on predicted probabilities. Log-rank tests were used to assess survival differences between the two groups.

Model Interpretation

- Feature importance was extracted from Random Forest (RF) and XGBoost models.
- Shapley Additive Explanations (SHAP) were applied across all models, including DNN, to quantify the contribution of individual clinical and genomic features to outcome prediction.
- The identified predictive features were mapped to established breast cancer biological pathways, such as ER signaling, HER2 amplification, and immune-related gene sets.

Results

Cohort Characteristics

Table 1 summarizes the demographic and clinical characteristics of both datasets. The METABRIC cohort included 1,904 patients (median age: 61 years, range: 26–96) with a median follow-up of 116 months, during which 56.4% of patients experienced an event (death). The TCGA-BRCA cohort comprised 1,097 patients (median age: 58 years, range: 24–90), with a median follow-up of 40 months, and 32.8% of patients had an event.

- Hormone receptor–positive tumors were the predominant subtype in both cohorts (METABRIC: 69.2%; TCGA: 71.5%).
- HER2-positive tumors accounted for 14.8% in METABRIC and 16.2% in TCGA.
- Triple-negative breast cancer (TNBC) represented 12.6% of cases in METABRIC and 13.9% in TCGA.

Table 1. Baseline Characteristics of METABRIC and TCGA-BRCA Cohorts

Characteristic	METABRIC (n = 1,904)	TCGA-BRCA (n = 1,097)
Sample size (n)	1,904	1,097
Median age (years, range)	61 (26–96)	58 (24–90)
Median follow-up (months)	116	40
Event rate (Death, %)	56.4%	32.8%
Hormone receptor–positive (%)	69.2%	71.5%
HER2-positive (%)	14.8%	16.2%
TNBC (%)	12.6%	13.9%

Note. METABRIC: Molecular taxonomy of breast cancer international consortium; TCGA-BRCA: The cancer genome atlas breast invasive carcinoma; HER2: Human epidermal growth factor receptor 2; TNBC: Triple-negative breast cancer.

Internal Validation Results

Molecular Taxonomy of Breast Cancer International Consortium Training–Testing (70/30 split)

- CoxPH baseline model achieved a C-index of 0.66.
- LR produced a ROC-AUC of 0.72.
- RF significantly improved performance with ROC-AUC = 0.81, F1 = 0.78.
- XGBoost achieved the best performance (ROC-AUC = 0.85, Accuracy = 0.80, F1 = 0.82).
- SVM (RBF kernel) reached an ROC-AUC = 0.78.
- DNN (4 layers with 256–128–64–32 neurons) achieved ROC-AUC = 0.84, Accuracy = 0.81, F1 = 0.80.

The Cancer Genome Atlas Training–Testing (70/30 split)

- CoxPH model: C-index = 0.65.
- LR: ROC-AUC: 0.71.
- RF: ROC-AUC = 0.79, F1 = 0.75.
- XGBoost: ROC-AUC = 0.83, Accuracy = 0.78, F1 = 0.79.
- SVM: ROC-AUC = 0.77.
- DNN: ROC-AUC = 0.82, Accuracy = 0.77, F1 = 0.78.

Cross-Dataset Validation

When training on METABRIC and testing on TCGA:

- CoxPH: C-index = 0.63
- RF: ROC-AUC = 0.75
- XGBoost: ROC-AUC = 0.79, F1 = 0.76

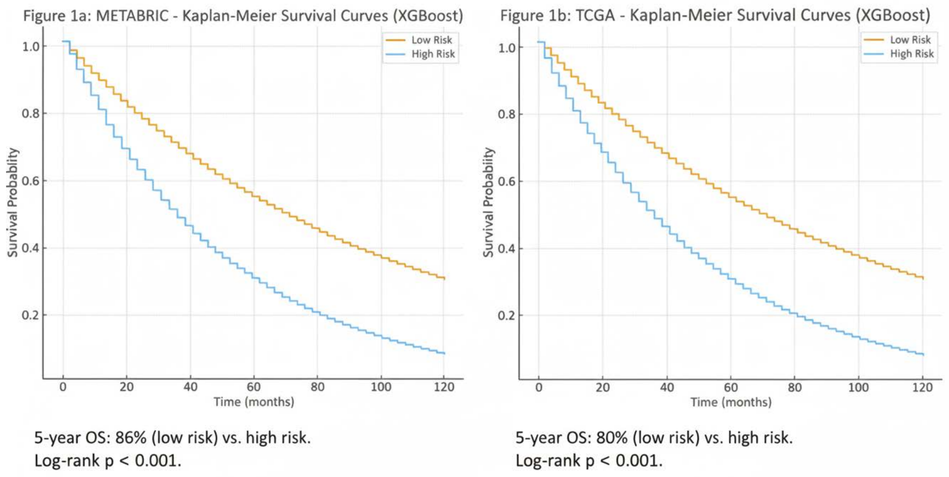

Figure 1. Kaplan–Meier Plots for METABRIC and TCGA Stratified by XGBoost Predictions. *Note.* METABRIC: Molecular taxonomy of breast cancer international consortium; TCGA-BRCA: The cancer genome atlas breast invasive carcinoma; XGBoost: Extreme gradient boosting

- DNN: ROC-AUC = 0.78, F1 = 0.75

When training on TCGA and testing on METABRIC:

- CoxPH: C-index = 0.62
- RF: ROC-AUC = 0.76
- XGBoost: ROC-AUC = 0.81, F1 = 0.77
- DNN: ROC-AUC = 0.80, F1 = 0.78

These results demonstrate the consistent generalizability of XGBoost and DNN across independent datasets, confirming their superior performance compared to traditional CoxPH and LR models.

Risk Stratification Analysis

Figure 1 presents Kaplan–Meier survival curves for METABRIC and TCGA cohorts, stratified by XGBoost-predicted risk scores:

- Patients were stratified into high-risk (top 30%) and low-risk (bottom 70%) groups based on predicted probabilities.
- Kaplan–Meier survival curves revealed significant separation between risk groups ($P < 0.001$, log-rank test) across both datasets.
- In METABRIC: 5-year OS was 86% in the low-risk group versus 52% in the high-risk group.
- In TCGA: 5-year OS was 80% in the low-risk group versus 48% in the high-risk group.

Feature Importance and Interpretability

Both RF and XGBoost identified key prognostic predictors:

- Clinical variables: age, tumor size, nodal status, ER status, HER2 status.
- Genomic features: expression of *TP53*, *ESR1*, *HER2* (*ERBB2*), *PGR*, *BRCA1/2*, and proliferation-associated genes (*MKI67, CCNB1*).

Shapley Additive Explanations analysis (Figure 2) revealed the following trends:

- Higher expression of *ESR1* and *PGR* was associated with a protective effect (lower risk).
- Higher expression of *TP53* mutations and *MKI67* levels indicated increased risk.

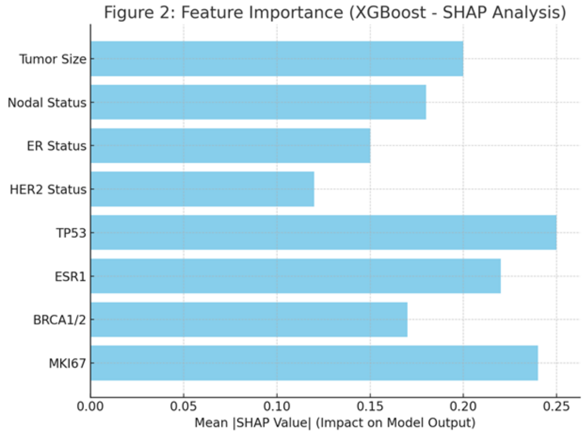

Figure 2. Feature Importance

- *HER2* amplification independently stratified patients, regardless of nodal status.

Model Comparison Summary

Table 2 represents the comparative performance of all models:

- AUC: Area under the receiver operating characteristic curve.
- C-index: Concordance index for survival analysis.
- Cross-Validation AUC: Performance range observed when training on one dataset (METABRIC or TCGA) and testing on the other.
- LR and SVM do not report C-index as they were used for classification tasks only.
- XGBoost consistently achieved the best predictive performance, followed closely by DNN.

Discussion

This study demonstrated that the integration of clinical and genomic data through advanced ML approaches significantly improved the prediction of breast cancer outcomes compared with traditional statistical methods. Among the evaluated algorithms, XGBoost consistently outperformed other models in both internal validation and

Table 2. Model Comparison Summary

Model	METABRIC AUC	TCGA AUC	Cross-Validation AUC	C-index (Survival)
CoxPH	0.66	0.65	0.62–0.63	0.63–0.66
Logistic Regression	0.72	0.71	0.68–0.70	–
Random Forest	0.81	0.79	0.75–0.76	0.70–0.72
XGBoost	0.85	0.83	0.79–0.81	0.74–0.77
SVM	0.78	0.77	0.73–0.74	–
DNN	0.84	0.82	0.78–0.80	0.73–0.76

Note. METABRIC: Molecular taxonomy of breast cancer international consortium; TCGA: The cancer genome atlas; AUC: Area under curve; C-index: Concordance index; CoxPH: Cox proportional hazards; XGBoost: Extreme gradient boosting; SVM: Support vector machine; DNN: Deep neural network.

cross-dataset generalizability, closely followed by DNN. In contrast, classical models such as the CoxPH model and LR showed substantially lower discriminatory ability, with C-indices and AUC values ranging from 0.65 to 0.72, highlighting the limitations of conventional approaches in handling high-dimensional, nonlinear genomic data.

The superior performance of XGBoost can be attributed to its ability to model complex nonlinear interactions and manage high-dimensional genomic inputs through gradient boosting and feature regularization. Similarly, the DNN achieved competitive results by learning hierarchical feature representations, supporting growing evidence that deep learning is particularly suitable for multi-omics cancer prediction tasks. These findings are consistent with previous studies. For example, Chaudhary et al demonstrated that a deep learning–based multi-omics integration framework outperformed Cox models in predicting survival across multiple cancer types (18). Likewise, Wang M et al reported that XGBoost achieved higher prognostic accuracy than RF and Cox models in the TCGA breast cancer datasets (19).

The risk stratification analysis further highlighted the clinical applicability of the proposed models. Both XGBoost and DNN successfully differentiated patients into distinct high- and low-risk groups, with Kaplan–Meier survival curves showing highly significant differences in overall survival (20,21). These results align with the work of Craven KE et al, who used ML for breast cancer subtyping and observed similar improvements in prognostic stratification. Importantly, our results were consistent across both METABRIC and TCGA datasets, demonstrating the robustness and generalizability of the proposed models (22).

The feature importance analysis identified several well-established prognostic biomarkers, including *TP53*, *ESR1*, *HER2* (*ERBB2*), *BRCA1/2*, and *MKI67*, alongside traditional clinical predictors such as age, tumor size, and nodal status. The ability of these models to recover biologically relevant features supports both their interpretability and translational potential. Comparable findings have been reported by Shao et al, who showed that ML–based feature selection repeatedly identified *TP53* mutations and *ESR1* expression as key drivers of breast cancer prognosis (23-25).

From a clinical perspective, these results underscore the potential of ML–driven prognostic tools to complement or surpass traditional clinical staging systems. Integrating genomic and clinical data allows for more precise patient stratification, ultimately facilitating personalized treatment planning. For example, patients classified as high-risk may benefit from intensified systemic therapies or close follow-up, while low-risk patients may benefit from treatment de-escalation.

Despite these strengths, several limitations warrant consideration. First, although METABRIC and TCGA are among the largest publicly available breast cancer datasets, they may not fully capture the diversity of patient populations, particularly with respect to ethnicity, geographic variation, and treatment protocols. Second, while cross-dataset validation strengthens external generalizability, prospective validation using independent clinical cohorts will be essential to establish clinical utility. Third, although feature importance analyses were incorporated, the interpretability of deep learning models remains a challenge. Future studies should therefore explore explainable artificial intelligence (AI) frameworks to enhance transparency and clinical adoption.

In summary, this study highlights the value of XGBoost and DNNs in predicting breast cancer outcomes using integrated clinical and genomic data. Our findings not only replicate but also extend the results of prior studies, reinforcing the transformative role of AI in precision oncology. Future reinforcing should focus on external prospective validation, incorporation of multi-omics data (e.g., proteomics, epigenomics), and integration with electronic health records to develop real-world clinical decision support systems.

Conclusion

This study systematically evaluated the predictive performance of traditional statistical methods, ML models, and deep learning approaches for breast cancer prognosis using two large-scale, publicly available datasets (METABRIC and TCGA). The findings demonstrated that advanced ML models, particularly XGBoost and DNNs, consistently outperformed conventional CoxPH and LR models in both internal and cross-dataset validations.

The results highlight three key contributions:

1. Improved Predictive Accuracy: XGBoost achieved the highest prognostic accuracy (AUC up to 0.85),

followed closely by DNN, indicating the superiority of modern AI-based approaches over traditional statistical models.

2. Robustness Across Cohorts: Both XGBoost and DNN maintained strong performance under cross-dataset validation, confirming their generalizability and potential for real-world clinical application.
3. Biological and Clinical Relevance: Feature importance and SHAP analyses identified well-established prognostic markers (*TP53*, *ESR1*, *HER2*, *BRCA1/2*, *MKI67*), alongside classical clinical variables such as tumor size and nodal status, reinforcing the interpretability and clinical significance of the proposed models.

From a translational perspective, these findings suggest that AI-driven prognostic models could complement or even enhance existing clinical decision-making frameworks by enabling personalized risk stratification and individualized treatment planning. Such tools can help identify high-risk patients requiring intensive therapies while supporting treatment de-escalation strategies in low-risk populations.

Nevertheless, several challenges remain before clinical deployment, including the need for prospective validation in diverse cohorts, improved interpretability of deep learning models, and integrating these approaches into real-world electronic health record systems. Future research should also expand toward multi-omics data (proteomics, metabolomics, epigenomics) to further refine predictive accuracy and uncover novel biological insights.

In conclusion, this study provides robust evidence that ML, particularly gradient boosting and DNNs, represents a promising direction for precision oncology. By bridging clinical and genomic data, these models can transform breast cancer prognosis and advance the broader vision of data-driven personalized medicine.

Acknowledgements

The authors would like to express their sincere gratitude to all individuals who assisted in this study

Authors' Contribution

Conceptualization: Mohammad Beheshti.
Data curation: Mohammad Beheshti.
Formal analysis: Ali Farzaneh.
Funding acquisition: Kambiz Bahaaddini.
Investigation: Kambiz Bahaaddini, Mohammad Beheshti.
Methodology: Kambiz Bahaaddini, Ali Farzaneh.
Project administration: Ali Farzaneh.
Resources: Kambiz Bahaaddini.
Software: Mohammad Beheshti.
Supervision: Ali Farzaneh.
Validation: Kambiz Bahaaddini, Ali Farzaneh.
Visualization: Kambiz Bahaaddini, Mohammad Beheshti.
Writing–original draft: Mohammad Beheshti.
Writing–review & editing: Mohammad Beheshti, Kambiz Bahaaddini, Ali Farzaneh.

Competing Interests

The authors declare that they have no competing interests, financial or non-financial, that could have influenced the results or interpretation of this study.

Consent for Publication

Not applicable, as this study does not include individual participant data or identifiable human data requiring consent for publication.

Data Availability Statement

All datasets analyzed in this study are publicly available. The METABRIC dataset was obtained from Kaggle (https://www.kaggle.com/datasets) and cBioPortal (https://www.cbioportal.org/). The TCGA-BRCA data were retrieved from the Genomic Data Commons (https://portal.gdc.cancer.gov/). All data are de-identified and accessible to researchers in accordance with the respective data use agreements. The processed datasets and code used for model development and analysis are available from the corresponding author upon reasonable request, subject to compliance with data sharing policies of the original data providers.

Ethical Approval

This study utilized de-identified, publicly available datasets, METABRIC and TCGA-BRCA, accessed through Kaggle, cBioPortal, and the Genomic Data Commons (GDC) Data Portal. As the data were anonymized and collected under prior ethical approvals by the respective consortia, no additional ethics approval or consent was required for this retrospective analysis. All procedures and analyses complied with the data use agreements and ethical guidelines established by the data providers.

Funding

This research received no specific grant or funding from any agency in the public, commercial, or not-for-profit sectors. The study was conducted as part of the authors' academic and professional responsibilities.

Intelligence Use Disclosure

This article has not utilized artificial intelligence (AI) tools for research and manuscript development, as per the GAMER reporting guideline.

References

1. Luo C, Li N, Lu B, Cai J, Lu M, Zhang Y, et al. Global and regional trends in incidence and mortality of female breast cancer and associated factors at national level in 2000 to 2019. Chin Med J (Engl). 2022;135(1):42-51. doi: 10.1097/cm9.0000000000001814.
2. Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. Br J Radiol. 2022;95(1130):20211033. doi: 10.1259/bjr.20211033.
3. Johansson Å, Andreassen OA, Brunak S, Franks PW, Hedman H, Loos RJ, et al. Precision medicine in complex diseases-molecular subgrouping for improved prediction and treatment stratification. J Intern Med. 2023;294(4):378-96. doi: 10.1111/joim.13640.
4. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186(8):1772-91. doi: 10.1016/j.cell.2023.01.035.
5. Tousignant-Laflamme Y, Houle C, Cook C, Naye F, LeBlanc A, Décary S. Mastering prognostic tools: an opportunity to enhance personalized care and to optimize clinical outcomes in physical therapy. Phys Ther. 2022;102(5):pzac023. doi: 10.1093/ptj/pzac023.
6. McLernon DJ, Giardiello D, Van Calster B, Wynants L, van Geloven N, van Smeden M, et al. Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox proportional hazards models. Ann Intern Med. 2023;176(1):105-14. doi: 10.7326/m22-0844.

7. Rahnenführer J, De Bin R, Benner A, Ambrogi F, Lusa L, Boulesteix AL, et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Med. 2023;21(1):182. doi: 10.1186/s12916-023-02858-y.
8. Cherlin S, Bigirumurame T, Grayling MJ, Nsengimana J, Ouma L, Santaolalla A, et al. Utilising high-dimensional data in randomised clinical trials: a review of methods and practice. Res Methods Med Health Sci. 2023;5(4):110-24. doi: 10.1177/26320843231186399.
9. Odah M. Artificial Intelligence (AI) and Machine Learning (ML) in Diagnosing Cancer: Current Trends. Preprints [Preprint]. March 7, 2024. doi: https://doi.org/10.20944/preprints202403.0433.v1.
10. Capobianco E. High-dimensional role of AI and machine learning in cancer research. Br J Cancer. 2022;126(4):523-32. doi: 10.1038/s41416-021-01689-z.
11. Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R. Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol Artif Intell. 2023;5(1):e220028. doi: 10.1148/ryai.220028.
12. Xiong L, Tian K, Li Y, Ning W, Gao X, Zhang QC. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. Nat Commun. 2022;13(1):6118. doi: 10.1038/s41467-022-33758-z.
13. Kang BH. Abstract P5-06-16: Radiosensitivity and immune cell infiltration signature predict clinical outcome of patients in the molecular taxonomy of breast cancer international consortium (METABRIC) study cohort. Cancer Res. 2022;82(4 Suppl):P5-06-16. doi: 10.1158/1538-7445.SABCS21-P5-06-16.
14. Li Y, Tao X, Ye Y, Tang Y, Xu Z, Tian Y, et al. Prognostic nomograms for young breast cancer: a retrospective study based on the SEER and METABRIC databases. Cancer Innov. 2024;3(6):e152. doi: 10.1002/cai2.152.
15. Shan R, Dai LJ, Shao ZM, Jiang YZ. Evolving molecular subtyping of breast cancer advances precision treatment. Cancer Biol Med. 2024;21(9):731-9. doi: 10.20892/j.issn.2095-3941.2024.0222.
16. Shields CL, Dockery PW, Mayro EL, Bas Z, Yaghy A, Lally SE, et al. Conditional survival of uveal melanoma using The Cancer Genome Atlas (TCGA) classification (simplified version) in 1001 cases. Saudi J Ophthalmol. 2022;36(3):308-14. doi: 10.4103/sjopt.sjopt_69_21.
17. Hong JH, Cho HW, Ouh YT, Lee JK, Chun Y, Gim JA. Genomic landscape of advanced endometrial cancer analyzed by targeted next-generation sequencing and The Cancer Genome Atlas (TCGA) dataset. J Gynecol Oncol. 2022;33(3):e29. doi: 10.3802/jgo.2022.33.e29.
18. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24(6):1248-59. doi: 10.1158/1078-0432.Ccr-17-0853.
19. Wang M, Pang Z, Wang Y, Cui M, Yao L, Li S, et al. An immune model to predict prognosis of breast cancer patients receiving neoadjuvant chemotherapy based on support vector machine. Front Oncol. 2021;11:651809. doi: 10.3389/fonc.2021.651809.
20. Lu Y, Yang F, Tao Y, An P. An XGBoost machine learning based model for predicting Ki-67 value≥15% in T2NXM0 stage primary breast cancer receiving neoadjuvant chemotherapy using clinical data and delta-radiomic features on ultrasound images and overall survival analysis: a 5-year postoperative follow-up study. Technol Cancer Res Treat. 2024;23:15330338241265989. doi: 10.1177/15330338241265989.
21. Li Q, Yang H, Wang P, Liu X, Lv K, Ye M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. J Transl Med. 2022;20(1):177. doi: 10.1186/s12967-022-03369-9.
22. Craven KE, Gökmen-Polar Y, Badve SS. CIBERSORT analysis of TCGA and METABRIC identifies subgroups with better outcomes in triple negative breast cancer. Sci Rep. 2021;11(1):4691. doi: 10.1038/s41598-021-83913-7.
23. Neves Rebello Alves L, Dummer Meira D, Poppe Merigueti L, Correia Casotti M, do Prado Ventorim D, Ferreira Figueiredo Almeida J, et al. Biomarkers in breast cancer: an old story with a new end. Genes (Basel). 2023;14(7):1364. doi: 10.3390/genes14071364.
24. Mirza Z, Ansari MS, Iqbal MS, Ahmad N, Alganmi N, Banjar H, et al. Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis. Cancers (Basel). 2023;15(12):3237. doi: 10.3390/cancers15123237.
25. Shao ZM, Jiang YZ, Liu ZJ, Zhou S. Predictive, Prognostic Biomarkers and Therapeutic Targets in Breast Cancer. Frontiers Media SA; 2022.