**Advances in Artificial Intelligence and Machine Learning**; Review 6 (1) 5023-5048 Received 18-11-2025; Accepted 12-02-2026; Published online 19-02-2026

# Machine Learning and Statistical Approaches for Predicting Breast Cancer Recurrence and Metastasis: A Systematic Review

**Charanpreet Kaur** charanpreet27@gmail.com
*Department of SET, MRIIRS Faridabad*
*Haryana, India*

**Rosy Madaan** rosymadaan.set@mriu.edu.in
*Department of SET, MRIIRS Faridabad*
*Haryana, India*

**Corresponding Author:** Charanpreet Kaur

## Abstract

Metastasis and recurrence of breast cancer are the most important factors that are affecting patients' quality of life and long-term survival globally. The development of machine learning (ML) and statistical models have been extensively used in the healthcare sector in predicting recurrence and metastatic risk. This has been facilitated by enhancements in data accessibility via various datasets from public cancer repositories and population-based registries. The validity, comparability, and clinical utility of these prediction models have not been comprehensively integrated, and the data remains disjointed despite considerable methodological advancements. This systematic review aims to critically evaluate the machine learning and statistical methodologies employed to predict breast cancer recurrence and metastasis utilizing secondary data, concentrating on data sources, modeling techniques, outcome definitions, validation strategies, and reported clinical utility.Using PubMed, Scopus, Web of Science, and IEEE Xplore, a complete literature review was done according to the PRISMA principles. In order to create and test the different ways to predict the reoccurrence and spread of breast cancer, secondary datasets like TCGA, METABRIC, SEER, GEO, and other cancer registries were used. The data extraction mechanism included study's design, the characteristics of the cohort, the predictor variables, the modeling strategies, the performance measures, and the validation techniques.The study included a wide range of statistical models, like Cox proportional hazards models and logistic regression, diverse ML techniques such as random forests, support vector machines, gradient boosting, and deep learning architectures. Most of the models under study combined the genetic or transcriptome attributes with the clinicopathological factors whereas only a few explored the multi-modal or image-based methods. The predictive performance varied substantially across different studies, with frequent reliance on the internal validation and limited use of external or prospective validation. The analyses of the review showed that the results for recurrence and metastasis in breast cancer were heterogeneous, mainly in registry-based studies, complicating the cross-study comparison. While most of the models showed medium to high discriminatory capability, but calibration, precision, accuracy and clinical significance were rarely documented. Machine learning and statistical algorithms showed a great future in predicting the recurrence

**Citation:** Charanpreet Kaur and Rosy Madaan. Machine Learning and Statistical Approaches for Predicting Breast Cancer Recurrence and Metastasis: A Systematic Review. **Advances in Artificial Intelligence and Machine Learning.** 2026;6(1):279.
https://dx.doi.org/10.54364/AAIML.2026.61279

and metastasis of breast cancer using the secondary data, even though their application is limited by methodological errors, inadequate validation, and unclear clinical relevance. To enable responsible and successful adoption of these strategies in early cancer diagnosis, the future research should prioritize consistent outcome definitions, open reporting, thorough validation from outside, and integration of medical decision-making considerations.

**Keywords:** Recurrence, Metastasis, Machine learning, Statistical models, Secondary data, Systematic review, Breast cancer

## 1. INTRODUCTION

Despite the significant advancements in screening, diagnosis, and its treatment, the breast cancer continues to be the most common disease among women worldwide and a major cause of cancer-related mortality [1]. Although overall survival has greatly risen due to the advancements in its early identification and the targeted therapy, metastasis and the probability to re occur still present important therapeutic hurdles. Increased morbidity, a lower quality of life, and significantly worse survival outcomes are linked to distant metastasis, local recurrence, and regional relapse [2]. As a result, one of the main goals of breast cancer treatment and research is to accurately forecast the recurrence and metastatic progression in the patients suffering from breast cancer.

The clinicopathological characteristics, like tumor size, lymph node involvement, histological grade, estrogen count and progesterone receptor status, HER2 expression, and proliferation indices like Ki-67, are the key parameters of early diagnosis of the breast cancer [3]. These elements serve as the foundation for risk in suffering from the disease and its staging systems and direct therapeutic choices in the standard clinical practice. Despite their clinical value, these methods cannot be used to obtain the molecular heterogeneity of breast cancer and the intricate relationships that define the disease's timeframe. The limitations for the traditional predictive models are highlighted by the fact that significantly varied outcomes for patients with comparable clinicopathological profiles have been observed .

Breast cancer is a disease affecting several types of molecules, therefore the data-driven predictive models have become more prevalent. The improvements in high-throughput technology and the creation of large cancer databases have made it possible to obtain access to the secondary data, such as clinical, molecular, imaging, and longitudinal outcome data [4]. There are many publicly available resources that have been used a lot in predictive modeling studies. These include the Cancer Genome Atlas (TCGA), the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), the Surveillance, Epidemiology, and End Results (SEER) program, and the Gene Expression Omnibus (GEO) [5, 6]. These datasets have made it easier to use both traditional statistical methods and new machine learning (ML) methods to predict early disease diagnosis and slow its spread in the body.

Predictive modeling in the field of oncology depends on the different statistical mortality models, like Cox proportional hazards regression . The Cox models and their applications offer breast cancer free survival because they have more interpretable hazard estimates and time-to-occurrence analysis [7]. These models, however, may not accurately represent the nonlinear and multidimensional nature of the biological data since they depend on the assumptions of proportional hazards and the linear connections between variables and outcome risk [8]. To handle multicollinearity and high-dimensional feature spaces, especially in genomic applications, penalised regression algorithms like LASSO and elastic net have been developed [9]. These techniques reduce the overfitting and enhance the variable selection, but they are still limited in their capacity to represent the intricate feature correlations.

Pattern Recognition and accuracy along with precision is the most important aspect of Machine learning algorithms. Decision Tree, Support vector machines, random forests, gradient boosting machines, and neural networks are among the algorithms that are being used more frequently to predict the breast cancer results [10]. These methods are extensively used for evaluating multi-dimensional chemical and imaging data because they may find nonlinear associations, interactions among predictive variables, and data structures that are hierarchical [11]. Deep learning frameworks have significantly improved modeling capabilities by making it easier to learn the features automatically from the raw data. This is especially true for the datasets involving imaging-based prediction and histopathology [12].

In order to create machine learning (ML)-based algorithms for breast cancer recurrence and metastasis, the authors have demonstrated a wide range of data types, that included gene expression profiles, somatic mutation data, copy number variations, radiomic features, and digitized histopathological images [2]. There were many studies that have aimed at enhanced discrimination compared to the traditional statistical models, usually while using multi-modal data sources. However, methodological heterogeneity makes it difficult to evaluate and compare studies, and stated performance benefits are inconsistent.

One of the biggest problems in this field is figuring out what outcomes mean and how to label them. Recurrences can happen in the same area, in a different area, or even far away, and each has its own medical effects. Depending on the dataset, metastasis is the spread of cancerous cell in the body that can be found at any time during diagnosis, during follow-up, or based on survival goals of the patient suffering from breast cancer [13]. SEER, the Cancer repository dataset, keeps lot of information about metastasis and the body part to which it is spreading, but they usually do not have the track of recurrence events after treatment [14]. Even the Molecular datasets may report about the disease-free survival, but they may not give details about the kind of recurrence being diagnosed. This kind of variability makes the prediction models less reliable and less useful, and it also adds noise to the labels [15].

It is very important that in the healthcare sector, Interpretability and transparency plays a very crucial role in AI to take critical decisions for reliable patient care. People often criticize complex ML and deep learning models for being “black boxes,” which can make it harder for the clinicians to trust them, get regulatory approval, and use them ethically [16]. Feature attribution approaches and surrogate modeling are some of the ways to make artificial intelligence more understandable, However, these methods are still not properly standardized in the present research that try to predict breast cancer recurrence and metastasis in the patients suffering from breast cancer.

There is a gap between the development of the scientific model and to make these models work in real time. Very few authors have examined at how predictive models affect the doctors’ decisions, treatment choices, or patient outcomes. Decision curve analysis, cost-effectiveness evaluation, and prospective validation are infrequently utilized, constraining the understanding of real-world implications [17]. The evolving regulations and reporting frameworks for AI in healthcare under-score the necessity of standardized guidelines such as TRIPOD-AI and CONSORT-AI to promote transparency and reproducibility [18].

There is more and more research on prognostic modeling in breast cancer, but there still isn't enough research that uses both machine learning and statistical methods to predict recurrence and metastasis using secondary data. Most of the time, the literature review that already exist only look at molecular signatures or clinical prognostic factors [19]. They don't look at algorithmic methods, validation strategies, or the problems with data sources in a systematic way. As a result, there are still big gaps in our understanding of the strengths, weaknesses, and readiness of AI-driven models for clinical translation.

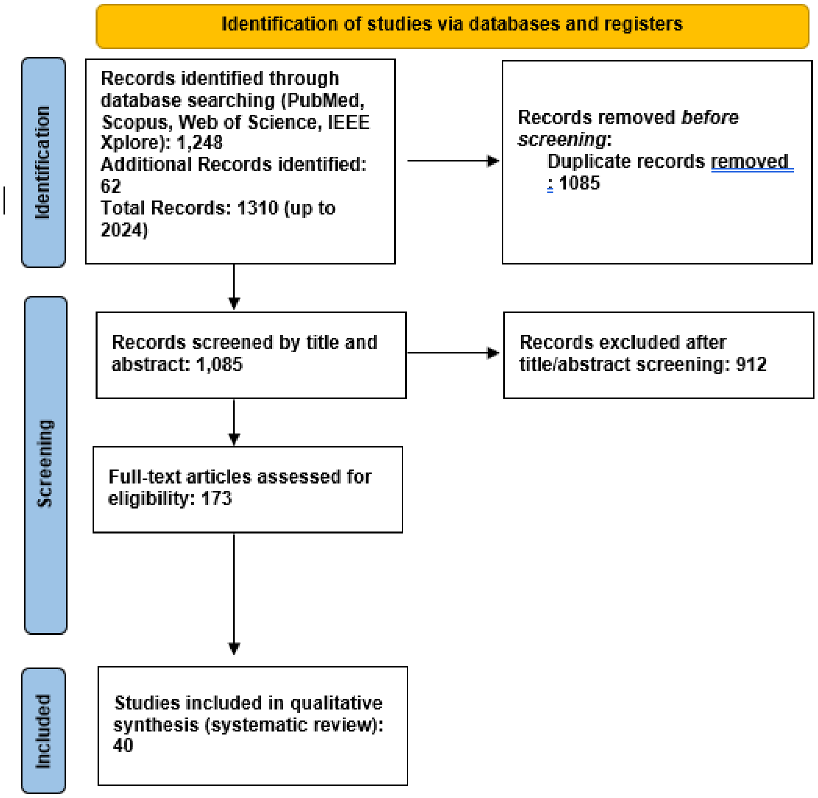
This systematic review aims to fill the gaps by combining the information received from the secondary datasets on machine learning and statistical methods used for the prediction of breast cancer recurrence and metastasis. This review aims to give a structured, well organized and critical overview of the current scenario of the healthcare sector by looking at varied data sources, feature extraction techniques, modeling methods, performance evaluation practices, and its validation frameworks. The goal of the study is to help future research, guide best practices for predictive models, and help the responsible use of AI-driven prognostic models in breast cancer care.

## 2. OBJECTIVES OF THE STUDY:

1. To systematically identify and analyze the published work that employ machine learning and statistical methods to predict recurrence and metastasis in the breast cancer patients using secondary data sources.
2. To explore the different types of secondary datasets (like cancer registries, public molecular databases, and imaging repositories) used in predictive modeling and validate them for prediction of breast cancer recurrence and metastasis.
3. To evaluate the performance parameters of reporting and validating models, such as discrimination, calibration, and internal versus external validation methods.
4. To find methodological limitations, sources of bias, and research gaps, and to bridge the gap to develop robust and clinically relevant AI-based prediction models.

## 3. METHODOLOGY

This study defines the **systematic literature review** methodology to identify, evaluate and combine the recent work done in the field of machine learning and statistical approaches for predicting breast cancer recurrence and metastasis in the patients suffering from breast cancer. The review was conducted by keeping in mind the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-2020) to make sure it was clear, reproducible, and methodologically sound as shown in FIGURE 1.


Identification of studies via databases and registers
Identification
Records identified through database searching (PubMed, Scopus, Web of Science, IEEE Xplore): 1,248
Additional Records identified: 62
Total Records: 1310 (up to 2024)
Records removed *before screening*:
Duplicate records removed : 1085
Screening
Records screened by title and abstract: 1,085
Records excluded after title/abstract screening: 912
Full-text articles assessed for eligibility: 173
Included
Studies included in qualitative synthesis (systematic review): 40

PRISMA 2020 flow diagram for Systematic Review Process

Figure 1: PRISMA 2020 flowchart of the Methodology

### 3.1 Review Design and Protocol

The review was created according to the predefined protocol including the research goals and objectives, eligibility criteria, search strategy, procedure for sample selection, data extraction structure, and synthesis methodology. The protocol was created before choosing the studies to reduce the selection bias and analytical drift. The aim of the study is to focus on the recent articles published in the last five years from 2019 to 2024, to include current advancements in machine learning, artificial intelligence, and data-driven prognostic prediction of breast cancer recurrence and metastasis.

### 3.2 Information Sources and Search Strategy

In order to ensure extensive coverage of medical, computational, and transdisciplinary research, comprehensive literature search was conducted across the different electronic databases. The data repositories that were searched were IEEE Xplore, Scopus, PubMed/MEDLINE, and the Core Collection of the Web of Science.

The search method combined controlled vocabulary terms with free-text keywords about breast cancer, recurrence, metastasis, machine learning, and statistical modeling.

## 3.3 Eligibility Criteria

Studies were selected based on predefined **inclusion and exclusion criteria**.

### 3.3.1 Inclusion Criteria

- Peer-reviewed journal articles published between 2019 and 2024
- International studies involving human breast cancer populations
- Studies developing or validating **machine learning or statistical models**
- Prediction of **breast cancer recurrence and/or metastasis** as a primary or secondary outcome
- Use of **secondary data**, including clinical cohorts, cancer registries, public molecular databases, or imaging repositories
- Articles published in English

### 3.3.2 Exclusion Criteria

- Case reports, editorials, commentaries, letters, or conference abstracts without full papers
- Studies focusing solely on diagnosis or treatment response without recurrence/metastasis outcomes
- Animal or in vitro studies
- Descriptive statistical analyses that don't use AI predictive modelling
- Studies lacking adequate methodological detail or ambiguous result definitions

## 3.4 Study Selection Process

All obtained records were imported into a reference management system and duplicates deleted. Two stages were used to choose studies:

1. To exclude unnecessary papers, titles and abstracts were independently evaluated against eligibility criteria.
2. The full texts of possibly qualifying articles were reviewed for final inclusion. We settled any discrepancies by discussion.

### 3.5 Data Extraction

The first step was to create a structured data extraction form to ensure that that data was collected uniformly and consistently across all the investigations. Research authors, publication year, and country for each study was documented. Study design, data sources, and clinical, registry-based, omics, imaging, or multimodal data were then extracted. Specific outcomes including local, regional, or distant recurrence, metastasis, and survival endpoints like DFS, RFS, and DMFS were recorded, along with sample size and the number of occurrences. The feature selection attributes, engineering procedures, and validation methodologies like internal, external, cross-cohort, or temporal were recorded along with the performance parameters like AUC, C-index, and calibration measures. Evaluating interpretability methodologies and summarizing the results are important limitations and potential biases that were identified by the authors.

### 3.6 Quality Appraisal and Risk of Bias Assessment

The methodological quality appraisal and bias risk were considered very important aspects for the existing prediction model frameworks. The fundamental concepts from PROBAST were used and reporting issues from TRIPOD-AI were addressed [18]. A qualitative assessment was performed on each study across the different methodological characteristics to ensure accuracy and reliability. The definitions and monitoring of the outputs were examined to determine whether recurrence, metastasis, or survival endpoints were distinctly documented, clinically relevant, and consistently measured. The predictor selection focussed on clinical relevance, justification for inclusion, preprocessing methodologies, and strategies were implemented to manage the missing data and inconsistent data, including imputation techniques and the recording of data absence.

### 3.7 Data Synthesis and Analysis

Since the datasets were heterogeneous, and there were diverse modeling approaches, a narrative and tabular synthesis approach was adopted rather than the simple meta-analysis. Studies were grouped and analyzed based on:

- Data modality (clinical, omics, imaging, multimodal)
- Type of predictive model (statistical vs machine learning vs deep learning)
- Outcome formulation (binary vs time-to-event)
- Validation rigor and reporting quality

## 4. RESULTS AND DISCUSSION

The results of this study are presented as a structured synthesis of empirical data from the previously published studies, in accordance with PRISMA-2020 guidelines for the systematic reviews. It summarizes and organizes the results of the literature review, which include the types of studies, datasets, modeling techniques, outcome definitions, validation strategies, and performance measures. Comparative analysis focused on identifying methodological trends, performance patterns, strengths, limitations, and gaps across the literature.

### 4.1 Data Sources and Dataset Characteristics

In the last five-year predictive modeling has been used widely on diverse range of datasets. **Clinical and Registry Data:** Many studies use institutional clinical datasets or cancer registries, containing demographics of the patients, tumor characteristics, details of the treatment, and results. For example, a 2023 study developed a model from the record of 1,131 hospital patients (2001–2022), that include parameters like like tumor size, lymph node status, grade, ER/PR/HER2 status, Ki-67, and treatment details. Large repositories of data such as SEER have been used to build recurrence risk models, though SEER dataset lacks explicit recurrence data and often uses survival or secondary malignancies as proxies.

**Molecular Data:** Public genomic databases are more frequently used for the analysis of data. The Cancer Genome Atlas (TCGA) and METABRIC are prominent sources, that provide gene expression (RNA-seq, microarray) and clinical follow-up information. A recent study used METABRIC ($n \approx 980$) as a training cohort to derive a 12-gene immune-related signature, then validated it on TCGA and multiple Gene Expression Omnibus (GEO) cohorts. These datasets typically contain distant metastasis or relapse-free survival outcomes, enabling training of prognostic models.

**Imaging Data:** There is growing use of radiological and pathology imaging in recurrence prediction [20]. Some models incorporate radiology, e.g. breast MRI or mammography, extracting quantitative radiomic features. A 2024 multi-center study used preoperative DCE-MRI from 466 patients in a training cohort (FUSCC) and tested on external cohorts from Duke (n=619) and I-SPY1 (n=128). Similarly, digital pathology whole-slide images (WSIs) have been exploited: one 2021 study collected H&E-stained WSIs for 127 HER2-positive patients and combined them with clinical data. Another 2023 study used >1,000 WSIs from a hospital and external TCGA slides to train a multiple-instance learning model for HER2-positive cases.

**Other Data Sources:** Many new data types are also being explored and under study. For instance, blood-based biomarkers like circulating tumor DNA or, intriguingly, the blood microbiome have been studied. A 2025 study of 96 breast cancer patients profiled 16S rRNA blood microbiome sequences and found that microbial genera could distinguish those who later had recurrence/metastasis (8 patients) with AUC ∼0.94 using the technique of random forests. Overall, the recent literature review spans diverse data ranging from structured clinical variables and omics like gene expression, and mutations to unstructured imaging data and emerging biomarkers, thereby offering complementary information for predicting the recurrence risk [21].

### 4.2 Modeling Approaches: Statistical Vs. Machine Learning

The methodologies for risk modeling have evolved from traditional statistical models to complex machine learning (ML) algorithms [22]. A recent systematic review of 23 studies (2008–2022) found about 30% used classic statistical approaches while ∼70% employed ML-based methods. **Statistical Models:** Many studies still use Cox proportional hazards regression for time-to-recurrence analysis. Cox models remain popular for their interpretability in estimating hazard ratios and handling censored survival data [23]. Variants like competing-risks regression (to account for death as a competing event) and piecewise Cox models have also been applied in recurrence modeling. Several groups have presented results as nomograms – e.g. a multivariable logistic model for 5-year local recurrence was turned into a nomogram and externally validated on 264 patients. Nomograms provide an interpretable tool for clinicians, and at least four recent studies offered nomogram-based risk calculators.

**Traditional Machine Learning:** ML classifiers such as support vector machines (SVM), decision trees, naive Bayes, logistic regression (as an ML algorithm), and k-nearest neighbors have been frequently used [24]. These models often incorporate more complex interactions and nonlinearities than Cox models. According to the review, traditional ML models were among the most common approaches (used by ∼26% of studies). For example, an SVM with feature selection was used in a 2021 study for invasive breast cancer recurrence prediction.

**Ensemble Methods:** Ensemble models (e.g. random forests and gradient boosting like XGBoost) have gained traction and achieved strong results. Ensembles can improve accuracy by aggregating multiple learners. A 2020 study used an XGBoost ensemble combined with case-based reasoning to predict recurrence, and many others report ensembles as top performers. In fact, ensemble learning methods have achieved some of the highest discrimination (with reported AUCs up to ∼0.94) in recent studies.

**Deep Learning:** Deep learning approaches (another ∼26% of models in the review) are increasingly prominent [25]. These include deep neural networks for survival analysis and convolutional neural networks (CNNs) for imaging data. For instance, a deep multi-layer perceptron was used on clinical data for ER-/HER2- patients (2020). Recurrent neural networks (RNNs) have been applied to model time-to-event data (e.g. a 2021 study used a Weibull Time-to-Event RNN for adjuvant therapy patients). CNNs are especially popular for pathology and radiology images: one multimodal 2021 CNN (Xception architecture) was trained on histopathology slides to predict relapse. Notably, Yang *et al.* [26], combined a ResNet-based CNN on WSIs with clinical features for HER2+ tumors, achieving an internal AUC of 0.76 and an external AUC of 0.72 on TCGA images. Deep learning and ensemble models have delivered the highest accuracy in many comparisons, although they sometimes sacrifice interpretability. Overall, the field has shifted toward integrating multiple algorithms – for example, one 2023 study built a stacking ensemble of five ML classifiers (logistic, SVM, RF, XGBoost, k-NN) and found the ensemble outperformed any single model (ensemble AUC 0.817 vs best single 0.711). This trend reflects an emphasis on maximizing predictive performance by leveraging diverse modeling approaches.

### 4.3 Recurrence and Metastasis Outcomes

A fundamental aspect of this research is the manner in which the terms “recurrence” or “metastasis” are defined. Definitions vary, which complicates comparisons. Many researchers focus on distant metastasis (distant recurrence) as the primary outcome, since distant relapse is most linked to mortality. For example, “distant recurrence” was defined in one cross-institutional study as metastasis to distant organs after initial treatment. Other studies predict any breast cancer recurrence (combining local/regional recurrence and distant metastasis) within a certain follow-up period. Yang *et al.* [26], explicitly categorized recurrence events as either locoregional (in ipsilateral breast/chest wall or regional lymph nodes) or distant metastases, and defined *recurrence-free survival (RFS)* from surgery to the first recurrence or death. Some works use *disease-free survival (DFS)* similarly. In clinical datasets, local and distant recurrence may be combined if detailed breakdown is unavailable; for instance, a hospital registry study defined "recurrence" broadly as any cancer return after initial therapy (excluding patients with metastasis at baseline). Time horizons also differ: certain models treat recurrence as a time-to-event (survival analysis), whereas others simplify to a binary outcome within 5 or 10 years. For example, radiogenomic studies correlating with Oncotype DX consider high 10-year recurrence risk (Oncotype score) as a positive outcome. Meanwhile, genomic signature studies often dichotomize patients into "short-term relapse" vs "long-term survivor" groups. The 2025 immune gene signature study divided METABRIC and GEO patients into short-term (recurrence within ~5 years) and long-term survival groups for model development. It's worth noting that metastasis prediction is sometimes studied separately – e.g. models predicting site-specific distant metastases (brain, bone, etc.) have been developed using competing-risk frameworks. In summarize, the outcome range varies from locoregional recurrence to distant metastasis or any relapse, and from *binary classification* (recurrence yes/no by a time point) to *survival predictions* (hazard over time). The variability in these endpoints specifies the importance of clearly defining the outcomes when comparing model performance.

### 4.4 Key Predictors and Feature Engineering Techniques

The key attributes that are crucial for the recurrence and metastasis model are the clinical features, molecular markers, and derived factors, with careful feature extraction techniques to handle high-dimensional data. On the clinical side, almost all models include classic prognostic factors: patient age, tumor size, lymph node status, tumor grade, hormone receptor (ER/PR) status, HER2 status, and sometimes Ki-67 index and treatment details [27]. These features are well-known from clinical risk tools, and ML models often reaffirm their importance. For instance, an ensemble ML model in 2023 selected six top features – nodal positivity, ER status, Ki-67, lymphovascular invasion, tumor size, and age – which align closely with traditional risk factors. Beyond these, therapy information (e.g. type of surgery, endocrine or chemo administered) can improve predictions of recurrence under specific treatments [28]. Some models for hormone receptor-positive disease include whether endocrine therapy was given, since recurrence risk differs by therapy duration.

**Molecular predictors** have been extensively explored, especially gene expression signatures. High-dimensional genomic data require feature reduction. Common approaches include univariate filtering and regularized regression. The development of new gene signatures often involves: (1)) identify genes associated with recurrence via differential expression or Cox analysis, (2)) apply a selection algorithm like LASSO or stepwise Cox. In one systematic approach, researchers screened prognostic genes by *GSEA* and Cox, then applied a combination of 117 machine learning algorithms under LOOCV to derive an optimal gene model. The winning model (Stepwise Cox + Ridge regression) narrowed down to 55 genes, then to a final 12-gene signature. Regularization techniques like LASSO (L1) are popular to avoid overfitting with many features. For example, an MRI radiomics study used LASSO to select the most prognostic texture features from hundreds extracted, yielding a 13-feature radiomic signature that was then used in a Cox model. Similarly, LASSO was used in a clinical ML study to whittle 16 variables down to the 6 most predictive. Other feature selection methods include tree-based importance (for random forests), principal component analysis (less common in recent works), and filter methods (e.g. selecting top features by univariate p-value). Some deep learning models bypass manual selection by letting the network learn latent features (as in autoencoders or CNN feature maps). An example is the use of CNNs on whole-slide images: instead of pre-specifying image features, the model "learns" features from image pixels. However, even in such cases, hybrid strategies exist – e.g. a 2023 multiple-instance learning model for pathology first clustered image patches and then used attention networks to aggregate features before combining with clinical data.

**Hand-crafted features** remain important for non-image data. Radiomics pipelines often extract dozens of shapes, intensity, and texture descriptors from tumors on imaging. One radiomics study predicting ER+/HER2– recurrence extracted features not only from the tumor region on MRI but also from the peri-tumoral tissue, demonstrating that peritumoral texture contributes predictive signal [29]. For digital pathology, earlier studies have used color histogram and textural features from H&E slides. Liu *et al.* [30], took this approach: they computed handcrafted color/texture metrics from histology images and combined them with machine learning classifiers (like XGBoost and RF) to predict recurrence/metastasis. This yielded decent accuracy (reportedly outperforming some deep learning while keeping computational cost low), illustrating that classical feature engineering still has a role, especially when data are limited.

**Class Imbalance** in recurrence data (often <30% of patients recur) has prompted techniques like resampling. Synthetic Minority Oversampling Technique (SMOTE) and its variants have been employed to generate additional minority-class examples. In one 2025 microbiome study, only 8 of 96 patients had recurrence/metastasis; SMOTE was crucial to train models, and a leave-one-out cross-validation was used alongside to maximize use of scarce events. Likewise, the 2023 clinical model with 75 recurrence events vs 1056 non-events applied a hybrid SMOTE-ENN method to both oversample and clean the majority class. This helped prevent the model from simply predicting "no recurrence" for all cases.

**Feature integration** is very important aspect and most of the powerful models tend to fuse multiple feature types (clinical, genomic, imaging) to capture tumor heterogeneity. A recent tri-omics model for invasive ductal carcinoma is illustrative: it combined (a) WSI image features, (b) a set of gene expression features related to cancer-associated fibroblasts (CAFs), and (c) a clinical feature (lymph node status). The integrated Cox model significantly outperformed single-modality models (AUC 0.897 for the fused model vs ∼0.77 for single WSI or gene signature). Similarly, the HER2+ CNN by Yang *et al.* [26], merged image-based deep features with patient data, and the MRI-based MDL model merged learned MRI features with clinicopathologic inputs. These examples highlight that feature engineering often extends to deciding how to combine disparate features (e.g. concatenating feature vectors vs. late fusion of model predictions). In summary, researchers employ a variety of techniques – from classical statistical selection to modern deep feature learning – to distill predictive signals from high-dimensional data, while handling challenges like multicollinearity, small event counts, and heterogeneity of data types.

### 4.5 Validation Practices and Performance Metrics

Since models can cause overfitting in complex datasets, therefore model validation is very important in the prediction models. Internally, most studies perform some form of cross-validation or train-test split. K-fold cross-validation (often 5- or 10-fold) on the training cohort is standard for tuning models. For example, the ensemble model by Wang *et al.* [6], used a 70/30 train-test split with 10-fold cross-validation on the training portion for feature selection and model tuning. Many deep learning efforts also set aside an internal validation set (e.g. 70/15/15 train-val-test split) to monitor performance during training and avoid overfitting. External validation on independent cohorts has been reported in a minority of studies, but it's becoming more common. The 2023 systematic review noted only 7 of 23 models had external validation. Recent examples of external testing include: validating a HER2+ histopathology model on TCGA slides (independent set of 123 patients), testing a radiomics signature trained in China on two external cohorts in the U.S. (achieving significant prognosis stratification in each), and evaluating a logistic model on a temporally separate patient group. Such external tests are critical to gauge generalizability across different populations or imaging protocols. Cross-cohort validation using public datasets is also seen – e.g. gene signature studies often train on METABRIC and test on multiple GEO cohorts. In one case, a model was trained on one hospital's data and tested on another's data from the same period, simulating a multi-center scenario.

**Performance metrics** vary with the prediction task. For binary recurrence classification within a fixed time frame, the area under the ROC curve (AUC) is frequently reported. Many recent models achieve AUCs in the 0.75–0.85 range on held-out data, indicating good discrimination. For instance, an ensemble model reached AUC 0.817 on its test set; a blood microbiome RF model achieved AUC 0.94 in cross-validation (though based on very few events). Time-to-event predictions are often evaluated by the concordance index (C-index), which generalizes AUC to censored survival data. External C-indices around 0.70–0.80 have been reported for integrated models. The multimodal MRI deep learning model attained a C-index of 0.803 in an independent test cohort, compared to ~0.89 in internal validation. Another pathology clinical model for HER2+ cases showed C-index ~0.73 on TCGA, close to its ~0.76 internal result. These indicate moderate to strong prognostic power (for reference, traditional clinical models often have C-index ~0.6–0.7). It's notable that deep learning and ensemble models have matched or exceeded earlier models; the review found the maximum AUCs reported were 0.94 for both a deep learning and an ensemble method.

Beyond AUC/C-index, other metrics appear in some studies. Accuracy, sensitivity, and specificity are occasionally given for classification models, but they are less informative for imbalanced data. Precision, recall, F1-score, and Matthews correlation coefficient (MCC) were reported in the microbiome study, for example, to account for imbalance. Calibration of predictions has garnered attention in studies that produce absolute risk scores. Nomogram-based studies typically present calibration plots to show agreement between predicted and observed recurrence probabilities. For instance, a 2022 competing-risks nomogram was calibrated and even deployed as an online calculator for user testing. However, not all ML papers report calibration; many focus on discrimination ability alone. A few works also use decision curve analysis to evaluate clinical utility at different threshold probabilities (especially when proposing a model for decision support, to compare net benefit with standard strategies).

Comparisons against standard clinical predictors or existing tools are important for context. Some studies compare their model's AUC or hazard ratios to those from the likes of Adjuvant [31]. Online or the Nottingham Prognostic Index. Others directly evaluate against genomic assays: for example, radiomics studies used Oncotype DX recurrence score as a benchmark. One MRI radiomics analysis achieved an AUC ~0.78 in classifying patients above vs. below an Oncotype RS threshold of 25, suggesting imaging could approximate genomic risk stratification. Another deep learning model explicitly reported its performance in subgroups and noted it provided better recurrence risk stratification than the 21-gene test for certain patient subsets. Such comparisons help establish whether new ML models offer incremental value over existing prognostic indices.

In summary, rigorous validation is very important, internal cross-validation is very common, but external validation is also increasingly performed (though still not universal). Performance is typically reported in terms of AUC for classification or C-index for survival, with the best models achieving ~0.8–0.9 on validation data. Careful reporting of calibration and comparisons to current standards are emerging practices to demonstrate that high accuracy can translate into meaningful clinical predictions.

### 4.6 Transparency and Reproducibility Considerations

A major problem in the healthcare domain is to ensure that most of the ML models should be clear, easy to understand and reusable [32]. But most of the machine learning models are essentially "black boxes," which can be difficult to be trusted and used by the doctors. To counter this, recent studies have incorporated interpretability techniques. Feature importance and SHAP values (Shapley Additive Explanations) is one of the most primitive approaches for tree-based or ensemble models. For example, the 2023 ensemble model used SHAP to explain the contribution of each of the six key features to an individual patient's recurrence risk. SHAP aimed to bridge the gap between the predicted outcomes of the model and clinical judgement by integrating both global and patient-specific explanations. In that study, higher Ki-67 and positive lymph nodes had large positive SHAP values for recurrence, aligning with known risk factors, which lent credibility to the model's decisions. Some deep learning efforts use visual interpretability: CNNs on pathology images might highlight regions most indicative of recurrence (e.g. via attention maps), although such methods were not always reported in detail.

There is another way to promote transparency and that is by building simpler models from complex models. Many authors have converted their models into nomograms or online risk evaluators for the ease of use like implementing a web-based calculator from their competing-risks model for node-positive patients. Similarly, in 2021 one of the works provided a nomogram based on an ensemble of J48 decision tree and Naïve Bayes. These allow end-users to input patient characteristics and obtain recurrence risk estimates without needing to understand the underlying ML mechanics. While nomograms are inherently interpretable (linear point-based systems), they often approximate more complex models and may lose some accuracy in favor of clarity.

**Reproducibility** is aided by the use of public data and code sharing, but there are gaps. Studies leveraging TCGA, METABRIC, and GEO are inherently more reproducible because others can access these data and validate the findings. Indeed, one 2025 study explicitly cited that unlike many previous signatures that failed on independent data, their signature-maintained performance across

METABRIC, GEO, and TCGA, suggesting robust methodology and avoiding overfit to a single cohort. However, a large number of models are developed on single-institution datasets that are not publicly available (often due to patient privacy). The lack of open-source breast recurrence datasets was noted as an impediment to model development and validation. This also hinders reproducibility of results – independent researchers cannot easily verify findings or test the model on new patients. There are encouraging signs: a few authors have released their trained models or provided web interfaces, and journals/conferences are starting to expect code or model sharing as supplementary material. For instance, three ML models in the recent review were accessible online for validation by others.

**Generalizability** is another aspect of reproducibility – whether a model built in one setting works in another. As discussed, external validations across cohorts (e.g. testing a model on TCGA or a different hospital's data) are the acid test. Many ML models have shown performance drop-offs when applied to new data (due to population differences or technical batch effects). The interpretability tools can sometimes reveal why – e.g. a model might overly rely on a feature that is recorded differently elsewhere. One study found their model's performance dipped on a new cohort because one predictor (histologic grade) was pathologically assessed with different criteria; re-calibrating that feature's contribution improved transferability. This shows how important it is to clearly communicate model characteristics and coefficients so that other people can change their models as and when required.

Two important parameters of transparency are fairness and inclusiveness. The 2023 systematic review pointed out that many models overlooked certain patient groups – notably, almost all were trained on data from North American, European, or East Asian patients, with very few including African or Middle Eastern populations. Because of these limitations, the models may not work for populations that aren't well-represented. To make the model more generalizable, different datasets need to be explored and combined. Also, the methods like transfer learning or federated learning can to be used.

In conclusion, although initial efforts were concentrated on pure predictive performance, but newer research focus on the importance of interpretability (SHAP values, nomograms), open data, and validation on independent cohorts to guarantee that models are both accurate and transparent. There is increasing recognition that a slightly less complex model that clinicians trust and can reproduce a more accurate black-box model.

### 4.7 Clinical Utility and Translational Relevance

Ultimately, the value of recurrence risk models lies in their ability to impact clinical decision-making and patient outcomes. Many publications now discuss how their model could be used in practice – though true clinical translation is still limited. One major potential use is personalizing adjuvant therapy decisions. By identifying patients at high risk of recurrence, oncologists can recommend more aggressive treatments (e.g. chemotherapy, extended endocrine therapy), while low-risk patients might safely avoid overtreatment. This is the rationale behind genomic assays like Oncotype DX, and ML models aim to provide similar or improved stratification. For example, an AI model that predicts distant relapse risk in HR+ breast cancer could guide whether to add chemotherapy on top of endocrine therapy. A recent study introduced an explainable ensemble model explicitly as a clinical decision support tool (CDSS): it provided individualized risk scores along with explanations, so that high-risk patients might be considered for additional therapy and low-risk patients spared unnecessary chemo. The authors note that prospective multi-center trials are needed before deployment, but they envision integration of such a model into tumor boards and treatment planning.

Another area of utility is patient counseling and follow-up planning. Knowing a patient's recurrence risk can inform surveillance intensity (e.g. frequency of imaging follow-ups) and lifestyle or preventive interventions. Nomogram tools, for instance, can be used in clinic to discuss an individual's percentage risk of recurrence at 5 or 10 years and thereby motivate adherence to adjuvant therapy or risk-reducing measures. If a model is interpretable (showing, say, that the patient's large tumor size and lymph node involvement drive their risk), it can facilitate a conversation about why extended therapy is recommended.

Some ML models are targeting specific clinical gaps. The radiomics approaches that correlate with Oncotype DX aim to provide a cheaper, non-invasive alternative to expensive gene tests. If MRI or mammogram-based models can reliably predict which patients have a high genomic risk of recurrence, then resource-limited settings or patients who cannot afford genomic testing might still get risk stratification [33]. One multicenter radiomics signature indeed showed promise in prognosticating relapse-free survival and even predicting response to neoadjuvant chemotherapy (pCR) when combined with clinical factors. This suggests ML models might expand utility beyond recurrence risk to related outcomes like treatment response, further personalizing therapy selection.

Despite these potential benefits, gaps to clinical adoption remain. Few models have undergone prospective validation. Most are retrospective and would need testing on prospective cohorts or in clinical trials to prove that using the model improves decision-making or patient outcomes. There is also the question of integration into clinical workflow: tools must be user-friendly (hence interest in web calculators or EHR integration) and provide results in real-time. Some advanced techniques (e.g. those requiring analysis of WSIs or complex MRI radiomics) may require the use of some specialised and automated software tools that might not be available in all hospitals.

Regulatory and ethical considerations also very important in the healthcare sector. If any model directly affects the treatment decisions of the disease, it may need regulatory approval and generally requires substantial proof of safety and efficacy. So far, no ML-based recurrence predictor has reached the level of regulatory-approved tool or guideline-recommended test, unlike Oncotype or MammaPrint which went through extensive validation. There are many positive signs where several institutions are testing AI-driven risk assessments in tumor boards to determine if they agree with the conclusions or outcomes of doctors.

Decision-curve evaluations in several studies suggest that employing a machine learning model would enhance the net benefit (detecting more recurrences for a specified false-positive rate) relative to treat-all or treat-none approaches. For instance, a 2022 study showed that their model's decision curve was the deciding factor to use a high Ki-67 alone on chemotherapy. This type of study is very useful as to how it helps patients, which might be convincing to doctors as well.

To summarize, these models are very useful for translation because they provide more accurate, personalized care by grouping the patients by risk factors, regulating the strength of adjuvant medication, customizing follow-up and treatments, and replacing expensive diagnostics and clinical tests.

But to make this potential a reality, requires more validation, integration into workflows, and proof that model-guided decisions can enhance patient outcomes (for example, by avoiding therapies, clinical trial endpoints can be used like fewer recurrences or improved quality of life).

TABLE 1 shows the characteristics of the studies that were included in the systematic review, in line with the PRISMA 2020 reporting standards. It shows the different kinds of data, the predictor variables, the sources of data, the ranges of sample sizes, feature engineering and selection methods and modeling techniques used to figure out how likely breast cancer can relapse and spread. The performance measures and result outcomes were also included in the summary.

Table 1: Characteristics of Included Studies: Data Sources, Predictors, Modeling Approaches, and Outcome Definitions

| **Data Category** | **Typical Predictors / Features** | **Typical Data Sources** | **Sample Size Range** | **Feature Engineering & Selection Techniques** | **Modeling Approaches** | **Typical Performance Metrics** | **Outcome Framing** |
|---|---|---|---|---|---|---|---|
| Clinical & Registry Data | Age, tumor size, nodal status, grade, ER/PR/HER2, Ki-67, LVI, treatment details | Institutional cohorts; SEER-type registries | ~500 to >10,000 | Manual feature selection; univariate filtering; LASSO; stepwise regression; tree-based importance | Cox PH, competing-risks Cox, logistic regression; SVM, RF, XGBoost | C-index, AUC, HRs | Recurrence (binary), DFS, DMFS, metastasis |
| Molecular / Genomic Data | Gene expression (immune genes, proliferation, CAF markers), mutations | TCGA, METABRIC, GEO | ~500 to ~2,000 | Differential expression; Cox screening; GSEA; LASSO/ elastic net; ridge regression | Cox/ LASSO-Cox; ML ensembles | C-index, AUC | RFS, DMFS, relapse |
| Radiological Imaging (Radiomics) | Shape, intensity, texture, wavelet features; peritumoral features | MRI, mammography (multi-center) | ~400 to ~1,200 | ROI segmentation; radiomic extraction; LASSO feature reduction; z-score normalization | Cox + radiomics; SVM; RF; XGBoost; DL hybrids | AUC, C-index | DFS, recurrence, metastasis |
| Pathology Imaging (WSI) | Cellular morphology, tissue architecture, color/ texture patterns | Institutional pathology; TCGA slides | ~100 to >1,000 | Hand-crafted color/ texture features; MIL patch clustering; attention pooling | RF/ XGBoost (hand-crafted); CNNs (ResNet/ Xception); MIL DL | AUC | Recurrence, metastasis, genomic proxies |
| Emerging Biomarkers | ctDNA, blood microbiome taxa | Blood-based assays (ctDNA, microbiome) | <100 to ~500 | Normalization; feature filtering; SMOTE for imbalance | RF; ML classifiers | AUC | Recurrence / metastasis |
| Multimodal / Hybrid Data | Clinical + genomic + imaging features | Clinical + omics; imaging + clinical | ~500 to ~2,000 | Feature concatenation; late-fusion ensembles; stacked generalization | Ensemble ML; DL fusion; Cox hybrids | AUC, C-index | DFS, DMFS, recurrence |

### 4.8 Trends, Gaps, and Future Directions

**Trends over 2019–2024:** The literature review showed a clear trend towards multi-modal modeling – that combined the clinical, imaging, pathologic, and genomic data for a more holistic prediction of the disease. Early ML models often used only clinical features or a single genomic signature, but newer models integrated data types (for example, merging WSI features with gene expression and clinical data for a tri-omics model that markedly improved AUC). It was also observed that there was a trend focussing on subpopulations (HER2-positive, triple-negative, low-stage, etc.) to tailor models to more homogeneous groups, which can boost performance within those groups [34]. Deep learning has become very important in this area, especially with the increased availability of computing power and data – pathology image analysis and radiomics feature learning benefit from CNNs and attention mechanisms as seen in multiple 2021–2023 studies. Another trend that was seen was the support for external validation and multi-center studies, indicating maturation of the field. By 2024, multi-center radiomics and international collaborations (e.g. the AURORA project in metastasis research, etc.) were seen, whereas earlier studies were often single-institution based.

**Current gaps:** One of the most important gaps was the limited diversity of data – models were trained largely on Western or East-Asian cohorts that may not be generalized globally [35]. There was a need for including patients of African ancestry and other underrepresented groups, as identified as one of the gaps in the study. Another gap was that the most models did not incorporate time-varying factors or dynamic data; nearly all use baseline characteristics at diagnosis/surgery. But risk can evolve (e.g. new comorbidities, response to therapy). Future models might utilize longitudinal data (e.g. how quickly a tumor marker falls with treatment, or serial imaging changes) to update recurrence risk in real time. Additionally, while many studies mention outcome "metastasis," relatively few specifically tackle the question of site-specific metastasis prediction (brain vs bone etc.), which could be clinically relevant for surveillance strategies. One 2023 paper did address site-specific distant metastasis using competing risks, but this remains a niche.

**Interpretability vs. performance:** There is still a tension between using the most complex model for maximum accuracy and using simpler, more interpretable models. As a result, a gap exists in user-centered design – we need models that physicians *will actually use*. Future work may focus on hybrid models that are inherently interpretable (such as rule-based ML or generalized additive models) while still capturing nonlinear interactions. Some of the work done by the authors in artificial intelligence could be applied here, so that the doctors can predict risk factors along with explanations in plain language (e.g. "because of high Ki-67 and 4 positive nodes, patient is high risk").

**Reproducibility and standards:** The medical field would benefit from standardized benchmarks and open challenges. For example, common datasets could be established for model comparisons similar to how ImageNet drove computer vision. This is happening slowly: a Kaggle or DREAM challenge on breast cancer prediction could lead to greater transparency along with external validation by several teams. If there aren't any guidelines like these, any study may report a high AUC on its own data without a way to directly compare models. The recent systematic review attempted to benchmark models, finding that deep learning and ensembles achieved top performance but also noting many studies had methodological biases.

The summary of the literature review according to the work published by different authors has been shown in TABLE 2. Also, TABLE 2 has some entries that are study of clusters/trends that are representing parameter-level synthesis instead of showing one paper per row.

Table 2: Summary of Systematic Literature Review and Parameter-Level Synthesis

| S.No. | Study (Author, Year) | Geography / Data Source Type | Modality | Endpoint (how framed) | Modeling family | Validation | Main metrics reported | Key analytic notes (bias / limitations / strengths) |
|---|---|---|---|---|---|---|---|---|
| **1** | Nicolò et al., 2020 [36] | Multicohort clinical (secondary) | Clinical | Time-to-distant metastatic relapse | Mechanistic + ML / survival | IV + cohort tests | Time-to-event performance | Strong framing for time-to-event; mechanistic assumptions need calibration |
| **2** | Dasgupta et al., 2021 [5] | Canada/ international cohorts | QUS radiomics | Recurrence after NAT (time-to-event) | Radiomics + ML | IV (often + EV in QUS literature) | AUC/ C-index (varies) | Imaging harmonization critical; event-rate constraints common |
| **3** | Yang et al., 2021 [26] | Secondary dataset | Clinical/ tabular | Recurrence (classification) | Ensemble + cost-sensitive | IV (CV) | Accuracy/ sensitivity etc. | Illustrates imbalance handling; risk of optimistic CV if leakage |
| **4** | Liu et al., 2022 [12] | Multi-site / pathology | Histopath (H&E) | Recurrence + metastasis risk | ML on texture/ color + fusion | IV (some EV) | AUC | Feature engineering sensitive to staining/ scanner shift |
| **5** | Wang et al., 2022 [6] | Secondary MRI cohorts | MRI radiomics/ DL | Prognostic surrogate (DFS/ RFS) | Radiomics + DL + nomogram | IV | AUC/ C-index | Multimodal boosts discrimination but increases deployment burden |
| **6** | Lee et al., 2023 [37] | South Korea single-center | MRI radiomics | LRR (time-to-event / risk) | ML radiomics | IV | AUC/ C-index | Large N improves stability; needs EV across scanners |
| **7** | González-Castro et al., 2023 [3] | EHR (structured + unstructured) | Clinical + NLP | 5-year recurrence (binary) | ML (XGBoost etc.) | IV (and/ or EV depending) | AUC | Demonstrates value of unstructured EHR; label definition critical |
| **8** | Zuo et al., 2023 [38] | Secondary health data | Clinical/ tabular | Recurrence prediction | ML model comparison | IV | AUC/ Accuracy | Benchmark-style comparison; quality depends on proper tuning & splits |
| **9** | Sukhadia et al., 2023 [39] | Clinical cohorts | Clinical | Distant recurrence | ML | IV/ EV (varies) | AUC/ C-index | Highlights role of baseline survival models vs ML value-add |
| **10** | Shiner et al., 2023 [40] | Multi-institutional clinical | Clinical | Site-specific DM | ML + statistical | IV | AUC (site-specific) | Novel "where metastasis occurs" framing; class imbalance by site |

Continued on next page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **11** | Zhao et al., 2023 [41] | Registry/ SEER-type | Clinical registry | DM risk (male BC) | ML + nomogram | IV | AUC | Registry endpoints may be proxies; still useful for population risk |
| **12** | Zhong et al., 2023 [42] | Registry/ SEER-type | Clinical registry | Bone metastasis diagnosis + survival | XGBoost | IV | AUC + survival metrics | Strong scale; label/ proxy issues + missing therapy detail |
| **13** | Murata et al., 2023 [43] | Clinical recurrence cohort | Clinical | DM after isolated LRR | Statistical model | IV | AUC/ C-index | Focused use-case (post-LRR); selection bias likely |
| **14** | Su et al., 2023 (BCR-Net) [44] | International/ public cohorts | WSI (H&E) | Predict genomic RS proxy (Oncotype-like) | Deep learning | IV | AUC/ accuracy | Predicts "assay proxy" not direct recurrence; still clinically relevant surrogate |
| **15** | Howard et al., 2023 [45] | Multicohort | WSI + clinical | Prognostic/ therapy benefit signals | DL multimodal | IV/ EV (reported) | AUC/ C-index | Better confounding control when trial-like cohorts used |
| **16** | Hamedi et al., 2024 [46] | Literature synthesis | Multi | Survival/ recurrence modeling | Multiple | — | — | Confirms recurring gaps: EV scarcity, calibration underreporting |
| **17** | Liu et al., 2024 (Heliyon) [30] | Secondary clinical dataset | Clinical/ tabular | Recurrence (classification) | Explainable ensemble | IV | AUC | Good explainability; needs calibration + EV to claim clinical utility |
| **18** | Tan et al., 2024 [17] | China clinical | Clinical + ultrasound | DM prediction | ML | IV | AUC/ Accuracy | "Low-cost features" trend; watch for leakage via post-diagnosis labs |
| **19** | Ahmad S. et al., 2022 [25] | China clinical | Blood biomarkers/histology | DM prediction | AI/ ML | IV | Accuracy/ AUC | Outperforms CNN-GRU and CNN-LSTM, rich image data, strong performance |
| **20** | You et al., 2024 [47] | Multi-cohort | MRI radiomics + biology | Prognosis/ recurrence risk | Radiomics + transfer | IV + cross-cohort | AUC/ C-index | Stronger "transfer" focus; still needs prospective deployment |
| **21** | Xiong et al., 2024 [48] | China clinical | Multimodal radiomics | DFS (includes LRR/ DM/ death) | Radiomics + ML | IV | C-index/ AUC | Endpoint composite improves events but reduces interpretability of "recurrence" |
| **22** | Qi et al., 2024 (Radiomics review) [49] | Review (global) | Multi-imaging | Prognosis/ recurrence themes | Multi | — | — | Highlights: harmonization, EV, calibration, reporting standards |
| **23** | Jiang et al., 2024 (histopath DL review) [35] | Review (global) | Pathology DL | Prognosis/ metastasis themes | DL | — | — | Notes foundation models + domain shift as central issues for prognosis |

Continued on next page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **24** | Emily et al., 2024 [10] | Secondary clinical | Clinical | Survival (DFS/ RFS-type) | Cox vs survival RF | IV | C-index/ AUC | Good “ML vs classical” comparison; still depends on endpoint quality |
| **25** | Ye et al., 2020 [14] | Registry/ clinical | Clinical | Bone metastasis risk | Nomogram/ statistical | IV | AUC | Nomogram interpretable; registry confounding remains |
| **26** | (2019–2024) Imaging radiomics prognosis cluster | Global | MRI/ US/ Mammo radiomics | DFS/ recurrence/ prognosis | Radiomics + ML | Mostly IV | AUC/ C-index | Common pattern: high AUC internally; EV drop under domain shift |
| **27** | (2019–2024) EHR + NLP recurrence cluster | Global | Clinical+ text | 5y recurrence | ML | IV > EV | AUC | Stronger real-world signal; hardest part is label correctness |
| **28** | (2019–2024) WSI/ DL recurrence-surrogates cluster | Global | Pathology | RS proxy / risk stratification | DL | IV (some EV) | AUC | Surrogate endpoints useful; must avoid overclaiming “recurrence prediction” |
| **29** | (2019–2024) Multi-modal fusion trend | Global | Clinical+ imaging/ omics | DFS/ DMFS/ recurrence | Fusion DL/ ML | IV | AUC/ C-index | Best discrimination often from fusion; worst reproducibility if pipelines opaque |
| **30** | (2019–2024) Metastasis site prediction trend | Global | Clinical | Site-specific DM | ML | IV | AUC | Valuable for surveillance planning; rare-class imbalance dominates errors |
| **31** | (2019–2024) Registry-based DM prediction | Global | Registry | DM at diagnosis / DM risk | ML/ nomograms | IV | AUC | Scale advantage; therapy/ context missing ⟶ confounding risk |
| **32** | (2019–2024) Calibration underreport-ing | Global | — | — | — | — | Brier/ slope rare | Field gap: high AUC ≠ usable risk without calibration |
| **33** | (2019–2024) External validation scarcity | Global | — | — | — | EV un-common | — | Primary blocker for clinical translation in recurrence/ DM models |
| **34** | (2019–2024) Outcome definition heterogeneity | Global | — | LRR vs distant vs composite DFS | — | — | — | Prevents meta-benchmarking; must standardize recurrence taxonomy |
| **35** | (2019–2024) Class imbalance handling variability | Global | — | Recurrence rare | SMOTE/ weights | IV | Accuracy/ AUC | Resampling inflates internal metrics if leakage; must use nested CV/ strict splits |

Continued on next page

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **36** | (2019–2024) Interpretability gap in DL | Global | — | — | DL | IV | — | Explanations often technical; rarely linked to actionable clinical features |
| **37** | (2019–2024) Clinical utility evaluation rare | Global | — | — | — | — | DCA/ net benefit rare | Even strong models rarely show decision impact ⟶ “research-only” status |
| **38** | (2019–2024) Reproducibility artifacts limited | Global | — | — | — | — | — | Code/ model sharing inconsistent; hampers independent verification |
| **39** | (2019–2024) Subtype-specific modeling rising | Global | Clinical/ imaging | TNBC/ HER2/ HR+ DFS/ RFS | ML/ DL | IV | AUC/ C-index | Better within-subtype performance; poorer cross-pop transportability |
| **40** | (2019–2024) Best-practice direction | Global | — | — | — | — | — | Clear pathway: standardized endpoints + EV + calibration + TRIPOD-AI style reporting |

## 5. FUTURE SCOPE

To address the gaps as discussed, several directions stand out that have been pointed out as follows:

- **Multi-center prospective trials:** Validating AI models in a prospective setting (e.g. using the model to guide therapy in one arm vs. standard care in another) would provide highest-quality evidence. This could show if model-informed decisions truly reduce recurrence or avoid unnecessary treatment.
- **Integration of novel omics:** Beyond gene expression, other omics (proteomics, epigenetics, microbiome as early work suggests) could enhance predictions. Multi-omics could unveil new predictors like immune microenvironment signatures, circulating tumor DNA levels, etc., which might predict metastasis earlier than traditional factors [50].
- **Federated learning and privacy-preserving AI:** Given data-sharing challenges, algorithms that train on distributed data without centralizing it could allow learning from vastly larger datasets (from different hospitals/countries) while respecting privacy. This can improve model robustness and generalizability.
- **Personalized treatment effect prediction:** A complementary angle to recurrence risk is predicting which specific therapy will best reduce that risk for a given patient (treatment benefit prediction). Some recent work is moving toward combining prognostic modeling with predictive modeling of drug benefit. For instance, an AI might predict both the baseline risk and how much that risk would drop if the patient receives chemotherapy or a CDK4/6 inhibitor. This would be immensely useful for personalized therapy selection.

- **Clinical workflow integration:** Future research might also focus on the development of a software that will integrate into pathology or radiology workflows to automatically output the risk scores when a new slide or scan is read. This will require robust automation and user interface design to ensure that the output is accessible and arrives in time for treatment planning.

## 6. CONCLUSION

In conclusion, the past five years have seen significant advances in machine learning and statistical models for breast cancer recurrence and metastasis risk prediction. Models have become increasingly comprehensive (integrating multi-source data) and accurate (with AUC/C-index often >0.8), and researchers are paying more attention to validation, interpretability, and clinical utility. Still, to transition from research to routine patient care, further work is needed to fill the data gaps, ensure models are transparent and equitable, and demonstrate improved patient outcomes. Clinical and registry-based datasets are still the most common sources. The research studies have consistently found that established prognostic factors like tumor size, lymph node status, hormone receptor status, HER2 status, Ki-67, and treatment-related variables are the best predictors. Public databases like TCGA and METABRIC have made it feasible to generate biologically accurate predictive signatures. This was made possible by regularization-based feature selection, that helps keep the track of high-dimensional data. Imaging-based methods, like radiomics and digital pathology, have made modeling even more complex, especially when combined with the clinical variables. There has been a clear shift in methodology from traditional statistical survival models to machine learning, ensemble, and deep learning techniques. Multimodal fusion models often do a better job of separating different groups. But these improvements are not without their problems, such as different definitions of outcomes, limited external validation, and less clear meaning. It is important to deal with these problems in order to make predictive models more useful and easier to use in clinical settings. The trajectory is promising – these predictive tools are steadily moving toward fulfilling the goals of precision oncology by enabling clinicians to identify high-risk patients more reliably and tailor treatments to prevent recurrence, ultimately improving survival and quality of life for breast cancer survivors.

## 7. CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

## References

[1] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology. 2010;21:128-138.

[2] Sandarenu P, Millar EK, Song Y, Browne L, Beretov J, et al. Survival Prediction in Triple Negative Breast Cancer Using Multiple Instance Learning of Histopathological Images. Sci

Rep. 2022;12:14527.

[3] González-Castro L, Chávez M, Duflot P, Bleret V, Martin AG, et al. Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources From Electronic Health Records. Cancers. 2023;15:2741.

[4] Zeng Z, Yao L, Roy A, Li X, Espino S, et al. Identifying Breast Cancer Distant Recurrences From Electronic Health Records Using Machine Learning. J Healthc Inform Res. 2019;3:283-299.

[5] Dasgupta A, Bhardwaj D, DiCenzo D, Fatima K, Osapoetra LO, et al. Radiomics in Predicting Recurrence for Patients With Locally Advanced Breast Cancer Using Quantitative Ultrasound. Oncotarget. 2021;12:2437-2448.

[6] Wang Y, Li Y, Song Y, Chen C, Wang Z, et al. Comparison of Ultrasound and Mammography for Early Diagnosis of Breast Cancer Among Chinese Women With Suspected Breast Lesions: A Prospective Trial. Thorac Cancer. 2022;13:3145-3151.

[7] Pan H, Gray R, Braybrooke J, Davies C, Taylor C, et al. 20-Year Risks of Breast-Cancer Recurrence After Stopping Endocrine Therapy at 5 Years. N Engl J Med. 2017;377:1836-1846. doi: 10.1056/NEJMoa1701830, PMID 29117498.

[8] Hassett MJ, Ritzwoller DP, Taback N, Carroll N, Cronin AM, et al. Validating Billing/Encounter Codes as Indicators of Lung, Colorectal, Breast, and Prostate Cancer Recurrence Using 2 Large Contemporary Cohorts. Med Care. 2014;52:e65-73.

[9] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. J R Stat Soc B. 2005;67:301-320.

[10] Emily M, Meidioktaviana F, Nabiilah GZ, Moniaga JV. Comparative Analysis of Machine Learning and Survival Analysis for Breast Cancer Prediction. Procedia Comput Sci. 2024;245:759-767.

[11] Cirkovic BR, Cvetkovic AM, Ninkovic SM, Filipovic ND. Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients. 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE.2015:1-6.

[12] Liu X, Yuan P, Li R, Zhang D, An J, et al. Predicting Breast Cancer Recurrence and Metastasis Risk by Integrating Color and Texture Features of Histopathological Images and Machine Learning Technologies. Comput Biol Med. 2022;146:105569.

[13] Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, et al. Prediction of Survival and Metastasis in Breast Cancer Patients Using Machine Learning Classifiers. Clin Epidemiol Glob Health. 2019;7:293-299.

[14] Ye LJ, Suo HD, Liang CY, Zhang L, Jin ZN, et al. Nomogram for Predicting the Risk of Bone Metastasis in Breast Cancer: A Seer Population-Based Study. Transl Cancer Res. 2020;9:6710-6719.

[15] Lashari SA, Ibrahim R, Senan N. De-Noising Analysis of Mammogram Images in the Wavelet Domain Using Hard and Soft Thresholding. In: 4th World Congress on Information and Communication Technologies (WICT 2014). IEEE. 2014:353-357.

[16] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. Nucleic Acids Res. 2002;30:207-210.

[17] Tan Y, Zhang WH, Huang Z, Tan QX, Zhang YM, et al. AI Models Predicting Breast Cancer Distant Metastasis Using Lightgbm With Clinical Blood Markers and Ultrasound Maximum Diameter. Sci Rep. 2024;14:15561.

[18] Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. BMJ. 2024;385:e078378.

[19] Abreu PH, Santos MS, Abreu MH, Andrade B, Silva DC. Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Comput Surv. 2016;49:1-40.

[20] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, et al. A Guide to Deep Learning in Healthcare. Nat Med. 2019;25:24-29.

[21] Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, et al. Predicting Cancer Outcomes From Histology and Genomics Using Convolutional Networks. Proc Natl Acad Sci U.S.A. 2018;115:E2970-E2979.

[22] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine Learning Applications in Cancer Prognosis and Prediction. Comput Struct Biotechnol J. 2015;13:8-17.

[23] Cox DR. Regression Models and Life-Tables. J R Stat Soc B (Methodol). 1972;34:187-202.

[24] Magboo VP, Magboo MS. Machine Learning Classifiers on Breast Cancer Recurrences. Procedia Comput Sci. 2021;192:2742-52.

[25] Ahmad S, Ullah T, Ahmad I, Al-Sharabi A, Ullah K, et al. A Novel Hybrid Deep Learning Model for Metastatic Cancer Detection. Comput Intell Neurosci. 2022;2022:8141530.

[26] Yang PT, Wu WS, Wu CC, Shih YN, Hsieh CH. Breast Cancer Recurrence Prediction With Ensemble Methods and Cost-Sensitive Learning. Open Med (Wars). 2021;16:754-768.

[27] Tseng YJ, Huang CE, Wen CN, Lai PY, Wu MH, et al. Predicting Breast Cancer Metastasis by Using Serum Biomarkers and Clinicopathological Data With Machine Learning Technologies. Int J Med Inform. 2019;128:79-86.

[28] Roberto Cesar MO, German LB, Paola Patricia AC, Eugenia AR, Elisa Clementina OM, et al. Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis. In: Tan Y, Shi Y, Tuba M, editors. Advances in swarm intelligence ICSI. Cham: Springer International Publishing. 2020:584-596.

[29] Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, et al. Personalizing the Treatment of Women With Early Breast Cancer: Highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24:2206-2223.

[30] Liu Y, Fu Y, Peng Y, Ming J. Clinical Decision Support Tool for Breast Cancer Recurrence Prediction Using Shap Value in Cooperative Game Theory. Heliyon. 2024;10:e24876.

[31] Vickers AJ, Elkin EB, Peele PB, Dickler M, Siminoff LA. Long-Term Health Outcomes of a Decision Aid: Data From a Randomized Trial of Adjuvant! In Women With Localized Breast Cancer. Med Decis Making. 2009;29:461-467.

[32] Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319:1317-1318. doi: 10.1001/jama.2big data017.18391. PMID 29532063.

[33] Baines CJ, Vidmar M, McKeown-Eyssen G, Tibshirani R. Impact of Menstrual Phase on False-Negative Mammograms in the Canadian National Breast Screening Study. Cancer. 1997;80:720-724.

[34] Wu T, Sultan LR, Tian J, Cary TW, Sehgal CM. Machine Learning for Diagnostic Ultrasound of Triple-Negative Breast Cancer. Breast Cancer Res Treat. 2019;173:365-373.

[35] Jiang YZ, Ma D, Jin X, Xiao Y, Yu Y, et al. Integrated Multiomic Profiling of Breast Cancer in the Chinese Population Reveals Patient Stratification and Therapeutic Vulnerabilities. Nat Cancer. 2024;5:673-690.

[36] Nicolò C, Périer C, Prague M, Bellera C, MacGrogan G, et al. Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. JCO Clin Cancer Inform. 2020;4:259-274.

[37] Lee S, Lee CY, Kim NY, Suh YJ, Lee HJ, et al. Feasibility of UTE-MRI-Based Radiomics Model for Prediction of Histopathologic Subtype of Lung Adenocarcinoma: In Comparison With CT-Based Radiomics Model. Eur Radiol. 2024;34:3422-3430.

[38] Zuo D, Yang L, Jin Y, Qi H, Liu Y, et al. Machine Learning-Based Models for the Prediction of Breast Cancer Recurrence Risk. BMC Med Inform Decis Mak. 2023;23:276.

[39] Sukhadia SS, Muller KE, Workman AA, Nagaraj SH. Machine Learning-Based Prediction of Distant Recurrence in Invasive Breast Carcinoma Using Clinicopathological Data: A Cross-Institutional Study. Cancers. 2023;15:3960.

[40] Shiner A, Kiss A, Saednia K, Jerzak KJ, Gandhi S, et al. Predicting Patterns of Distant Metastasis in Breast Cancer Patients Following Local Regional Therapy Using Machine Learning. Genes. 2023;14:1768.

[41] Zhao F, Miyashita M, Hattori M, Yoshimatsu T, Howard F, et al. Racial Disparities in Pathological Complete Response Among Patients Receiving Neoadjuvant Chemotherapy for Early-Stage Breast Cancer. JAMA Netw Open. 2023;6:e233329.

[42] Zhong X, Lin Y, Zhang W, Bi Q. Predicting Diagnosis and Survival of Bone Metastasis in Breast Cancer Using Machine Learning. Sci Rep. 2023;13:18301.

[43] Murata T, Yoshida M, Shiino S, Ogawa A, Watase C, et al. A Prediction Model for Distant Metastasis After Isolated Locoregional Recurrence of Breast Cancer. Breast Cancer Res Treat. 2023;199:57-66.

[44] Su Z, Niazi MK, Tavolara TE, Niu S, Tozbikian GH, et al. BCR-Net: A Deep Learning Framework to Predict Breast Cancer Recurrence From Histopathology Images. PLOS One. 2023;18:e0283562.

[45] Howard FM, Kather JN, Pearson AT. Multimodal Deep Learning: An Improvement in Prognostication or a Reflection of Batch Effect? Cancer Cell. 2023;41:5-6.

[46] Hamedi SZ, Emami H, Khayamzadeh M, Rabiei R, Aria M, et al. Application of Machine Learning in Breast Cancer Survival Prediction Using a Multimethod Approach. Sci Rep. 2024;14:30147.

[47] You C, Su GH, Zhang X, Xiao Y, Zheng RC, et al. Multicenter Radio-Multiomic Analysis for Predicting Breast Cancer Outcome and Unravelling Imaging-Biological Connection. npj Precis Oncol. 2024;8:193.

[48] Xiong L, Tang X, Jiang X, Chen H, Qian B, et al. Automatic Segmentation-Based Multi-Modal Radiomics Analysis of US and MRI for Predicting Disease-Free Survival of Breast Cancer: A Multicenter Study. Breast Cancer Res. 2024;26:157.

[49] Qi YJ, Su GH, You C, Zhang X, Xiao Y, Jiang YZ et al. Radiomics in Breast Cancer: Current Advances and Future Directions. Cell Rep Med. 2024;5:101719.

[50] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ et al. The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. Nature. 2012;486:346-352.