CASE REPORT

DOI: 10.37047/jos.galenos.2026.2025-9-2

J Oncol Sci

A Survey on Machine and Deep Learning Techniques for Breast Cancer Prediction

Pallavi JADHAV, Ajit S. PATIL

D. Y. Patil Education Society Deemed to be University, Kolhapur, India

ABSTRACT

Breast cancer is a leading cause of cancer-related mortality among women worldwide, making early and accurate diagnosis essential for effective treatment and improved patient outcomes. In recent years, machine learning (ML) and deep learning (DL) techniques have emerged as promising tools for predicting and classifying breast cancer using gene expression and clinical data. However, existing studies face several limitations. Many rely solely on ML or DL approaches, lack comprehensive strategies for feature selection or extraction, and demonstrate inconsistent performance across datasets. These gaps result in models that are insufficiently accurate, uninterpretable, or unable to generalize well to unseen data. This work aims to address these challenges by conducting a detailed literature survey of existing ML and DL models applied to breast cancer prediction. The objectives include identifying common datasets, performance metrics, model types, and feature-engineering techniques. A structured methodology was followed to analyze peer-reviewed studies and extract trends in performance and limitations. Findings show that, while DL models outperform traditional ML in terms of accuracy, they often lack transparency and robust feature engineering. In conclusion, a unified approach combining advanced feature selection and extraction methods with DL techniques is necessary to develop accurate, generalizable breast cancer prediction systems.

Keywords: Breast cancer; machine learning; deep learning; feature selection; feature extraction; prediction models

INTRODUCTION

Cancer is one of the leading causes of death worldwide and is characterized by the uncontrolled growth and spread of abnormal cells in the body. These cells can invade surrounding tissues and metastasize to distant organs, making the disease highly complex and difficult to treat in its advanced stages. Cancer originates from genetic mutations, which are often triggered by environmental factors, lifestyle choices, hereditary predispositions, or a combination of these.[1,2] As the disease progresses, it disrupts the normal functioning of vital organs and systems, ultimately resulting in significant morbidity and mortality. There are many different types of cancer, each named after the organ or tissue where it originates. The most common types include lung cancer, prostate cancer, breast cancer, colorectal cancer, and liver cancer. Others, such as pancreatic, ovarian, and brain cancers, are less common but often more aggressive. Each type of cancer has unique characteristics, progression patterns, and treatment protocols. Among these, breast cancer is the most commonly diagnosed cancer in women globally and is also a leading cause of cancer-related mortality in women.[3,4]

Breast cancer originates in the breast tissue, typically in the ducts or lobules. The disease begins when cells in the breast mutate and grow uncontrollably, forming a tumor. In many cases, these tumors can become malignant, meaning they have the potential to spread to other parts of the body. Factors contributing to the occurrence of breast cancer include age, genetic mutations (such as *BRCA1* and *BRCA2*), hormonal imbalances, lifestyle factors (e.g., alcohol consumption, obesity), and family history.[5] According to the World Health Organization (WHO), breast cancer has surpassed lung cancer as the most diagnosed cancer globally. In its 2021 report, WHO estimated that in 2020, 2.3 million women worldwide were diagnosed with breast cancer and 685,000 died from

Correspondence: Pallavi JADHAV MD,
D. Y. Patil Education Society Deemed to be University, Kolhapur, India
E-mail: pallavijadhav18@gmail.com
ORCID ID: orcid.org/0000-0001-5080-0813

Received: 16.09.2025 **Accepted:** 07.01.2026 **Epub:** 22.01.2026

Cite this article as: Jadhav P, Patil AS. A survey on machine and deep learning techniques for breast cancer prediction. J Oncol Sci. [Epub Ahead of Print]

Available at journalofoncology.org

it.[6] These alarming statistics highlight the pressing need for improved diagnostic and prognostic tools for early detection and effective treatment planning. Moreover, breast cancer can be identified using various forms of data. Imaging techniques such as mammography, ultrasound, and magnetic resonance imaging are widely used in clinical settings for tumor detection and localization.[7] Recently, gene expression data have emerged as a powerful source for understanding the molecular mechanisms underlying breast cancer.[8] Gene expression profiling (GEP) offers a more granular view by identifying genes that are overexpressed or underexpressed, aiding early diagnosis, subtype classification, and survival prediction.

However, most current research is primarily focused on imaging-based analysis using machine learning (ML) and deep learning (DL).[8-10] In contrast, comparatively little attention has been given to gene expression data, despite its rich potential for molecular-level insights. Among the existing studies that utilize gene expression data, a significant proportion rely on traditional ML models such as support vector machines (SVM), random forests (RF), and Naïve Bayes (NB), with limited exploration of novel or hybrid approaches. Moreover, only a few studies have employed DL for gene expression-based breast cancer prediction or survival analysis, and those that have done so have not demonstrated consistently high performance or robustness. This gap indicates the need for a focused review and analysis of existing methods in this area. To address this, the present study conducts a comprehensive review of recent works published between 2024 and 2025, selected from a pool of 150 papers retrieved from IEEE, Google Scholar, Web of Science, and Scopus. After outdated and less relevant papers were discarded, 25 papers were chosen for in-depth analysis. The main goal of this work was to explore and evaluate current methodologies, datasets, performance metrics, and predictive accuracies in breast cancer detection and gene expression analysis. The contributions of this work are as follows:

- Conducts a comprehensive review of recent (2024-2025) literature focused on breast cancer detection using gene expression data.
- Twenty-five carefully selected papers were analyzed from an initial pool of 150 retrieved from reputable sources such as IEEE, Web of Science, Scopus, and Google Scholar.
- Highlights the research gap where most studies are focused on imaging data, while gene expression data remains underexplored.
- Demonstrates that existing gene expression studies largely use basic ML models with limited application of novel or advanced techniques.
- Shows that DL methods applied to gene expression have not achieved high predictive accuracy (ACC) or model robustness.
- Categorizes existing approaches based on datasets, performance metrics, and models used, offering clear insight into current research trends.
- Provides a foundation for future work, encouraging the development of novel DL architectures specifically designed for gene expression-based breast cancer prediction and survival analysis.

This manuscript is structured to provide a comprehensive analysis of ML and DL approaches in breast cancer prediction. Section II presents the literature review, covering existing methodologies and related work. Section III discusses the findings from the reviewed studies, divided into four subsections: 3.1 Literature Survey Findings, 3.2 Datasets Used, 3.3 Performance Metrics Used, and 3.4 ML and DL Models with Feature Extraction and Selection Approaches. Section IV highlights the issues and challenges identified in current research. Finally, section V concludes the study with a summary of insights.

Literature Survey

The literature survey explores recent advancements in breast cancer prediction, classification, and survival analysis using ML and DL approaches. Various studies utilized multi-omic datasets, such as the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas-Breast Invasive Carcinoma (TCGA-BRCA), integrating clinical, genomic, and imaging data to enhance predictive ACC and personalize treatment. Various strategies were applied to feature selection, classification, and survival prediction. For instance, Mahmoud et al.[11] to develop an advanced genomics-based architecture for predicting breast cancer survival in order to address disease variability and complexity. In this study, the multi-omic METABRIC dataset,[12] which includes clinical data, somatic mutations, and gene expression from a large patient cohort, was integrated and pre-processed. This work employed DL approaches, including Graph-Convolutional Networks (GCN), Long Short-Term Memory (LSTM), and Variational Autoencoders (VAE), which were trained using a stochastic gradient descent optimization approach with an 80:20 train-test split. Evaluations were conducted using specificity, sensitivity/recall (REC), and ACC. The findings showed that among VAE, GCN, and LSTM, LSTM achieved 98.7% ACC. The findings show that integration of multi-omic data within an optimized DL approach improves the ACC of survival prediction and enables more effective, personalized treatment strategies.

Bharanidharan et al.[13], aimed at designing computational accurate/efficient system for cancer detection across five regions, i.e., renal (kidney), prostate, lung, liver and breast. For this study, the microarray dataset from 1027 patients, sourced from CuMiD,a[14] was considered. This work utilized Sparse Auto-Encoder, Independent Component Analysis and Principal Component Analysis (PCA) for dimensionality reduction of the dataset, and employed Remora-Optimization, guided by a local entropy-based fitness function, to enhance feature transformation and classification ACC. For classification, SVM, NB, decision tree (DT), and RF were utilized. For the evaluation of this study, six performance metrics, i.e., REC, balanced ACC score (BAC), F-score (FS), precision (PRE), Cohen's kappa coefficient (KAP), and Matthews correlation coefficient (MCC), were utilized. Results show that dimensionality was reduced from 36,805 to 80 features, with average balanced ACC improving to 93.4%, compared with 82.7% without the proposed approach. Kishore Khan et al.[15] aimed to develop an effective breast cancer classification approach using the METABRIC dataset.[12] This work included data preprocessing, dimensionality reduction using PCA, and MCC-based feature selection. Further, a deep neural network (DNN) was used; it was enhanced with dropout layers, and early stopping was applied during training on selected features using MCC. The performance was evaluated using ACC, PRE, REC, and FS, alongside gene expression visualization. Results show that dimensionality reduction provided a boost in classification performance, resulting in higher ACC.

Das et al.[16] focused on enhancing breast cancer staging and classification by integrating ML with bioinformatics analyses using gene expression data from TCGA-BRCA dataset.[17] The methodology of this work involved the identification of differentially expressed genes and their analysis using protein-protein interaction, regulatory-network, and signaling-pathway approaches to uncover potential therapeutic targets. The ML models used included RF, SVM, DT, Gaussian NB (GNB), K-Nearest Neighbors (KNN), and eXtreme Gradient Boosting (XGB) for classification of cancer stage and cancer subtype. For evaluation, ACC, PRE, REC, FS, and specificity were considered. Evaluations showed that RF and XGB achieved better results, reaching 97.19% and 95.23%, respectively. Findings show that key proteins and micro ribonucleic acids (miRNAs) are potential biomarkers, demonstrating the method's potential to advance personalized treatment approaches. Hu et al.[18], aimed at enhancing identification of cancer-driver genes by addressing limitation in existing approaches related to feature relationships and noise in protein-protein interaction (PPI) data. This work utilized a dynamic-incentive-model (DIM) to construct a hypergraph to minimize false positives in PPI networks. Gene importance within hyperedges was assessed using Network Functional score (NFS), and DIM integrates NFS with miRNA and messenger RNA (mRNA) differential expression scores. The DIM was evaluated on pan-cancer, prostate cancer, lung cancer, and breast cancer datasets. The evaluation was performed using the area under the receiver operating characteristic curve (AUC-ROC), with DIM outperforming existing approaches and demonstrating strong cross-cancer generalization, thereby improving targeted gene discovery.

Kurniadi and Saputri[19] aimed to investigate breast-cancer survivability using multi-modal data from the METABRIC dataset.[12] Because the dataset is high-dimensional, this work used XGB to select top-k features. This work further utilized ML classifiers (XGB, RF, SVM, and KNN) using selected features. The evaluation metrics used for the study included ACC, PRE, REC, and FS. Findings show that XGB and RF achieved the highest ACC (72.7%). Findings show that feature optimization is important for achieving good performance in survival prediction. Brahmatej Rupavath et al.[20] aimed to improve metastasis prediction in breast cancer by using a recursive neural network (RecNN) and the METABRIC dataset[12], which provides comprehensive genomic information. The approach included data preprocessing using named-entity recognition for structured classification and feature selection using Least Absolute Shrinkage and Selection Operator (LASSO). In this work, the RecNN was applied to predict metastasis on the basis of selected features. For evaluation, ACC, PRE, REC, and FS were considered; the RecNN achieved 98.69% ACC, outperforming approaches such as CNN.

Puttegowda et al.[21], aimed at improving breast-cancer and personalized treatment by predicting key-clinical attributes, i.e., cancer subtype, tumor stage and progesterone-receptor status utilizing ML. For this study, the METABRIC dataset was utilized, which provides extensive clinical and genomic information. In this study, classification was performed using logistic-regression (LR), SVM, RF, and an ensemble of these approaches. Evaluations were conducted with respect to ACC, with SVM-radial basis function achieving 99.79% ACC, SVM achieving 97.93% ACC, RF achieving 97.59% ACC, and LR achieving 89.45% ACC. Findings show the effectiveness of non-linear models in capturing complex patterns, achieving prognostic ACC, and supporting personalized cancer treatment planning. Ghosh et al.[22] focused on the identification of subtype-specific gene biomarkers for breast cancer, utilizing gene expression data to support precise treatment and classification. For this study, the TCGA-BRCA dataset[17] was utilized. The methodology involved feature selection using LASSO, which was combined with four ML approaches (i.e., NB, KNN, SVM, and RF) to determine the best approach, with SVM achieving the best performance.

Furthermore, a modified Compact Genetic Approach (mCGA) was employed to refine biomarker selection for subtypes [basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal B, and Luminal A]. Evaluations showed AUC-ROC values of 100% for HER2 and Basal, 97.31%, and 98.78% for Luminal A. The pathway and enrichment analyses demonstrated biological relevance.

Turova et al.[23], aimed at improving breast-cancer subtyping by introducing breast-cancer classifier (BCC), a ML-based approach which utilizes RNA-sequence data for addressing limitation in current approaches like Immunohisto-Chemistry and PAM50, which are particularly focused in identifying HER2-low sub-type. In their study, data from TCGA-BRCA,[17] SCAN-B cohorts,[24] and METABRIC[12] were considered, from which the BCC was developed. The approach involved training BCC to classify breast-cancer subtypes, with a focus on distinguishing HER2-low as a unique group and reclassifying PAM50's normal subtype. Statistical analysis showed that BCC had high ACC. Findings show prognostic similarities between HER2-low and basal subtypes. Findings show that BCC's approach has potential to enhance treatment stratification and to deepen molecular understanding of breast cancer. Asfaw and Tegaw[25] aimed to compare survival outcomes of breast cancer patients undergoing mastectomy versus breast-conserving surgery (BCS) using ML approaches. This study utilized the METABRIC dataset,[12] which was first preprocessed using an imputation approach, then subjected to a Synthetic Minority Oversampling Technique (SMOTE)-based class-balancing approach, and finally underwent feature selection. For classification, DT, XGB, LR, GNB, RF, KNN, SVM, AdaBoost, and Gradient-Boosting (GB) were used; GB achieved 95.4% training and 86.4% testing ACC for mastectomy class. For the BCS class, GB achieved training and testing ACCs of 94.6% and 82.8%, respectively. The important features included age, the Nottingham Prognostic Index, and relapse-free status. Findings showed that younger patients derived greater benefit from BCS, supporting personalized treatment approaches.

Yaqoob and Verma[26] aimed to enhance breast cancer classification using gene-expression data by introducing a hybrid feature-selection approach that combined the Kashmiri-Apple Optimization Approach (KAO) and Armadillo-Optimization Approach (AOA), followed by an SVM classifier. The KAO was employed for global exploration of informative genes, while AOA performed local refinement to reduce redundancy and prevent premature convergence. The KAO-AOA-SVM was applied to breast cancer datasets, achieving 98.97% ACC using only 15 genes. The approach demonstrated consistent performance across gene subsets, indicating robustness and potential for clinical and cross-cancer applications. Kallah-Dagadu et al.[27], aimed at enhancing breast-cancer prediction by identification of key-genes using ML and explainable AI (XA) approaches. This study used the TCGA-BRCA[17] dataset, which contained 1,208 samples and 3,602 gene features. In this work, KNN, SVM, and RF were applied with feature selection. The XAI approaches included accumulated local effects, partial dependence plots, and SHapley Additive exPlanations (SHAP) values, which were used to interpret model outputs and assess gene importance. The leaving-one-covariate-in approach was used to identify the top ten predictive genes, with SVM and RF rankings closely aligned. The Findings showed the value of explainability in ML-driven cancer diagnosis for improving clinical decision-making.

Aliouane et al.[28] aimed to improve breast cancer classification by integrating a DL approach with SHAP to interpret gene expression data. The approach involved training a DL model and evaluating it using 5-fold cross-validation and ensemble learning on the ArrayExpress (E-MTAB-3732) dataset,[29] achieving a mean ACC of 99.64%. To assess generalizability, the CuMiDa dataset[14] was utilized; it comprises only three databases (GSE42568, GSE7904, and GSE45827), which achieved ACCs of 99.14%, 100%, and 98.67%, respectively. SHAP analysis identified key genes, including KRT5, ESR1, KRT19, and DSCAM-AS1. Further validation using the MalaCards[30] database demonstrated the relevance of genes, providing evidence of the method's effectiveness for biomarker interpretability and discovery. Li et al.[31] aimed to develop a stable and accurate approach for breast cancer prognosis that addressed data distribution shifts across diverse datasets; they presented a model called Deep-Global Balancing-Cox Regression (DGBCox), which integrated causal inference with DP. The gene-expression data were first transformed into latent representations using a deep autoencoder, and the resulting representations were then balanced using a causality-based approach. Causal features were selected using balanced representations for survival prediction. The DGBCox was evaluated on twelve breast cancer datasets, on which it outperformed existing benchmark approaches in both stability and predictive ACC, demonstrating effectiveness in heterogeneous data scenarios and improving prognostic reliability.

Rabah et al.[32] aimed to enhance noninvasive breast cancer subtype classification by developing a multi-modal DL approach that combined mammography images with clinical metadata. Utilizing the Chinese Mammography Database[33], which contains 4,056 mammography images from 1,775 patients, the approach classifies breast lesions into five classes: triple-negative, HER2-enriched, Luminal B, Luminal A, and benign. The approach integrated image

and clinical data, and its performance was evaluated using AUC, achieving 88.78% ACC. Sridharan and Ghosh[34] aimed to enhance breast cancer survival prediction by integrating GEP data with agent-based modelling (ABM). The approach first identified key genes involved in cancer progression using GEP, and then constructed a model representing how genes influence cellular behavior. These insights were incorporated into ABM to simulate tumor growth and treatment response under various conditions. The predictive performance of the GEP-ABM was validated using actual patient data and benchmarked against existing approaches using ACC-based metrics. Findings suggest that GEP-ABM integration improves survival predictions and supports more personalized, data-driven breast cancer treatment strategies.

Kunta and Lepakshi[35] aimed at developing ascalable, non-invasive solution for breast cancer detection using mRNA gene expression data. The approach involved transforming one-dimensional mRNA sequences into two-dimensional images to capture spatial information. After standard preprocessing and applying SMOTE for class balancing, features were extracted using AlexNet and ResNet101 to mitigate issues such as local feature dependence and vanishing gradients. These features were further combined and used to train an Ensemble-of-Ensemble classifier, which incorporated XGB, RF, AB, bagging, and extra-trees for consensus-based prediction. When evaluated on gene expression data, the model achieved 99.91% ACC, confirming its robustness and applicability. Li et al.[36] presented a novel bi-clustering approach, Bi-clustering differential-sparsity-constraints and dynamic-graph-regularization (BCDD), designed to enhance cancer subtype classification by addressing limitations in existing sparse singular-value decomposition (SVD)-based approaches. The approach incorporated differential sparsity constraints, applying an L1/2-norm to genes and an L1-norm to samples, to reflect the inherent sparsity imbalance in cancer gene expression data. Additionally, a dynamic graph regularization strategy was proposed, which enabled iterative updates to the graph adjacency matrix based on changes in SVD to avoid bias introduced by previously extracted biclusters. For evaluations, the five datasets from TCGA[37] were considered; these included the TCGA-BRCA dataset.[17] The BCDD demonstrated superior bi-clustering ACC and robustness compared to state-of-the-art methods, confirming its effectiveness in identifying biologically relevant gene modules.

Goidescu et al.[38], aimed at exploring contribution of moderate/low risk gene mutations for hereditary breast cancer using multi-gene panel testing. Next-generation sequencing was used to analyze 255 breast cancer patients who met clinical criteria for genetic testing. Among the 104 identified pathogenic variants, 21 were found in moderate-risk genes (notably CHEK2 and ATM), three were found in low-risk genes (MSH1 and MLH1), and eight were found in genes with insufficient evidence of risk. The analysis emphasized the clinical relevance of reporting less-penetrant mutations to enhance genetic risk assessment. Findings support expanding genetic screening to improve diagnostic precision, personalize treatment strategies, and refine breast cancer risk prediction models across populations. Rezaei et al.[39], aimed to evaluate role of AI in enhancing breast cancer diagnosis and management through transcriptomic data analysis. A systematic search across databases including PubMed, Scopus, WoS, Embase, and IEEE Xplore identified 7,287 studies, of which 54 were selected for final analysis: 24 focused on RNA sequencing and 30 on GEP. The methodology involved screening by multiple reviewers and extraction of data on AI models and molecular techniques. Common AI methods included RF, CNNs, SVMs, and LASSO. These approaches demonstrated high potential in biomarker identification, prognosis prediction, and drug response optimization, though further large-scale validation and interdisciplinary research are needed.

Thâalbi and Akhloufi[40] aimed at enhancing breast cancer gene expression prediction by introducing EMGP-Net, a novel DL architecture combining EfficientFormer and MambaVision. EMGP-Net was trained using a leave-one-patient-out method on the HER2+ dataset (8 patients)[41] and validated externally on the STNet dataset (23 patients),[42] with training alternating between the two datasets. The model integrated features from both architectures using attention mechanisms and dense layers to predict the expression of 250 selected genes. Evaluation using the Pearson correlation coefficient (PCC) showed superior performance, achieving a maximum PCC of 0.7903 for the *PTMA* gene. Chowdhury and Kamal[43] aimed to develop an interpretable ML framework for classifying BRCA subtypes using RNA-sequencing data. The approach utilized the TCGA transcriptomic dataset,[37] incorporating dimensionality reduction and performing hyperparameter tuning via grid search to optimize classification models. SHAP values were employed to identify significant transcriptomic markers relevant to subtype differentiation. The model's performance was evaluated using metrics such as ACC, PRE, and FS, thereby demonstrating enhanced classification ACC and interpretability compared with existing approaches. Additionally, gene set enrichment analysis revealed key molecular pathways linked to BRCA subtypes, highlighting the method's potential to support personalized prognosis and treatment planning in clinical settings.

Nasarudin et al.[44], focused on developing an interpretable DL model for predicting breast cancer survival using METABRIC dataset.[12] The approach integrated bidirectional (BiLSTM)

and CNN architectures with minimum redundancy maximum relevance (MRMR) for feature selection. Evaluations were conducted using METABRIC (n=1980) and TCGA-BRCA (n=1080) datasets, incorporating clinical data, copy number alterations, and gene expression profiles. Performance was assessed using ACC and AUC-ROC metrics. The model achieved 98% ACC on METABRIC and 96% on TCGA, outperforming existing methods. These findings suggest the model's robustness and potential to support personalized treatment decisions in breast cancer care. Maigari et al.[45] aimed to review advancements in multimodal DL approaches for breast cancer survival prediction, focusing on architectures that integrated imaging, genomic, and clinical data. A systematic literature review was conducted using databases and search engines such as Google Scholar, Web of Science, and Scopus, from which 19 relevant studies were selected. These studies employed DL methods, particularly CNNs, to handle high-dimensional, heterogeneous data. Evaluation metrics included predictive ACC and model interpretability. Findings revealed that CNNs and hybrid models, including Graph Neural Networks, significantly improved prognostic ACC. However, gaps remain in dynamic modeling, multimodal integration, and explainability, underscoring the need for robust and interpretable solutions in PRE oncology.

Findings

This section presents the key findings derived from an extensive review of recent research on breast cancer prediction using ML and DL techniques. The studies were analyzed based on their methodologies, datasets, performance metrics, and model architectures. Emphasis was placed on understanding how different approaches handle data preprocessing, feature selection, and model evaluation. The findings are categorized to clarify trends and limitations in the existing literature. By identifying common practices and shortcomings, this section lays the foundation for recognizing research gaps and justifying the need for more robust, interpretable, and generalizable DL-based frameworks.

Literature Survey Findings

This section presents and discusses key findings from the reviewed literature; these findings are systematically summarized in Table 1. The table summarizes outcomes of various studies on breast cancer prediction, classification, and survival analysis that employed ML, DL, and hybrid techniques. It highlights the use of diverse datasets, such as METABRIC and TCGA-BRCA, and advanced models, such as LSTM, XGB, and DNN. The findings provide insights into the performance of different approaches in terms of ACC, interpretability, feature selection, and into their potential for clinical application in personalized cancer treatment.

Datasets Used

The literature review reveals that METABRIC[12], TCGA-BRCA[17], and CuMiDa[14] are the most commonly used datasets for breast cancer research and analysis, as presented in Table 2. These datasets provide extensive genomic, transcriptomic, and clinical information, making them highly valuable for developing ML and DL models focused on prediction, classification, survival analysis, and personalized treatment planning.

The METABRIC[12], TCGA-BRCA[17], and CuMiDa[14] datasets are described in detail below.

• METABRIC[12]: METABRIC is a widely used breast cancer dataset that includes clinical and genomic data from approximately 2,000 patients. It provides gene expression profiles, somatic mutation data, copy number aberrations, and survival outcomes. The dataset is instrumental in building models for prognostic analysis, metastasis prediction, and clinical feature classification.

• TCGA-BRCA[17]: TCGA-BRCA dataset contains comprehensive multi-omic profiles, including mRNA expression, DNA methylation, copy number variations, and clinical annotations for over 1,000 breast cancer patients. It supports subtype classification, survival analysis, and biomarker discovery. It is a benchmark dataset for breast cancer ML/DL research because of its size, richness, and the availability of follow-up data.

• CuMiDa[14]: The Curated Microarray Database (CuMiDa) is a collection of microarray gene expression datasets covering various types of cancer, including breast, liver, lung, prostate, and kidney. It contains over 1,000 patient samples and is primarily used for multi-class classification, dimensionality reduction, and benchmarking optimization-based ML approaches.

Performance Metrics Used

Performance evaluation plays a crucial role in assessing the effectiveness of ML and DL models in breast cancer prediction and analysis. Various studies have employed a range of metrics depending on the problem type, data balance, and model objective. The most commonly used metrics include ACC, PRE, REC, and FS, particularly for classification tasks. Other metrics, such as balanced accuracy (BAC), kappa (KAP), MCC, and AUC-ROC, are used for imbalanced datasets and multiclass classification tasks. PCC was adopted for expression-level prediction. Table 3 summarizes the performance metrics used in the reviewed studies.

Below are common performance metrics used in the literature and how they are calculated:

ACC indicates the ratio of correctly predicted observations to total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where *TP=True Postive, TN=True Negative, P=False Positive* and *N=False Negative.*

PRE: Measures how many of the positively predicted instances are actually positive.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

REC: Measures how many actual positive instances were correctly identified.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

FS: The harmonic mean of precision and recall, balancing both.

$$FS = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Balanced Accuracy (BAC): Used when data is imbalanced. It is the average of sensitivity and specificity.

TABLE 1: Literature survey findings.

Reference	Findings
11	An LSTM model achieved 98.7% accuracy using the multi-omics METABRIC dataset for survival prediction.
13	Dimensionality was reduced from 36,805 to 80 features, and balanced accuracy improved to 93.4%.
15	PCA- and MCC-based feature selection with a DNN improved classification performance.
16	RF and XGB achieved accuracies of 97.19% and 95.23%, respectively, and identified key proteins and miRNAs.
18	DIM model improved cancer-driver gene identification with strong cross-cancer generalization.
19	XGB and RF models achieved 72.7% accuracy in survival prediction using selected METABRIC features.
20	RecNN achieved an accuracy of 98.69% in predicting metastasis using LASSO-selected features.
21	SVM-RBF achieved 99.79% accuracy in predicting key clinical attributes from METABRIC data.
22	SVM performed best, with subtype-specific AUC-ROC scores of up to 100% for HER2 and Basal subtypes.
23	BCC improved subtyping accuracy and reclassified HER2-low as a distinct subtype.
25	The Gradient Boosting model achieved 86.4% test accuracy for the mastectomy class; age and relapse were key predictors.
26	KAO-AOA-SVM achieved an accuracy of 98.97% using only 15 genes to classify breast cancer.
27	XAI methods, such as SHAP, helped identify the top predictive genes; SVM and RF showed concordant rankings.
28	DL with SHAP achieved up to 100% accuracy and validated key genes, including ESR1 and KRT5.
31	DGBCox improved prognostic reliability across 12 datasets by addressing distributional shifts in the data.
32	Multimodal DL using mammograms and clinical data achieved an accuracy of 88.78%.
34	Integration of GEP with ABM improved survival prediction and tumor simulation accuracy.
35	An E2E classifier using mRNA image features achieved 99.91% accuracy.
36	The BCDD biclustering approach outperformed traditional SVD for subtype classification.
38	Identified moderate/low-risk gene mutations, supporting extended genetic screening.
39	A systematic review confirmed ML's promise in diagnosis, biomarker discovery, and treatment prediction.
40	EMGP-net achieved high gene expression prediction accuracy, with a PCC of 0.7903 for PTMA.
43	An interpretable ML model using SHAP and enrichment analysis improved BRCA subtype classification.
44	BiLSTM-CNN with MRMR achieved accuracies of 98% (METABRIC) and 96% (TCGA).
45	The review identified CNNs and hybrid DL models as top performers in multimodal survival prediction.

METABRIC: The Molecular Taxonomy of Breast Cancer International Consortium; LSTM: Long short-term memory; PCA: Principal component analysis; MCC: Mathew's correlation coefficient; XGB: eXtreme gradient-boosting; RF: Random forest; DIM: Dynamic-incentive-model; LASSO: Least-absolute-shrinkage and selection-operator; RecNN: Recursive-neural-network; SVM: Support vector machines; RBF: Radial basis function; AUC-ROC: Area under curve-receiver operating characteristic; HER2: Human epidermal growth factor receptor 2; KAO: Kashmiri-Apple optimization approach; AOA: Armadillo-optimization approach; SHAP: SHapley Additive exPlanations; DL: Deep learning; DGBCox: Deep-global balancing-cox regression; GEP: Gene expression profiling; ABM: Agent-based modelling; E2E: Ensemble-of-ensemble; BCDD: Bi-clustering differential-sparsity-constraints and dynamic-graph-regularization; SVD: Singular-value decomposition; ML: Machine learning; BiLSTM: Bi-directional long short-term memory; CNN: Convolutional neural network; MRMR: Minimum redundancy maximum relevance; TCGA: The cancer genome atlas.

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (5)$$

Kappa Coefficient (KAP): Measures agreement between predicted and actual labels while considering chance agreement.

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (6)$$

where P_0 denotes observed accuracy, P_e denotes expected accuracy by chance.

MCC: Provides a balanced measure even for imbalanced datasets.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

AUC-ROC (Area Under Curve - Receiver Operating Characteristic): Measures the ability of a classifier to distinguish between classes. Higher AUC indicates better model performance.

$$AUC - ROC = \int_0^1 TPR(FPR)dFPR \quad (8)$$

where PR=True Postive Rate, FPR=False Positive Rate.

The Pearson correlation coefficient (PCC) measures the linear correlation between predicted and true gene expression levels.

$$PCC = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (9)$$

TABLE 2: Dataset used in existing works.

Reference	Dataset used
11,15, 19, 20, 21, 23, 25, 44	METABRIC[12]
13, 28	CuMiDa[14]
16, 22, 23, 27, 30, 36, 43, 44	TCGA-BRCA[17]
24	SCAN-B
32	CMMD[33]
31	12 breast cancer datasets
38	Custom gene panel sequencing
41, 42	HER2+, STNet
29	ArrayExpress (E-MTAB-3732)
37	TCGA (multiple, including BRCA)

METABRIC: The Molecular Taxonomy of Breast Cancer International Consortium; HER2: Human epidermal growth factor receptor 2; TCGA: The cancer genome atlas; BRCA: Breast invasive carcinoma; CuMiDa: The curated microarray database; CMMD: Chinese mammography database; SCAN-B: The Sweden Cancerome Analysis Network-Breast.

x_i where and y_i are predicted and actual values respectively.

ML and DL Models and Feature Extraction and Selection Approaches Used

ML and DL techniques have been widely adopted in breast cancer research for tasks such as classification, survival prediction, and biomarker identification. These models are often complemented with feature selection and feature extraction methods to enhance performance, reduce dimensionality, and improve interpretability. Feature selection techniques, such as LASSO, MCC, and MRMR, help identify the most relevant variables, whereas extraction methods, such as PCA and autoencoders transform raw data into meaningful representations. Table 4 provides a comprehensive overview of the reviewed studies, highlighting the use of ML, DL, feature selection, and feature extraction techniques across different research efforts in this domain.

Issues and Challenges

This section discusses the issues and challenges. Table 5 summarizes the key issues and challenges identified in existing ML and DL approaches for breast cancer prediction and analysis. While many studies have demonstrated high performance, they often exhibit limitations, such as the exclusive use of ML or DL without cross-paradigm validation. Additionally, several works rely on basic or outdated feature selection methods, which may not adequately capture complex biological interactions. In some studies, feature extraction methods such as PCA or VAE, though effective for dimensionality reduction, can lead to a loss of interpretability or of biologically relevant information. Another significant issue across studies is the limited generalizability and lack of external validation, particularly when using small datasets or omics-specific models. Furthermore, imbalanced datasets, overfitting, and inconsistent benchmarking between ML and DL approaches affect the deployment of robust models in clinical settings.

TABLE 3: Performance metrics used in reviewed studies.

Reference	Performance metrics used
11, 15, 16, 19-22,25-28, 43, 44	Accuracy, precision, recall, F1-score
13	Balanced accuracy score, F1-score, precision, recall, kappa, Matthews correlation coefficient
23, 36	Area under the curve-receiver operating characteristic
31, 34, 38, 39	Custom metrics, statistical validation, interpretability-based assessments
40	Pearson correlation coefficient

TABLE 4: Overview of ML, DL, feature selection, and feature extraction usage by existing literature review.

Reference	ML	DL	Feature selection	Feature extraction
[11]	No	Yes (GCN, LSTM, VAE)	Yes (preprocessing)	Yes (VAE)
[13]	Yes (SVM, NB, DT, RF)	No	Yes (RO+entropy)	Yes (PCA, ICA, SAE)
[15]	No	Yes (DNN)	Yes (MCC-based)	Yes (PCA)
[16]	Yes (RF, SVM, DT, GNB, KNN, XGB)	No	Yes (gene analysis)	Yes (pathway & network analysis)
[18]	No	No	Yes (NFS-based DIM)	Yes (hypergraph+PPI)
[19]	Yes (XGB, RF, SVM, KNN)	No	Yes (top-k XGB features)	No
[20]	No	Yes (RecNN)	Yes (LASSO)	Yes (NER)
[21]	Yes (SVM, LR, RF, Ensemble)	No	Yes (preprocessing)	No
[22]	Yes (NB, KNN, SVM, RF)	No	Yes (LASSO, mCGA)	No
[23]	Yes	Yes (BCC)	Yes (subtype refinement)	Yes (RNA-seq embedding)
[25]	Yes (DT, XGB, SVM, LR, etc.)	No	Yes (SMOTE+feature selection)	No
[26]	Yes (SVM)	No	Yes (KAO+AOA)	Yes (gene subset search)
[27]	Yes (KNN, SVM, RF)	No	Yes (LOCI+SHAP)	No
[28]	No	Yes (DL+SHAP)	Yes (SHAP genes)	Yes (ensemble learning)
[31]	No	Yes (DAE, DGBCox)	Yes (causal feature balancing)	Yes (latent representation via DAE)
[32]	Yes (meta-classification)	Yes (DL with imaging)	Yes (metadata analysis)	Yes (CNN for images)
[34]	Yes	No	Yes (key gene identification)	Yes (agent-based modeling)
[35]	Yes (E2E: XGB, RF, AB, etc.)	Yes (AlexNet, ResNet)	Yes (SMOTE)	Yes (image transformation)
[36]	Yes	No	Yes (sparsity-based)	Yes (SVD+graph regularization)
[38]	No	No	Yes (gene panel filtering)	No
[39]	Yes	Yes	Yes (reviewed LASSO, RF, etc.)	No
[40]	No	Yes (EMGP-net)	Yes (top 250 genes)	Yes (efficientformer+mambavision)
[43]	Yes	No	Yes (SHAP+hyperparameter tuning)	Yes (dimensionality reduction)
[44]	Yes	Yes (BiLSTM+CNN)	Yes (MRMR)	Yes (multi-omic integration)
[45]	No (review)	Yes	Yes (reviewed techniques)	Yes (imaging+genomic fusion)

DL: Deep learning; ML: Machine learning; SVM: Support vector machines; GCN: Graph-convolutional networks; LSTM: Long short-term memory; VAE: Variational autoencoders; NB: Naïve bayes; RF: Random forest; DT: Decision tree; GNB: Gaussian NB; KNN: K-nearest neighbors; XGB: eXtreme gradient-boosting; LR: Logistic-regression; E2E: Ensemble-of-ensemble; AB: AdaBoost; KAO: Kashmiri-apple optimization approach; AOA: Armadillo-optimization approach; LOCI: Leaving-one-covariate-in; SHAP: SHapley additive exPlanations; SMOTE: Synthetic minority oversampling technique; CNN: Convolutional neural network; SVD: Singular-value decomposition; MRMR: Minimum redundancy maximum relevance; BiLSTM: Bi-directional long short-term memory; RO: Remora-optimization; MCC: Mathew's correlation coefficient; NFS: Network-functional-score; DIM: Dynamic-incentive-model; PCA: Principal component analysis; SAE: Sparse auto-encoder; ICA: Independent component analysis; PPI: Protein-protein interaction; LASSO: Least-absolute-shrinkage and selection-operator.

The reviewed literature reveals notable progress in breast cancer prediction using ML and DL, but also highlights critical limitations in existing approaches. Many studies rely on either ML or DL alone, missing opportunities to leverage the strengths of both. ML models often offer high interpretability but may struggle with complex, non-linear patterns in omics data. On the other hand, DL models like CNNs, RNNs, and AEs provide superior feature representation and predictive ACC, but are frequently criticized for their black-box nature and high computational complexity. Feature selection techniques are often rudimentary, leading to suboptimal model performance, whereas feature extraction methods may reduce interpretability. Moreover, several approaches demonstrate promising results in limited datasets, but fail to generalize across diverse cohorts due to inadequate validation strategies. This underscores the growing need for DL-based frameworks that not only capture high-dimensional, nonlinear patterns in multi-omic and imaging data but also integrate explainability and domain knowledge. Combining DL with advanced feature selection and robust external validation could pave the way for more accurate, interpretable, and clinically applicable cancer prediction models, ultimately contributing to personalized and PRE oncology.

TABLE 5: Issues and challenges in existing approaches.

Reference	Issues and challenges
11	DL-only method; no ML-to-DL comparison; potential loss of interpretable features during feature extraction using VAE.
13	Despite extensive feature extraction, ML-based classifiers may reach a performance plateau; BAC increased, but remained subpar for all classes.
15	PCA may ignore biologically significant features; however, the DL technique employed lacks diversity in classifiers.
16	Solely employs machine learning; fails to investigate DL models, which could more effectively capture non-linear dependencies; although gene-level network analysis is intricate, it might overlook more profound patterns.
18	Depends on network-based scoring (DIM), which might not generalize to noisy datasets; lacks DL/ML categorization.
19	The accuracy is comparatively low (72.7%); the feature-extraction approach is not robust; the feature selection is restricted to the top-k features via XGB.
20	RecNN is used, but no ML comparison is made. LASSO may fail to detect interactions among nonlinear features.
21	The study is ML-only; feature selection is straightforward, and DL is not used for deeper representation learning.
22	LASSO and mCGA were employed; however, no DL comparison was conducted. The biomarker finding was robust; however, there was no evidence of generalizability.
23	DL-based subtyping, but there isn't any obvious external validation; RNA-seq embeddings might vary depending on the dataset.
25	ML models were used; GB performed well, but test accuracy declined, suggesting overfitting.
26	DL is not integrated by ML with hybrid feature selection, which is restricted to classification without biological interpretability.
27	No DL model is employed; explainability is prioritized, yet predictive power may be weak; feature selection may overlook latent features.
28	Although the DL model is reliable, it does not integrate biological pathway information, and its generalizability has been validated only on a small number of datasets.
31	When DL is applied to causal inference, its complexity increases, and its interpretability and clinical applicability are constrained.
32	Multi-modal DL may require improved feature fusion, although its accuracy (88.78%) is lower than that of DL-only methods.
34	No DL model; ABM lacks real-time flexibility and is strong for simulation but not predictive.
35	High performance can be achieved using complex ensemble methods and DL; however, model interpretability and computational cost remain significant obstacles.
36	There is no DL; bi-clustering and graph regularization are heavily used; interpretability is good but not predictively validated.
38	Gene panel analysis may overlook new biomarkers in more recent datasets because it is not inherently predictive.
39	Review; draws attention to the lack of extensive validation across datasets and the inconsistency in ML/DL model comparison.
40	The validation dataset for the gene expression-focused DL model is modest (8 and 23 patients), raising concerns about its generalizability.
43	Strong interpretability ML model without DL benchmarking; robustness may be impacted by gene expression variability.
44	Although BiLSTM+CNN works effectively, it is complex and difficult to interpret, and MRMR selection may exclude synergistic genes.
45	Multimodal DL techniques are reviewed; however, the incorporation of dynamic patient data and explainability remain two main gaps.

DL: Deep learning; ML: Machine learning; VAE: Variational autoencoders; BAC: Balanced accuracy score; PCA: Principal component analysis; DIM: Dynamic-incentive-model; XGB: eXtreme gradient-boosting; LASSO: Least-absolute-shrinkage and selection-operator; RecNN: Recursive-neural-network; CNN: Convolutional neural network; BiLSTM: Bi-directional long short-term memory; ABM: Agent-based modelling; MRMR: Minimum redundancy maximum relevance; GB: Gradient-boosting.

CONCLUSION

Breast cancer remains one of the most critical health challenges affecting women worldwide, with early and accurate diagnosis being essential for effective treatment and improved survival rates. This work began with a comprehensive review of ML and DL approaches applied to breast cancer prediction and classification. Although numerous studies have attempted to use ML and DL models with various genomic, transcriptomic, and clinical datasets, significant limitations persist. Common issues include over-reliance on either ML or DL models, lack of generalization, inadequate feature selection or extraction techniques, and inconsistent performance metrics across datasets. The research identified key gaps such as limited integration of multi-modal data, poor interpretability, and the absence of robust, unified frameworks capable of handling

complex and high-dimensional gene expression data. In response, the problem statement was formulated to address the need for a more accurate and generalizable approach to breast cancer classification. The objectives included analyzing existing techniques, identifying their limitations, and proposing a way forward. A systematic methodology was adopted, including a literature review, dataset exploration, evaluation of performance metrics, and comparison of ML and DL models. Findings revealed that DL models generally offer superior performance but suffer from a lack of transparency and consistency when applied across different datasets. Future work will involve developing a novel DL-based framework that incorporates advanced feature extraction and selection methods. The proposed system will be trained and validated using diverse datasets, such as CuMIDA, METABRIC, and TCGA-BRCA. The goal is to accurately predict and classify various subtypes of breast cancer while ensuring high interpretability, robustness, and clinical relevance.

Footnotes

Authorship Contributions

Concept: A.S.P., Design: A.S.P., Data Collection or Processing: A.S.P., Analysis or Interpretation: P.J., Literature Search: P.J., A.S.P., Writing: P.J., A.S.P.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Sathishkumar K, Chaturvedi M, Das P, Stephen S, Mathur P. Cancer incidence estimates for 2022 & projection for 2025: result from National Cancer Registry Programme, India. Indian J Med Res. 2022;156(4&5):598-607. [Crossref] [PubMed] [PMC]
2. Tian F, Liu D, Wei N, et al. Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. Nat Med. 2024;30:1309-1319. [Crossref]
3. Zhang Y, Ji Y, Liu S, et al. Global burden of female breast cancer: new estimates in 2022, temporal trend and future projections up to 2050 based on the latest release from GLOBOCAN. J Natl Cancer Cent. 2025;5(3):287-296. [Crossref] [PubMed] [PMC]
4. Cai Y, Dai F, Ye Y, et al. The global burden of breast cancer among women of reproductive age: a comprehensive analysis. Sci Rep. 2025;15:9347. [Crossref]
5. García-Sancha N, Corchado-Cobos R, Pérez-Losada J. Understanding susceptibility to breast cancer: from risk factors to prevention strategies. Int J Mol Sci. 2025;26(7):2993. [Crossref] [PubMed] [PMC]
6. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. 2022;66:15-23. [Crossref] [PubMed] [PMC]
7. Abu Abeelh E, AbuAbeileh Z. Comparative effectiveness of mammography, ultrasound, and MRI in the detection of breast carcinoma in dense breast tissue: a systematic review. Cureus. 2024;16(4):e59054. [Crossref]
8. Carriero A, Groenhoff L, Vologina E, Basile P, Albera M. Deep learning in breast cancer imaging: state of the art and recent advancements in early 2024. Diagnostics (Basel). 2024;14(8):848. [Crossref] [PubMed] [PMC]
9. Jiang B, Bao L, He S, Chen X, Jin Z, Ye Y. Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis. Breast Cancer Res. 2024;26(1):137. [Crossref] [PubMed] [PMC]
10. Khan K, Awang S, Talab MA, Kahtan H. A comprehensive review of machine learning and deep learning techniques for intraclass variability breast cancer recognition. Franklin Open. 2025;11:100296-100296. [Crossref]
11. Mahmoud A, Alhussein M, Aurangzeb K, Takaoka E. Breast cancer survival prediction modeling based on genomic data: an improved prognosis-driven deep learning approach. IEEE Access. 2024;(12):119502-119519. [Crossref]
12. Breast Cancer Gene Expression Profiles (METABRIC). Available from: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric (accessed 06 Aug. 2025) [Crossref]
13. Bharanidharan N, Sanassi Chakravarthy SR, Vankatesan VK, Abbas M, Mahesh TR, Mohan E. Local entropy based remora optimization and sparse autoencoders for cancer diagnosis through microarray gene expression analysis. IEEE Access. 2024;(12)39285-39299. [Crossref]
14. Feltes BC, Chandelier EB, Grisci BI, Dorn M. CuMiDa: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. J Comput Biol. 2019;26(4):376-386. [Crossref] [PubMed]
15. Kishore Khan S, Kanamarlapudi A, Singh AR. Diagnosis and classification of breast cancer using data visualization and deep learning techniques. 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India. 2024;1-6. [Crossref]
16. Das SC, Tasnim W, Rana HK, Acharjee UK, Islam MM, Khatun R. Comprehensive bioinformatics and machine learning analyses for breast cancer staging using TCGA dataset. Brief Bioinform. 2024;26(1):bbae628. [Crossref] [PubMed] [PMC]
17. Cancer.gov, 2024. Available link: https://portal.gdc.cancer.gov/projects/TCGA-BRCA
18. Hu Z, Li G, Luo X, et al. Identification of cancer driver genes based on dynamic incentive model. IEEE/ACM Trans Comput Biol Bioinform. 2024;21(6):2371-2381. [Crossref] [PubMed]
19. Kurniadi FI, Saputri HA. Feature selection using XGBoost on METABRIC dataset for survivability breast cancer detection. 2024 International Conference on Information Management and Technology (ICIMTech), Bali, Indonesia. 2024;259-262. [Crossref]
20. Brahmatej Rupavath RVSS, Shnain AH, Pushpalakshmi GRG, Hemalatha K. A deep learning based metastasis prediction system using multi omics data and recursive neural network. 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkuru, India. 2024;1-5. [Crossref]
21. Puttegowda K, Kumar DA, Ravi V, et al. Advanced machine learning techniques for prognostic analysis in breast cancer. The Open Bioinformatics Journal. 2025;18(1). [Crossref]
22. Ghosh N, Kumar Mridha S, Paul R. LASSO-mCGA: machine learning and modified compact genetic algorithm-based biomarker selection for breast cancer subtype classification. IEEE Access. 2025;13:17673-17682. [Crossref]

23. Turova P, Kushnarev V, Baranov O, et al. The breast cancer classifier refines molecular breast cancer classification to delineate the HER2-low subtype. NPJ Breast Cancer. 2025;11(1):19. [Crossref] [PubMed] [PMC]
24. Saal LH, Vallon-Christersson J, Häkkinen J, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med. 2015;7(1):20. [Crossref] [PubMed] [PMC]
25. Asfaw BB, Tegaw, EM. Explainable machine learning to compare the overall survival status between patients receiving mastectomy and breast conserving surgeries. Scientific Reports. 2025;15(1). [Crossref]
26. Yaqoob A, Verma NK. Feature selection in breast cancer gene expression data using KAO and AOA with SVM classification. J Med Syst. 2025;49(1):40. [Crossref] [PubMed]
27. Kallah-Dagadu G, Mohammed M, Nasejje JB, et al. Breast cancer prediction based on gene expression data using interpretable machine learning techniques. Sci Rep. 2025;15(1):7594. [Crossref] [PubMed] [PMC]
28. Aliouane SE, Chehili H, Boulahrouf K, Abdelaziz A, Khlifa N, Hamidechi MA. Integrating deep learning and SHAP for breast cancer classification and biomarker discovery using gene expression data. IEEE Access. 2025;13:49693-49709. [Crossref]
29. Signol F, Arnal L, Navarro-Cerdán JR, Llobet R, Arlandis J, Perez-Cortes JC. SEQENS: an ensemble method for relevant gene identification in microarray data. Comput Biol Med. 2023;152:106413. [Crossref] [PubMed]
30. "MalaCards - human disease database," Malacards.org, 2017. Available link: https://www.malacards.org/
31. Li X, Liu L, Li J, Le TD. Stable breast cancer prognosis. in IEEE Transactions on Computational Biology and Bioinformatics. 2025;22(2):721-731. [Crossref]
32. Rabah CB, Sattar A, Ibrahim A, Serag A. A multimodal deep learning model for the classification of breast cancer subtypes. Diagnostics. 2025;15(8):995-995. [Crossref]
33. NgX T. The Chinese Mammography Database (CMMD) 2022. [Crossref]
34. Sridharan P, Ghosh M. Gene expression and agent-based modeling improve precision prognosis in breast cancer. Scientific Reports. 2025;15(1). [Crossref]
35. Kunta JPKC, Lepakshi VA. Deep residual transfer ensemble model for mRNA gene-expression-based breast cancer. IEEE Access. 2025;13:105608-105628. [Crossref]
36. Li D, Song P, Wang J. A novel biclustering algorithm based on differential sparsity constraints and dynamic graph regularization for cancer gene expression data. IEEE Access. 2025;13:94681-94695. [Crossref]
37. National Cancer Institute. The Cancer Genome Atlas Program (TCGA) – NCI. Available link: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [Crossref]
38. Goidescu IG, Rotar IC, Nemeti G, et al. Moderate-low risk breast cancer gene expression in a romanian population. Int J Mol Sci. 2025;26(11):5313. [Crossref] [PubMed] [PMC]
39. Rezaei S, Hamedani Z, Ahmadi K, et al. Role of machine learning in molecular pathology for breast cancer: a review on gene expression profiling and RNA sequencing application. Crit Rev Oncol Hematol. 2025;213:104780. [Crossref] [PubMed]
40. Thâalbi O, Akhloufi MA. EMGP-net: a hybrid deep learning architecture for breast cancer gene expression prediction. Computers. 2025;14(7):253-253. [Crossref]
41. Andersson A, Larsson L, Stenbeck L, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. Nat Commun. 2021;12(1):6012. [Crossref] [PubMed] [PMC]
42. He B, Bergenstråhle L, Stenbeck L, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng. 2020;4(8):827-834. [Crossref] [PubMed]
43. Chowdhury TM, Kamal AMR. An efficient and interpretable machine learning model for classifying breast cancer subtypes using gene expression profiles. Engineering Technology & Applied Science Research. 2025;15(4):24196-24203. [Crossref]
44. Nasarudin NA, Al-Jasmi F, Abdul Aziz NH, et al. An improved deep learning algorithm for breast cancer survival prediction based on multi-omics data. F1000Research. 2025;14:765-765. [Crossref]
45. Maigari A, XinYing C, Zainol Z. Multimodal deep learning breast cancer prognosis models: narrative review on multimodal architectures and concatenation approaches. Journal of Medical Artificial Intelligence. 2025;8:61-61. [Crossref]