

Machine learning for predicting breast cancer risk and outcomes: a scoping review

Dini Alyani, Soehartati Gondhowiardjo



pISSN: 0853-1773 • eISSN: 2252-8083
<https://doi.org/10.13181/mji.rev.268261>
Med J Indones. 2026;35:151–62

Received: June 13, 2025

Accepted: March 04, 2026

Published online: May 19, 2026

Authors' affiliations:

*Department of Radiation Oncology,
 Faculty of Medicine, Universitas
 Indonesia, Cipto Mangunkusuma
 Hospital, Jakarta, Indonesia*

Corresponding author:

Dini Alyani
 Department of Radiation Oncology,
 Faculty of Medicine, Universitas
 Indonesia, Cipto Mangunkusuma
 Hospital, Jalan Diponegoro No. 71,
 Central Jakarta 10430, DK Jakarta,
 Indonesia
 Tel/Fax: +62-21-8148393
 E-mail: alyanidini19@gmail.com

ABSTRACT

Breast cancer (BC) is the most frequently diagnosed malignancy worldwide and a leading cause of death in women. Machine learning (ML) enables data-driven risk stratification and outcome prediction, potentially improving screening, treatment choices, and overall care. This scoping review cataloged ML models used to predict BC risk and post-treatment outcomes. A PubMed search using MeSH terms (“breast neoplasms,” “breast cancer,” “machine learning,” and “predictive value of tests”) retrieved 44 records, with 9 added by manual searching; 24 studies were included. We excluded 29 reports limited to imaging/radiomics or biomolecular/omics, plus other ineligible items. Model usage among included studies: 13 random forest (RF); 13 support vector machine; 7 decision tree; 5 k-nearest neighbors; 4 deep learning; 3 naïve Bayes; 2 artificial neural networks. Average RF accuracy surpassed 90% in most reports, indicating RF is both common and consistently strong for predictive modeling, supporting targeted clinical translation.

KEYWORDS breast neoplasm, machine learning, predictive value of tests

Artificial intelligence (AI) comprises computational methods designed to emulate human judgment and problem solving. Within AI, machine learning (ML), including deep learning (DL), learns patterns from data to generate predictions. In healthcare, its principal applications are oncology prognostication, encompassing overall survival, recurrence, treatment-related toxicities, biomarker discovery, and tumor detection/classification.¹

Breast cancer (BC) is the most prevalent malignancy among women and is a leading cause of cancer-related mortality owing to its high incidence. Marked inter-patient heterogeneity yields a wide

variation in 5-year survival, and a large number of cancer deaths occur in low- and middle-income countries,² where constrained resources limit access to optimal care. Accurate risk and survival predictions can inform therapeutic selection, prevent overtreatment, support clinical trial eligibility assessments, and improve cost efficiency. Numerous ML approaches have been evaluated for the diagnosis, risk stratification, recurrence, and symptom prediction. Although some investigations identify support vector machines (SVMs) as top performers on benchmark datasets, others report superior results for artificial neural networks (ANNs)

or random forests (RFs) models, and rigorous external validation and calibration remain limited.²⁻⁴

ML algorithms are generally categorized into three main types: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. This classification is primarily based on the nature of the desired output. In supervised learning, the algorithm is guided by a ‘teacher’ using labeled training data. Each input is paired with a corresponding output label, enabling the model to learn the underlying relationships and apply these learned patterns to classify or predict outcomes for new unseen data. In unsupervised learning, the examples provided are unlabeled, allowing the ML model to identify patterns within the data independently. Examples of unsupervised ML include self-organizing feature maps, hierarchical clustering, and k-means clustering algorithms. These methods create clusters from raw, unlabeled, or unclassified data. These clusters can later be used to develop classification schemes. Reinforcement learning is an ML technique in which an agent learns to make decisions by interacting with the environment, taking action, and receiving feedback in the form of rewards or penalties. The goal of the

agent is to learn a policy that maximizes cumulative reward over time by exploring and exploiting its environment.

Considering the increasing significance of identifying prognostic factors and the expanding application of ML in predictive modeling, it is pertinent to review the existing studies focusing on ML approaches for predicting cancer incidence and prognosis. This study aimed to examine the prevailing trends in the types of ML algorithms employed, nature of integrated data sources, specific outcomes targeted for prediction, and overall predictive performance of ML models. Such an evaluation is necessary because consolidated evidence comparing the performances of different ML models in predicting BC risk and outcomes remains limited.

METHODS

We searched PubMed (Medical Literature Analysis and Retrieval System Online [MEDLINE]) from database inception to January 24, 2023, using a combination of Medical Subject Headings (MeSH) and free-text terms for BC, ML/AI, and prediction/prognosis. Core

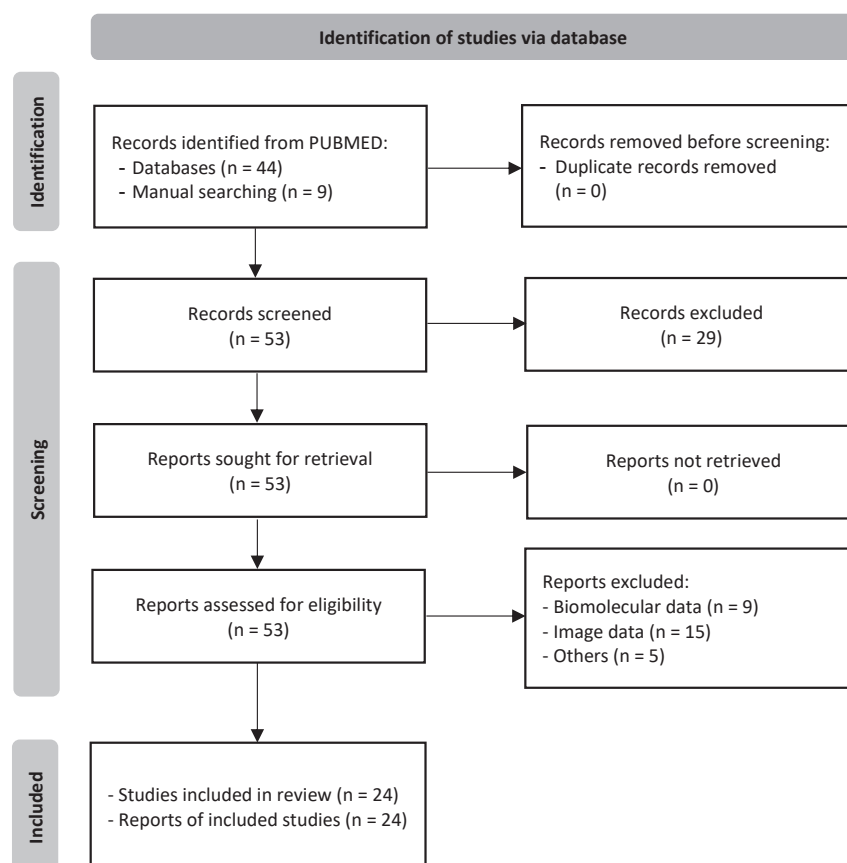


Figure 1. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 flow diagram

MeSH terms included “Breast Neoplasms,” “Machine Learning,” “Artificial Intelligence,” “Neural Networks (Computer),” “Deep Learning,” “Prognosis,” and “Treatment Outcome,” supplemented with title/abstract synonyms (e.g., “random forest,” “support vector machine,” “XGBoost,” “risk stratification,” “recurrence,” “survival,” “toxicity,” “prediction model”). We also included the MeSH term “Predictive Value of Tests.” No language or date limitations were imposed during the search. Publications were selected for review if they used one or more ML methods, at least two risk factors, or an experimental or observational study design. Imaging or molecular-only studies were excluded from the analysis.

This scoping review adhered to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 reporting guidelines (with PRISMA-extension for scoping reviews elements where applicable). Figure 1 summarizes the study selection process, showing 15 records identified through database searching and 9 through manual searching, for a total of 24 studies included. We pre-specified the review and eligibility criteria (Population: BC; Index/Intervention: ML models for risk and post-treatment outcomes; Outcomes: risk, survival, recurrence, toxicity, and others; Study design: original research) for this study. The largest database was searched (PubMed/MEDLINE and manual/reference checking). Two-stage screening (title/abstract, then full text) followed predefined criteria. We chart the algorithmic families (e.g., RF, SVM, ANN/DL), data sources, endpoints, performance metrics (area under the receiver operating characteristic [AUROC], accuracy, sensitivity, and others), and validation approaches (internal/external). Data items were screened and extracted, and both applicability and risk of bias (RoB) were evaluated using the checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS), with all processes conducted independently by the two reviewers (DA and SG).⁵ The reporting of this review was guided by the majority of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist items.⁶

RESULTS

Table 1 summarizes the study characteristics and Table 2 presents the methodological assessment

and model performance metrics. The assessment of model performance utilized a range of measures and was not consistently reported across the included studies.^{3,4,7–28} The results are summarized in Table 2. The most frequently reported measure was the AUROC, with accuracy reported in 19 of the 24 studies (79%). When available, the supplementary materials provided by the corresponding authors were used for data extraction. If such materials were not provided, data were extracted solely from the text of the published article. Confidence intervals for the model performance measures have not always been reported.

We identified 15 studies through database searching and nine additional studies through manual searching, resulting in a total of 24 relevant studies. RF was used in 13 studies. SVM was applied in 13 studies, followed by decision tree (DT) in seven studies, the k-nearest neighbors (kNNs) algorithm in five, DL in four, naïve Bayes in three, and ANN in two. The average accuracy of RF reported in most studies exceeded 90% (Figure 2).

DISCUSSION

ML, a subfield of AI, leverages a range of statistical, probabilistic, and mathematical optimization techniques to “learn” from historical data. These methods enable systems to classify new inputs, recognize previously unseen patterns, and forecast emerging trends. Although ML shares certain similarities with traditional statistical approaches in terms of data analysis AND, OR, NOT interpretation, it extends beyond these by incorporating Boolean logic (e.g., AND, OR, NOT), causal reasoning (e.g., IF, THEN, ELSE), conditional probabilities (e.g., $P[X|Y]$), and nontraditional optimization strategies to model complex data relationships and perform pattern classification. While classical approaches, such as multivariable regression or correlation, often assume linearity and predictor independence, and thus may miss complex structures, ML models can flexibly accommodate nonlinear effects and interactions. When appropriately specified, classical models can address some of these features but typically require manual transformations and interaction terms.

However, the accuracy of ML models is not universally guaranteed. Similar to conventional analytical methods, a comprehensive understanding of

Table 1. Studies with ML model for BC outcome prediction

First author, year	ML technique	Variable assessed	Development (with internal validation)/development with external validation/validation only	Outcome measured
Ding, ⁷ 2013	SVM	12 clinicopathological features (patient age, tumor size, histological type and grade, vascular invasion, hormonal receptor status, etc.)	Training and testing set	Predicting non-SLN status in patients with SLN-positive BC.
Montazeri, ⁴ 2016	Naïve Bayes, trees RF, 1NN, AdaBoost, SVM, radial basis function network, and MLP	Sex, age, morphology, grade, surgery, radiotherapy, chemotherapy, and status	10-fold cross technique	Alive or dead
Ganggayah, ⁸ 2019	DT, RF, neural network, extreme boost, LR, and SVM	23 independent/predictor variables and 1 dependent/categorical variable	Training and testing set	Mortality
Xu, ⁹ 2019	Forest, DT, logistic, GradientBoost, and LightGBM	Age, laboratory, intraoperative nonsteroidal anti-inflammatory drug, chemotherapy, tumor size, axillary surgery, breast surgery, HER2, and molecular subtype	Training set and verification set	Mortality
Whitney, ¹⁰ 2018	RF, SVM, linear discriminant analysis, and neural network	216 features corresponding to the nuclear shape and architecture categories from each pathologic image.	3-fold cross validation	Digitized H&E pathology images of early-stage ER+ BC might be able predict the corresponding oncotype DX risk categories
Kim, ¹¹ 2019	Azure ML (2 class decision forest, 2 class decision jungle, 2 class Bayes point machine, 2 class SVM, and 2 class neural network)	Age, histopathology, and molecular status	Internal and external validation	High and low risk oncotype DX criteria RS
Zhu, ¹² 2021	DL model	Age, BMI, dense breast volume, family history of BC, history of breast biopsy, race, clinical BI-RADS breast density, and automated BI-RADS density category	Training, validation, test set, and final independent test	DL models to estimate the risk of interval and screening-detected BCs with and without clinical risk factors
Pawloski, ¹³ 2022	RF	Clinicopathologic data	Training and validation	Supervised ML model to predict oncotype DX risk category
Lindsay, ¹⁴ 2019	RF, LR, and DT	Patient demographic, medical/surgical history, tumor characteristic, radiotherapy treatment history, and adverse effect	10-fold cross validation	Post radiotherapy treatment failure and adverse event
Wu, ¹⁵ 2019	DT with the CHAID	Classification of molecular subtype, imaging characteristics, and disease history	10-fold cross validation	Molecular subtype (luminal A, luminal B HER2, and basal-like BC)
Akselrod-Ballin, ¹⁶ 2019	Combination of ML and DL	Age, BMI, age first menstruation, post menopause, 1 year outcome, biopsy positive for cancer, biopsy negative for cancer, and BI-RADS category 2	Training, validation, and testing	Detecting early BC
Kim, ¹⁷ 2020	SVM and RF	Expression profile of 1,227 genes from TNBC samples, classification of TNBC to 4 subtypes, and primary metastasis	Training and validation	Response to cisplatin

Table continued on next page

Table 1. (Continued)

First author, year	ML technique	Variable assessed	Development (with internal validation)/development with external validation/validation only	Outcome measured
Nicolò, ¹⁸ 2020	Scikit-learn python package: LR, SVM, RF, KNN, and gradient boosting	21 clinicopathological variables	10-fold cross validation (internal validation)	Time to metastatic relapse
Boeri, ³ 2020	ANN and SVM	More than 30 factors contribute to recurrence risk	Training and testing set	Cancer recurrence (locoregional), cancer recurrence (systemic), and death from disease
Mosayebi, ¹⁹ 2020	RF, LVQ, Bayesian, C5.0 (DT), MLP, KPCA-SVM, and SVM	23 variables: clinical and pathological factors	Nested 5-fold cross validation	Recurrence of BC
Lou, ²⁰ 2020	ANN, KNN, SVM, NBC, and COX proportional-hazards model	Demographic, clinical characteristic, quality of care within 10 years, readmission in 30 days, and preoperative QoL	Internal and external validation	BC recurrence
Liu, ²¹ 2020	LASSO and SVM-RFE	7-lncRNA prognostic signature was constituted and validated, and stratified analysis was used to test the universal adaption of the signature in multiple BC groups.	Internal and external validation	OS
Chen, ²² 2020	Autoencoder unsupervised learning technique	Screen optimal genes for constructing a radiotherapy sensitivity estimation	Training and test	1-, 3-, and 5-year OS of radiation treated BC patient
Pellegrini, ²³ 2021	CVN	Clinical, histological, and molecular factors	Training, test, and validation	Predictions of survival at 5 years for BC patients using gene expression level
Wu, ²⁴ 2023	10 common ML algorithms	Age at diagnosis, tumor size, M status, regional lymph node dissection, N status, breast subtype, radiotherapy, grade, marital status at diagnosis, race, chemotherapy, surgery, gender, tumor site, and origin of primary	9-fold cross validation method	Estimation of survival of BC patients
Hou, ²⁵ 2020	XGBoost, RF, DNN, and LR	Main residence, number of live births, menopause status, age, age at first birth, BMI, duration of reproductive span, pregnancy history, and family history of BC	5-fold cross validation	BC occurrence
Najji, ²⁶ 2021	SVM, RF, LR, C4.5 (DT), and KNN	The features of data set are computed from a digitized image of a BC sample obtained from fine needle aspirate	Training and testing	BC occurrence
Rabiei, ²⁷ 2022	RF, gradient boosting tree, and MLP	24 demographic, laboratory, and mammographic features	K-fold validation and test and train data	BC occurrence
Saleh, ²⁸ 2022	DT, KNN, naive Bayes, RF, and SVM	30 features and 1 class label	Training and testing set	ML vs. DL which in 1 is better in prognostic

1NN=1-nearest neighbor; ANN=artificial neural network; BC=breast cancer; BI-RADS=breast imaging-reporting and data system; BMI=body mass index; CHAID=chi-square automatic interaction detection; CVN=coherent voting network; DL=deep learning; DNN=deep neural network; DT=decision tree; ER=estrogen receptor; GBM=gradient boosting machine; H&E=hematoxylin and eosin; HER2=human epidermal growth factor receptor 2; KNN=k-nearest neighbor; KPCA=kernal principal component analysis; LASSO=least absolute shrinkage and selector operation; lncRNA=long non-coding RNA; LR=logistic regression; LVQ=learning vector quantization; ML=machine learning; MLP=multilayer perceptron; NBC=naive Bayes classifier; OS=overall survival; QoL=quality of life; RF=random forest; RFE=recursive feature elimination; RS=recurrence score; SLN=sentinel lymph node; SVM=support vector machine; TNBC=triple-negative breast cancer; XGBoost=extreme gradient boosting

Table 2. Model performance indicators

First author, year	ML model	AUROC (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 score (%)
Ding, ⁷ 2013	SVM	NA	97.7	85.7	91.8	NA	NA
Montazeri, ⁴ 2016	- Naïve Bayes	95	64	98	95	94	NA
	- Trees RF	98	80	98	97	96	
	- 1-nearest neighbor	89	100	78	81	91	
	- AdaBoost	95	68	96	94	94	
	- SVM	82	65	98	95	94	
	- Radial basis function network	95	67	98	95	95	
	- MLP	96	79	99	97	95	
Whitney, ¹⁰ 2018	- RF	78	NA	NA	76 (SVM)	NA	NA
	- Neural network	73					
	- SVM	83					
	- Linear discriminant analysis	81					
Kim, ¹¹ 2019	Azure ML (2 class decision forest, 2 class decision jungle, 2 class bayes point machine, 2 class SVM, and 2 class neural network)	91.7	17	96	88	25	NA
Lindsay, ¹⁴ 2019	- RF	78.1	NA	NA	NA	NA	NA
	- LR	71.0					
	- DT	73					
Ganggayah, ⁸ 2019	- DT	72	82	75	79	91	NA
	- RF	86	83	81	82	93	
	- Neural network	84	83	79	82	93	
	- Extreme boost	87	84	75	81	89	
	- LR	85	82	78	81	92	
	- SVM	85	81	84	81	95	
Wu, ²⁴ 2019	DT	NA	79.4	NA	74.1	NA	77.4
Akselrod-Ballin, ¹⁶ 2019	Combination of ML and DL	91	87	77.3	NA	NA	NA
Xu, ⁹ 2019	- Logistic	71	NA	NA	73	57	21
	- DT	9			74	56	38
	- Forest	71			73	100	6
	- Gradient boosting	73			73	55	10
	- GBM	70			72	50	4
Kim, ¹⁷ 2020	SVM	NA	92	97	96	NA	NA
Nicolò, ¹⁸ 2020	- Mechanistic model	73	75	67	68	NA	30
	- Random survival forest	73	64	70	69		28
	- RF	75	71	66	66		28
	- LR	75	42	87	83		31
	- kNN	62	2	100	91		5
	- Gradient boosting	71	11	98	90		17
	- SVM	64	9	95	87		11
	- Cox regression	71	66	73	72		31
Boeri, ³ 2020	- ANN	91	35	98	96	NA	NA
	- SVM	89	41	99	96		
Hou, ²⁵ 2020	- XGBoost	74	65	68	67	NA	NA
	- RF	72	65	67	66		
	- Deep neural network	72	64	67	66		
	- LR	63	49	66	57		

Table continued on next page

Table 2. (Continued)

First author, year	ML model	AUROC (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 score (%)
Mosayebi, ¹⁹ 2020	- RF	72	76	68	71	NA	72
	- LVQ	63	65	68	66		63
	- Bayesian	69	57	72	65		69
	- C5.0 (DT)	76	86	77	81		76
	- MLP	62	50	74	61		62
	- KPCA-SVM	77	83	74	78		77
	- SVM	74	78	69	72		74
Liu, ²¹ 2020	LASSO and SVM-RFE	74.8 (3-year OS), 75.2 (5-year OS), and 77.1 (10-year OS)		NA	NA	NA	NA
Lou, ²⁰ 2020	- ANN	97	88	95	94	NA	NA
	- kNN	61	46	76	67		
	- SVM	82	70	94	89		
	- NBC	50	100	0	19		
	- COX proportional-hazards model	10	20.9	6	5		
Chen, ²² 2020	Autoencoder unsupervised learning technique	69 (1-year OS), 70 (3-year OS), and 73 (5-year OS)		NA	NA	NA	NA
Pellegrini, ²³ 2021	CVN	63	80	84	NA	NA	NA
Naji, ²⁶ 2021	- SVM	96	99	NA	98	97	98
	- RF	96	98		99	97	97
	- LR	94	99		95	95	97
	- C4.5 (DT)	94	97		98	96	96
	- kNN	95	96		94	95	95
Zhu, ¹² 2021	DL	66	NA	NA	NA	NA	NA
Pawloski, ¹³ 2022	RF	NA	48.3	96.3	NA	NA	NA
Saleh, ²⁸ 2022	- DT	NA	NA	NA	94.24	94.48	94.4
	- kNN				89.85	90.63	89.51
	- Naïve Bayes				80.74	81.22	79.85
	- RF				96.57	96.72	96.45
	- SVM				93.74	93.98	93.68
Rabiej, ²⁷ 2022	- RF	53	93	83	79	NA	NA
	- Gradient boosting tree	59	63	87	62		
	- MLP	56	78	85	71		
Wu, ²⁴ 2023	- Naïve Bayes	80.5	71.6	74.9	74.1	NA	57.9
	- Linear discriminant analysis	81.3	66.5	80.6	77.1		59.1
	- Quadratic discriminant analysis	77.8	62.5	79.1	75.0		55.5
	- kNN	76.2	59.6	80.7	75.4		54.7
	- SVM	77.9	69.2	79.0	76.6		59.6
	- Classification and regression trees	79.3	70.2	74.8	73.6		57.0
	- RF	81.0	76.3	70.1	71.7		57.3
	- Multivariate adaptive regression splines	83.1	73.7	77.2	76.4		60.8
	- LR	81.5	73.7	74.0	73.9		58.5
	- XGBoost	75.0	67.7	72.4	71.3		53.9

ANN=artificial neural network; AUROC=area under the receiver operating characteristic; CVN=coherent voting network; DL=deep learning; DT=decision tree; GBM=gradient boosting machine; kNN=k-nearest neighbor; KPCA=kernel principal component analysis; LASSO=least absolute shrinkage and selector operation; LR=logistic regression; LVQ=learning vector quantization; ML=machine learning; MLP=multilayer perceptron; NA=not available; NBC=naïve Bayes classifier; OS=overall survival; RF=random forest; RFE=recursive feature elimination; SVM=support vector machine; XGBoost=extreme gradient boosting

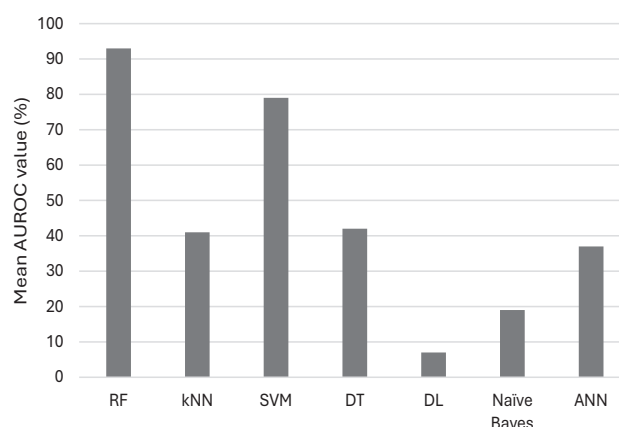


Figure 2. The mean AUROC value from 24 studies relevant to this review. AUROC=area under the receiver operating characteristic; ANN=artificial neural network; DL=deep learning; DT=decision tree; kNN=k-nearest neighbor; RF=random forest; SVM=support vector machine

the fundamental problems and constraints inherent in data is critical for achieving reliable and meaningful ML outcomes. Equally important is a clear understanding of the assumptions and limitations of the selected algorithm. Therefore, careful design, accurate implementation, and rigorous validation are likely to produce reliable and accurate results. Naturally, the quality of the input data directly influences the reliability of the output, underscoring the principle of “garbage in, garbage out.” Additionally, when the number of predictor variables exceeds the number of observations, there is an increased risk of overfitting, whereby the model captures noise rather than the underlying patterns in the data. This type of error is not exclusive to ML, but can occur in conventional statistical models. A common solution to address overfitting is to either reduce the number of variables (features) or increase the size of the training dataset. The widely accepted rule of thumb suggests that the sample-to-variables ratio should exceed 5:1. In addition to the sample size, the diversity and representativeness of the training data were equally critical. Excessive training on a small homogeneous dataset may result in overfitting, wherein the model performs well on training data but poorly on unseen data, analogous to a person trained narrowly on a single task who struggles when confronted with novel situations. In some cases, conventional statistical methods may outperform ML models. This may arise from inaccurate initial assumptions such as the presumption of variable interdependence or linearity, which may not hold

true for a given dataset. Importantly, this should not be viewed as a shortcoming of ML itself but rather as a reflection of data suitability, emphasizing the importance of selecting an appropriate analytical method for a specific problem.

Similarly, not all ML algorithms exhibit equal performance across different problem domains. Certain algorithms are better suited to specific task types, whereas others may be more effective in alternative contexts. For example, some ML methods demonstrate robust performance in biological applications, whereas others are less appropriate for such domains. Consequently, selecting the most suitable algorithm for a given research question is not always straightforward. To account for this variability, it is advisable to apply multiple ML algorithms during the analysis of training datasets to ensure a more comprehensive and reliable evaluation.

The popularity of RF across the included studies likely reflects the practical advantages of tabular clinical data: strong out-of-the-box discrimination with minimal preprocessing, the ability to learn nonlinear effects and high-order interactions, built-in internal error estimation (out-of-bag), and accessible variable importance summaries that facilitate clinical interpretation and feature selection. Nevertheless, the prevalence of use should not be conflated with superiority; model selection should prioritize external validation, calibration, and clinical utility rather than the popularity of the algorithm. RFs combine many weakly correlated DTs via bagging and random feature subsampling, reducing variance, and often yielding strong performance with minimal tuning across heterogeneous tabular datasets. In this review, RF was found to be among the most frequently applied and best-performing algorithms for BC prediction and prognosis. This observation is consistent with previous studies reporting that ensemble tree-based methods demonstrate strong performance in heterogeneous clinical datasets.^{4,7,14,26,28} However, no algorithm is universally optimal across problem spaces; neither RF nor SVM can be assumed to be superior *a priori*. Performance depends on the data-generating process, feature representation, and objective and should be verified (ideally with external validation) in the target cohort.

High area under the curve (AUC) does not guarantee usefulness at the bedside. There are still multiple stages ahead, and decisions should be

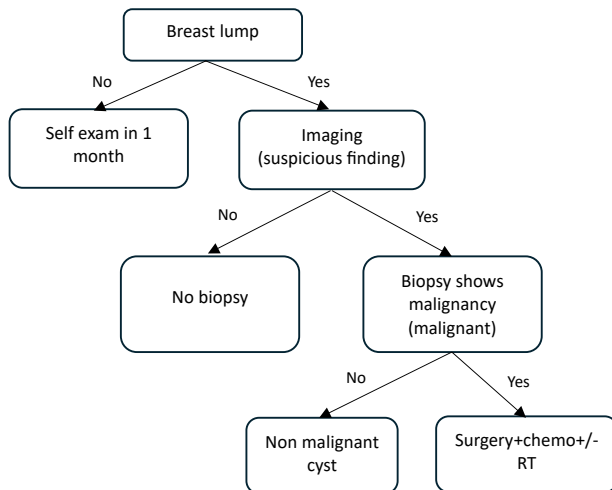


Figure 3. An example of a simple DT used in the diagnosis and treatment of BC. The DT is formulated based on expert assessment. A similar tree structure can be generated by ML. BC=breast cancer; DT=decision tree; ML=machine learning; RT=radiotherapy

guided by calibrated risk estimates and net benefits at clinically relevant thresholds. Decision curve analysis (DCA) quantifies benefit-harm trade-offs and should accompany discrimination and calibration reporting to support triage and treatment decisions. Poor calibration can negate clinical value, even when discrimination is good. Thus, AUC quantifies discrimination, whereas DCA quantifies the net clinical benefit at actionable thresholds; reporting both (with calibration) provides a complete assessment of predictive performance and decision impact.

A DT is a structured model, represented as a flowchart, composed of decision nodes and outcome branches that systematically guide the decision-making process to achieve a specific objective. It organizes complex decisions into a sequence of logical steps, thereby enabling clear and interpretable predictions. Although widely adopted in contemporary ML, DTs have longstanding historical roots and remain integral to various diagnostic and decision support protocols (Figure 3).

DTs are employed for data classification, wherein the terminal nodes (leaves) represent class labels, and the internal branches correspond to combinations of features that lead to these classifications. The construction of a DT involves recursively partitioning labeled training data into subsets based on numerical or logical criteria and continuing this process until further splitting is not feasible or a definitive

classification is reached. DTs offer several advantages: they are interpretable, easy to visualize, require minimal preprocessing, can handle diverse data types (numerical, nominal, and categorical), generate stable classifications, and can be validated using statistical methods. However, compared to ANNs, DTs generally demonstrate reduced effectiveness in classifying highly complex data patterns. In the reviewed literature, Saleh et al¹⁴ applied five ML models to identify key variables associated with BC diagnosis using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. According to their performance evaluation, the DT algorithm ranked third in terms of accuracy, followed by SVM and RF.

RF is an ensemble-based ML algorithm that builds a collection of DTs during the training phase and aggregates their outputs to produce a final prediction using majority voting for classification tasks or averaging for regression tasks. The algorithm introduces stochasticity through two mechanisms: (1) bootstrap sampling of the training data (i.e., drawing subsets with replacement), and (2) random selection of feature subsets at each node split within individual trees. This ensemble approach leverages the strengths of multiple uncorrelated DTs, thereby reducing overfitting, increasing generalizability, and improving overall predictive performance and robustness. The RF can handle high-dimensional data, accommodate both numerical and categorical variables, and provide estimates of feature importance, making it widely applicable across various domains in classification and regression problems. Montazeri et al⁴ demonstrated that the tree RF model, which is a rule-based classification algorithm, achieved the highest level of accuracy among the models tested. Consequently, this model is recommended as a valuable tool for predicting BC survival and for supporting medical decision-making.

Naïve Bayes is a probabilistic ML algorithm that utilizes Bayes' theorem to perform classification tasks under the assumption of conditional independence among features. This "naïve" assumption simplifies computation but is often violated in practical applications. Despite this limitation, the naïve Bayes remains a robust and computationally efficient classifier that is widely applied in domains such as text categorization and medical diagnosis. In a study by Montazeri et al,⁴ a hybrid predictive model was developed by integrating rule-based strategies with

multiple ML techniques including naïve Bayes and RF to estimate BC survival outcomes. Their findings revealed that, while achieving a high accuracy of 95%, it was marginally outperformed by the RF model.

The kNN algorithm classifies a data point based on the classes of its k closest neighbors within the dataset. As a supervised ML technique, kNN is applicable to both classification and regression tasks. For each new input, the algorithm calculates the distance between the test point and all data points in the training set, and selects the k closest instances to inform the prediction. Features exhibiting greater variability contributed more significantly to distance calculation. In a study conducted by Lou et al,²⁰ kNN was used to predict BC recurrence within a 10-year period after surgery. This study compared several ML models and found that kNN ranked third in terms of predictive accuracy. Consistent with previous research, the tumor stage was confirmed to be a strong predictor of recurrence. However, this study also identified additional variables, such as surgeon and hospital volumes as significant predictors, underscoring the potential of ML to reveal novel prognostic factors beyond those typically captured by conventional statistical models.

SVM is a well-established supervised ML algorithm, although its application in cancer prediction and prognosis studies remains relatively limited compared

to other models. SVM operates by identifying the optimal hyperplane that best separates data points from different classes, thereby maximizing the margin between them. For example, in a dataset in which tumor volume is plotted against the number of axillary lymph node metastases in patients with BC, SVM can be utilized to distinguish between individuals with favorable and poor prognoses (Figure 4). In addition to classification tasks, SVM is widely applied to model numerical (continuous) data in regression contexts, particularly in biomedical research, where high-dimensional datasets are common.

This illustration clearly demonstrates that the two data groups can be separated effectively. The SVM algorithm functions by identifying the optimal decision boundary, referred to as a hyperplane, that maximally distinguishes between two clusters. When additional data points are plotted for variables, such as tumor volume, the slope of this boundary tends to decrease, resulting in a flatter orientation. As more features are incorporated into the model, the dimensionality of the separation increases and the classification is determined by a multidimensional hyperplane. This hyperplane is constructed based on a subset of critical data points from each class, called ‘support vectors’, which play a pivotal role in defining the margin. The primary objective of the SVM algorithm is to establish a hyperplane that maximizes the margin between the two classes, thereby enhancing classification accuracy and generalizability. In a study utilizing the WDBC dataset, Naji et al²⁶ employed five different ML algorithms to predict BC diagnoses and outcomes and reported that the SVM demonstrated a performance comparable to that of the RF algorithm.

ANNs have demonstrated substantial capabilities in addressing a wide array of classification and pattern recognition tasks. A notable strength of ANNs lies in their ability to perform diverse statistical analyses, including linear, logistic, and nonlinear regression, as well as to execute logical functions such as AND, OR, XOR, NOT, and IF-THEN, which support their classification performance. ANNs are inspired by the human brain and comprise numerous interconnected processing units (neurons) linked via weighted pathways, analogous to biological axons. These connection weights are adaptively modified through iterative training, thereby simulating the learning processes of biological neural systems. Information within a network is propagated through

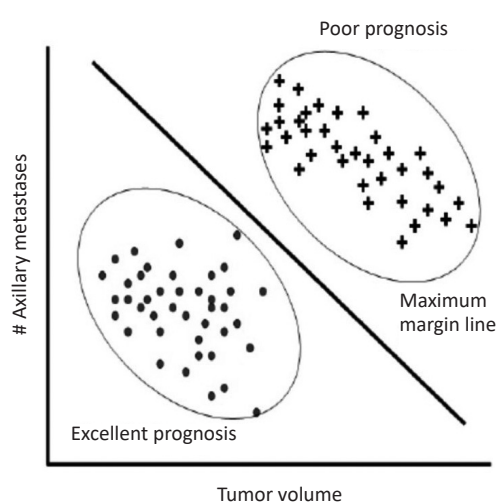


Figure 4. A simplified illustration of how support vector machine (SVM) can differentiate between two prognostic outcomes based on a data distribution. In this simple case, the SVM identifies a hyperplane (a line) that maximizes the separation between the two clusters

multiple hidden layers, each of which incrementally transforms the input before forwarding it to the next stage. To fulfill the mathematical requirements of these operations, both the input and output data were represented as vectors of numerical values. In the study reviewed herein, Lou et al²⁰ assessed the predictive accuracy of various models in forecasting BC recurrence over a 10-year postoperative period and concluded that ANNs achieved the highest accuracy among the evaluated ML approaches.

DL is an advanced subset of ANNs. Traditional ANN architectures are characterized by one or two hidden layers, and are predominantly applied within the framework of supervised ML. By contrast, DL distinguishes itself by incorporating a deeper architecture comprising multiple hidden layers, enabling it to handle both supervised and unsupervised ML tasks. DL has been predominantly employed in classification and segmentation tasks, particularly in medical imaging. However, their application in predicting post-radiotherapy outcomes remains relatively limited. In a study conducted by Zhu et al,¹² the predictive value of combining C-statistics with DL was explored to assess BC risk. This study compared cohorts based on the availability of both mammographic imaging data and clinical risk factors, with clinical data alone. The findings demonstrated that integrating DL-based analysis of mammographic abnormalities with clinical risk factors significantly improved risk prediction in patients with radiographic abnormalities.

Prior reviews in BC prediction typically report that tree ensembles and SVMs dominate on structured/tabular clinical data, whereas DL is more prevalent in imaging-centric tasks; across reviews, consistent gaps include limited external validation and sparse calibration/utility reporting. Our results are concordant with these patterns and reinforce the need for standardized reporting and decision-analytic evaluations. Our findings showed that RF appeared most frequently and often with strong accuracy, which aligns with its variance reduction design and suitability for mixed clinical predictors. However, frequency is an ecosystem signal (e.g., ease of implementation, sensible defaults), not a proof of universal optimality, and conclusions should emphasize like-for-like comparisons of shared endpoints and externally validated, well-calibrated models rather than algorithm labels. Nonetheless, consistent with

its ensemble architecture, RF has demonstrated promising performance in predicting BC survival and performing treatment-related analyses, although its applicability remains dependent on local calibration and threshold selection.

We limited the search to PubMed (MEDLINE) because of its focus on biomedical literature and resource constraints, which may have missed engineering/technical publications (Excerpta Medica Database, Scopus, or Institute of Electrical and Electronics Engineers/Association for Computing Machinery) and gray literature. The PRISMA extension for scoping reviews emphasizes transparent, comprehensive multisource strategies; hence, the review is vulnerable to selection bias and may miss gray literature and non-indexed studies. Across the studies, common weaknesses included reliance on internal validation, limited calibration reporting, heterogeneous endpoints and time horizons, small event counts, and inconsistent adherence to reporting standards. Researchers should register protocols and search multiple databases, report according to TRIPOD/TRIPOD+AI, extract and compare models using CHARMS, assess the RoB and applicability with the prediction model RoB assessment tool, prioritize external (temporal/geographic) validation with full calibration reporting and DCA, test transportability, and update models for local use where needed.^{5,6}

ML models, which represent the processes of knowledge extraction from data and the discovery of hidden relationships (correlations), have been increasingly applied in healthcare in recent years to predict disease outcomes. Numerous prognostic models have been developed to predict BC outcomes after surgery, including those that estimate the risk of recurrence within 10 years postoperatively, each with its own strengths and limitations. Given the rapid expansion of real-world data, the identification of widely adopted and highly accurate predictive models for assessing BC risk and posttreatment outcomes is of critical importance.

This review found that RF was the most frequently utilized ML algorithm for predicting cancer incidence and prognosis, and was consistently associated with high predictive performance across studies. Ensemble-based approaches are commonly applied to heterogeneous clinical datasets and have demonstrated robust classification capabilities. These findings suggest that RF is a

practical and reliable method for cancer prediction. Future research should emphasize standardized reporting, external validation, and methodological transparency to improve the reproducibility and clinical applicability.

Conflict of Interest

The authors affirm no conflict of interest in this study.

Acknowledgment

The authors thank the peer reviewers and editors for constructive feedback, which substantially strengthened this scoping review.

Funding Sources

None.

REFERENCES

- Santoro M, Strolin S, Paolani G, Della Gala G, Bartoloni A, Giacometti C, et al. Recent applications of artificial intelligence in radiotherapy: where we are and beyond. *Appl Sci*. 2022;12(7):3223.
- Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. *BMC Cancer*. 2019;19(1):230.
- Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, et al. Machine learning techniques in breast cancer prognosis prediction: a primary evaluation. *Cancer Med*. 2020;9(9):3234–43.
- Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technol Health Care*. 2016;24(1):31–42.
- Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
- Ding X, Xie S, Chen J, Mo W, Yang H. A support vector machine model for predicting non-sentinel lymph node status in patients with sentinel lymph node positive breast cancer. *Tumour Biol*. 2013;34(3):1547–52.
- Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak*. 2019;19(1):48.
- Xu Y, Ju L, Tong J, Zhou C, Yang J. Supervised machine learning predictive analytics for triple-negative breast cancer death outcomes. *Onco Targets Ther*. 2019;12:9059–67.
- Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer*. 2018;18(1):610.
- Kim I, Choi HJ, Ryu JM, Lee SK, Yu JH, Kim SW, et al. A predictive model for high/low risk group according to oncotype DX recurrence score using machine learning. *Eur J Surg Oncol*. 2019;45(2):134–40.
- Zhu X, Wolfgruber TK, Leong L, Jensen M, Scott C, Winham S, et al. Deep learning predicts interval and screening-detected cancer from screening mammograms: a case-case-control study in 6369 women. *Radiology*. 2021;301(3):550–8.
- Pawloski KR, Gonen M, Wen HY, Tadros AB, Thompson D, Abbate K, et al. Supervised machine learning model to predict oncotype DX risk category in patients over age 50. *Breast Cancer Res Treat*. 2022;191(2):423–30.
- Lindsay WD, Ahern CA, Tobias JS, Berling CG, Chinniah C, Gabriel PE, et al. Automated data extraction and ensemble methods for predictive modeling of breast cancer outcomes after radiation therapy. *Med Phys*. 2019;46(2):1054–63.
- Wu M, Zhong X, Peng Q, Xu M, Huang S, Yuan J, et al. Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. *Eur J Radiol*. 2019;114:175–84.
- Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*. 2019;292(2):331–42.
- Kim J, Yu D, Kwon Y, Lee KS, Sim SH, Kong SY, et al. Genomic characteristics of triple-negative breast cancer nominate molecular subtypes that predict chemotherapy response. *Mol Cancer Res*. 2020;18(2):253–63.
- Nicolò C, Périer C, Prague M, Bellera C, MacGrogan G, Saut O, et al. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO Clin Cancer Inform*. 2020;4:259–74.
- Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLoS One*. 2020;15(10):e0237658.
- Lou SJ, Hou MF, Chang HT, Chiu CC, Lee HH, Yeh SJ, et al. Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: a prospective cohort study. *Cancers*. 2020;12(12):3817.
- Liu Z, Mi M, Li X, Zheng X, Wu G, Zhang L. A lncRNA prognostic signature associated with immune infiltration and tumour mutation burden in breast cancer. *J Cell Mol Med*. 2020;24(21):12444–56.
- Chen X, Zheng J, Zhuo ML, Zhang A, You Z. A six-gene-based signature for breast cancer radiotherapy sensitivity estimation. *Biosci Rep*. 2020;40(12):BSR20202376.
- Pellegrini M. Accurate prediction of breast cancer survival through coherent voting networks with gene expression profiling. *Sci Rep*. 2021;11(1):14645.
- Wu R, Luo J, Wan H, Zhang H, Yuan Y, Hu H, et al. Evaluation of machine learning algorithms for the prognosis of breast cancer from the surveillance, epidemiology, and end results database. *PLoS One*. 2023;18(1):e0280340.
- Hou C, Zhong X, He P, Xu B, Diao S, Yi F, et al. Predicting breast cancer in Chinese women using machine learning techniques: algorithm development. *JMIR Med Inform*. 2020;8(6):e17364.
- Naji MA, Filali SE, Aarika K, Benlahmar EL, Abdelouahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*. 2021;191:487–92.
- Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. *J Biomed Phys Eng*. 2022;12(3):297–308.
- Saleh H, Abd-El Ghany SF, Alyami H, Alosaimi W. Predicting breast cancer based on optimized deep learning approach. *Comput Intell Neurosci*. 2022;2022:1820777.