Machine learning techniques in breast cancer preventive diagnosis: a review

Giada Anastasi[1,4] · Michela Franchini[1] · Stefania Pieroni[1] · Marina Buzzi[2] ·
Maria Claudia Buzzi[2] · Barbara Leporini[3] · Sabrina Molinaro[1]

Abstract

Breast cancer (BC) is known as the most prevalent form of cancer among women. Recent research has demonstrated the potential of Machine Learning (ML) techniques in predicting the five-year BC risk using personal health data. Support Vector Machine (SVM), Random Forest, K-NN (K-Nearest Neighbour), Naive Bayes, Neural Network, Decision Tree (DT), Logistic Regression (LR), Discriminant Analysis, and their variants are commonly employed in ML for BC analysis. This study investigates the factors influencing the performance of ML techniques in the domain of BC prevention, with a focus on dataset size and feature selection. The study's goal is to examine the effect of dataset cardinality, feature selection, and model selection on analytical performance in terms of Accuracy and Area Under the Curve (AUC). To this aim, 3917 papers were automatically selected from Scopus and PubMed, considering all publications from the previous 5 years, and, after inclusion and exclusion criteria, 54 articles were selected for the analysis. Our findings highlight how a good cardinality of the dataset and effective feature selection have a higher impact on the model's performance than the selected model, as corroborated by one of the studies, which gets extremely good results with all of the models employed.

Keywords Breast cancer · Machine learning · Preventive diagnosis · Random forest · Support vector machine

Abbreviations

FFDM	Full-Field digital mammography
ABUS	Automated breast ultrasound screening
CT	Computed tomography
UWB	Ultra-WideBand
H&E	Hematoxylin and eosin
MRI	Magnetic resonance imaging
DCE-MRI	Dynamic contrast-enhanced MRI
mpMRI	MultiParametric MRI
ADH	Atypical ductal hyperplasia
US	UltraSound
QUS	Quantitative US

Extended author information available on the last page of the article

qCT	Quantitative CT
FFPE	Formalin-fixed paraffin-embedded
BCIMS	Breast cancer information management system
CEDM	Contrast-enhanced digital mammography
WDBC	Wisconsin diagnostic breast cancer
CBC	Coimbra breast cancer

1 Introduction

1.1 Historical review

Cancer is recognized as a significant healthcare challenge by the Horizon Europe program [1]. Among female cancers, Breast Cancer (BC) is the most prevalent, with an incidence rate of 5 cases per 1,000 women, as extensively documented in the literature [2–8]. In the European Union (EU) in 2020, 2.7 million BC cases were diagnosed, resulting in 1.3 million deaths. The World Health Organization (WHO) guidelines strongly recommend optimising cancer treatment and care [9]. The "European Commission Cancer Plan" highlights the crucial role of cancer prevention and treatment optimization [10]. It also provides information on the allocation of funds for cancer research on early detection and introduces a new "EU supported Cancer Screening Scheme" aiming to offer screening to 90% of the EU population by 2025. As an immediate objective, the European Commission plans to propose an update to the Council Recommendation on cancer screening by 2022, incorporating the most recent scientific evidence. The updated recommendation suggests expanding cancer screening campaigns beyond breast, colorectal, and cervical cancer to include prostate, lung, and gastric cancer. Furthermore, the Commission proposes identifying criteria to target screening based on personal risk and characteristics rather than just age.

BC is categorised into three subtypes based on the presence or absence of molecular markers for estrogen receptor (ER) or progesterone receptor (PR) and human epidermal growth factor 2 (ERBB2 or HER2). Specifically, hormone receptor positive/ERBB2 negative cancers account for 70% of all BCs, ERBB2 positive accounts for 15%-20%, and triple-negative for 15% [11]. More than 90% of BC cases are non-metastatic at the time of diagnosis, and the therapeutic goals in such cases include tumor eradication and prevention of recurrence.

As shown in Table 1, BC mortality exhibits significant geographical variability [12–15], influenced by factors such as population structure, lifestyle, genetics, and the environment [16]. The 5-year net survival rate after BC diagnosis also varies and reaches 87% in developed countries where screening and early diagnosis are practised [17].

Risk factors for BC incidence and mortality can be classified into two groups: genetic risk factors (such as BRCA1 and BRCA2) and non-genetic risk factors (age at menarche, menopause, childbearing, breastfeeding, mammography density, overweight and obesity, physical inactivity, alcohol consumption, and lifestyle choices).

Breast Cancer Risk Models [18] utilise a model-driven analysis that incorporates a combination of several factors. However, aside from female gender and increasing patient age, certain risk factors have shown weak effects on BC, necessitating a large amount of data for accurate evaluation [19]. Data-driven analysis approaches in the field of Artificial Intelligence (AI) offer the potential to more effectively identify combinations of risk factors that contribute to increased BC incidence. These approaches leverage AI techniques to analyse

Table 1 Mean (SD) for breast cancer mortality rate for each IHME super region from 1995 to 2015, from [15]

Super Region	Year					
	1990	1995	2000	2005	2010	2015
Sub-Saharan Africa	6.63 (2.52)	6.67 (2.46)	6.99 (2.76)	7.11 (3.16)	7.18 (3.00)	7.66 (3.24)
North Africa Middle East	6.91 (3.15)	7.58 (3.43)	8.17 (3.67)	8.70 (3.99)	9.12 (4.06)	9.72 (4.00)
South Asia	6.12 (3.49)	6.79 (4.29)	7.35 (4.81)	7.80 (5.17)	8.30 (5.20)	9.14 (5.25)
Southeast Asia East Asia Oceania	9.06 (4.40)	10.11 (4.90)	11.35 (5.43)	12.53 (5.63)	13.72 (5.95)	14.70 (6.47)
Latin America Caribbean	12.12 (7.16)	13.53 (8.01)	14.94 (8.92)	16.08 (9.36)	17.95 (10.39)	19.64 (11.25)
Central Europe Eastern Europe Central Asia	20.43 (9.74)	22.63 (10.76)	24.27 (11.45)	25.33 (11.49)	26.02 (11.25)	26.99 (11.61)

and derive insights from extensive datasets, allowing for the identification of complex relationships and interactions among various risk factors. By harnessing the power of AI, we can enhance our understanding of BC risk and potentially improve the accuracy of risk prediction models.

1.2 Brief summary of the most recent studies

In recent years, Machine Learning (ML) techniques have emerged as one of the most prominent topics in the fields of Information Technology (IT) and Artificial Intelligence (AI). ML has experienced continuous growth and its applications extend across various domains, including pattern recognition, computer vision, finance, entertainment, computational biology, as well as biomedical and medical applications [20, 21]. ML represents an engineering approach that aims to enhance the ability to extract valuable information from data itself, without relying heavily on external inputs or prior knowledge. The primary objective of ML is to develop and refine models that can be trained using context-specific data, enabling decision-making without complete knowledge of external factors. The process of ML involves two essential steps: training and inference. During the training phase, an ML algorithm processes a dataset and identifies the function that best captures the underlying patterns in the data. This function is then encoded and referred to as the model, which is subsequently employed to extract knowledge from new data instances [22].

1.3 Opening problems under investigation

In recent years, significant advancements have been made in applying ML techniques to healthcare, as extensively documented in the literature [23]. Previous studies have demonstrated that augmenting the widely-used Gail risk model with additional inputs improves its ability to predict BC risk.

The Gail model incorporates six breast cancer risk factors, namely: age, age at menarche, age at first live birth, number of breast biopsies, history of atypical hyperplasia, and number of first-degree relatives with breast cancer. Based on this information the model provides the individual estimate of BC risk. Based on the Gail model, women with a breast cancer risk of > 1.66% were considered as high-risk according to the estimated 5-year breast cancer- risk assessment [24].

However, these models, including Gail, typically rely on simple statistical architectures and incorporate inputs obtained from expensive and/or invasive procedures. In contrast, recent studies [25] have presented ML models that utilise readily available personal health data to predict BC risk over a five-year period. Many of these studies have compared the accuracy of different models based on various ML algorithms and techniques, such as Random Forest (RF), K-Nearest Neighbour (K-NN), Naive Bayes (NB), Neural Network (NN), Decision Tree (DT), Logistic Regression (LR), Discriminant Analysis (DA), or Support Vector Machine (SVM) [26–28].

ML methods employed for tumor identification, classification, detection, or differentiation have demonstrated highly competitive results [29].

This review primarily focuses on the potential role of Artificial Intelligence (AI) in supporting BC prevention and the challenges that need to be addressed to enhance operational quality. Specifically, this work aims to analyse various ML techniques applied in the field of early detection of BC. To achieve this objective, recent papers (from 2017 to 2022) employing these techniques were collected and compared to determine the optimal combination of data types, feature extraction methods, and models that yield the most accurate results. Additionally, a secondary goal is to investigate the reasons behind the preference for certain ML techniques while neglecting others.

Aside from the primary goal of investigating the use of ML techniques in research studies, a secondary goal of this research initiative is to conduct a thorough investigation into the underlying factors that lead researchers to prefer certain ML methodologies while ignoring others. The availability and quality of training data, computational resource constraints, the established body of previous research in the respective field, the inherent complexity of the problem to be addressed, and the potential interpretability and explainability of the chosen ML models are all factors to consider when selecting machine learning techniques. The ML models used by all the authors of the publications included in this review will be examined in the discussion chapter, with an emphasis on how the most commonly used models have changed over time.

2 Methods

2.1 Search and selection of literature

The studies included in this review were identified through a systematic literature review conducted on PubMed and Scopus databases until December 2022. The search included articles published between 2017 and 2022. The search terms used were: "[Model name]" AND "machine learning" AND "breast cancer" AND "validation" AND ("prevention" OR "diagnosis" OR "risk analysis") AND "AUC" AND "accuracy" AND PUBYEAR > 2016. Only full-text documents written in English that defined validation methods and presented performance results in terms of Area Under the Curve (AUC) and accuracy were considered eligible for inclusion. Some articles have been excluded if they did not present either

of the two-performance metrics mentioned above, articles suggesting the use of Deep Learning (DL) models instead of ML models, articles unrelated to BC, and articles that did not specify the dataset size.

DL is a subclass of ML that is a data-driven technique for learning features and tasks. The term 'deep' refers to the various layers of algorithms that data passes through during computing to construct a neural network. This study decided to remove DL algorithms, which we know require a lot bigger quantity of data than typical ML algorithms, in order to compare datasets with a higher cardinality to each other. The distinctions between the two modalities are adequately highlighted in Section 2.1 of the paper [30], where it is stated that there are significant disparities between the two modalities in both the approach and the description of the data required.

Figure 1 illustrates the search queries used in the PubMed and Scopus databases to retrieve articles related to the early detection and prevention of breast cancer using ML algorithms tested between 2017 and 2022.

Model validation is a critical step in the process that ensures the effectiveness of the developed model. It involves evaluating the model's performance using an external dataset known as the validation set. The validation set is separate from the training data and is used to assess the quality and fit of the model's results. In the reviewed papers, the majority of studies employed the cross-validation method for model validation.

Cross-validation is a resampling technique where different subsets or partitions of the data are used for training and testing the model. There are several variations of cross-validation based on how the data is divided and utilised. One of the most commonly used approaches is the tenfold cross-validation, as depicted in Fig. 2 in [31]. In this method, the data is divided into 10 equal-sized subsets or folds. The model is then trained on 9 folds and tested on the remaining fold. This process is repeated 10 times, with each fold serving as the test set once. The results from each iteration are aggregated to assess the overall performance of the model.

Figure 2 in [31] provides an illustration of the tenfold cross-validation approach, highlighting the repeated training and testing steps with different subsets of the data.
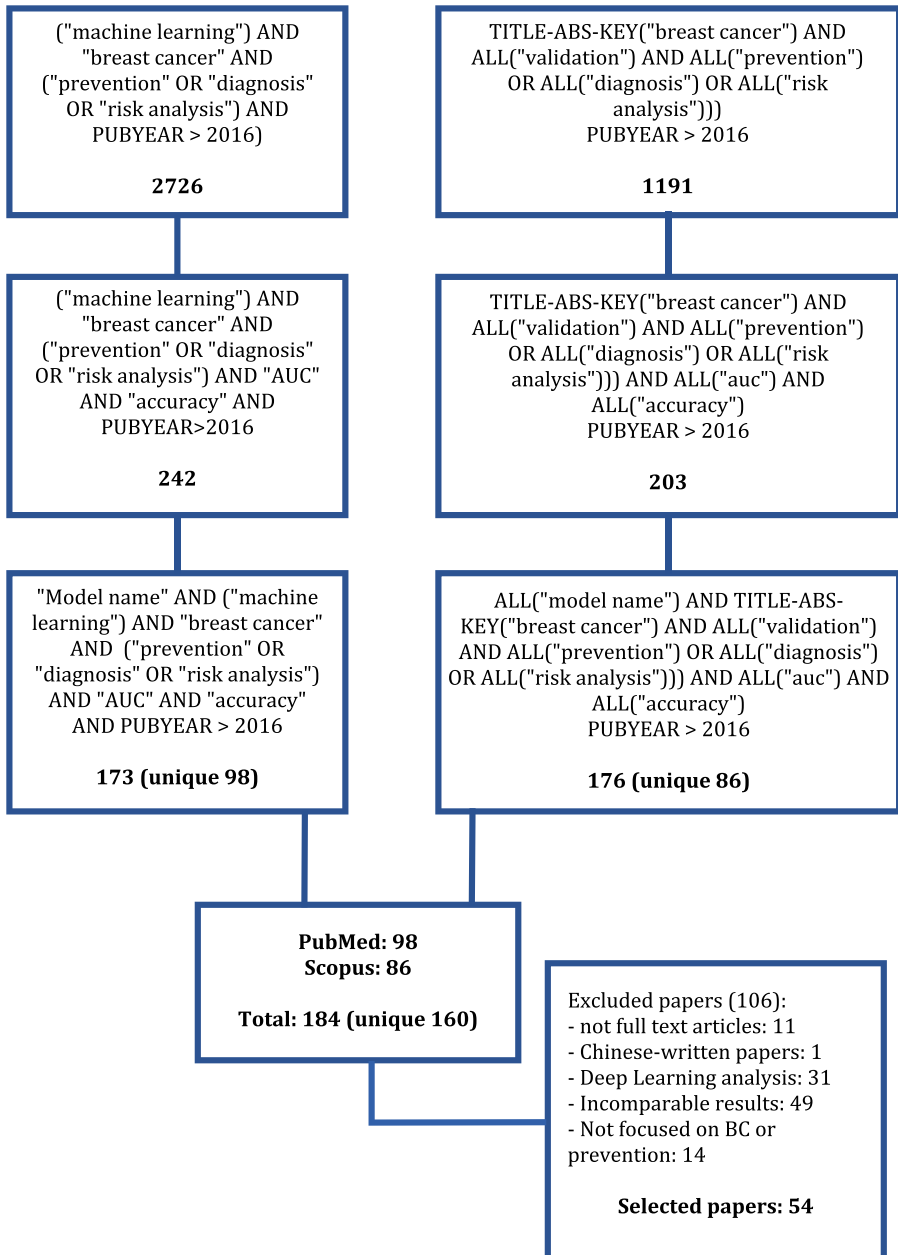
2.2 ML performance metrics considered for the paper selection

The performance evaluation of the ML techniques was conducted by comparing them in terms of two key metrics: Area Under the Curve (AUC) and accuracy.

Accuracy is an important and intuitive performance measure in evaluating classification models. It represents the ratio of correctly predicted observations (True Positives and True Negatives) to the total number of observations. In the context of a binary classification problem, True Positives (TP) and True Negatives (TN) correspond to the correctly classified instances of the positive and negative classes, respectively. False Positives (FP) and False Negatives (FN) represent the instances that are incorrectly classified as positive and negative, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Although accuracy is a regularly used metric to evaluate classifier performance, it may be insufficient to provide a thorough evaluation, particularly in circumstances involving unbalanced datasets. When dealing with imbalanced classes, a classifier that predicts the majority class for all occurrences can nevertheless generate a high rate of

```
┌─────────────────────────────┐         ┌─────────────────────────────┐
│  ("machine learning") AND   │         │ TITLE-ABS-KEY("breast cancer")│
│    "breast cancer" AND      │         │  AND ALL("validation") AND  │
│ ("prevention" OR "diagnosis"│         │ ALL("prevention") OR         │
│ OR "risk analysis") AND     │         │ ALL("diagnosis") OR ALL("risk│
│   PUBYEAR > 2016)           │         │   analysis")))              │
│                             │         │   PUBYEAR > 2016            │
│          2726               │         │          1191               │
└─────────────────────────────┘         └─────────────────────────────┘
```

```
┌─────────────────────────────┐         ┌─────────────────────────────┐
│  ("machine learning") AND   │         │ TITLE-ABS-KEY("breast cancer")│
│    "breast cancer" AND      │         │  AND ALL("validation") AND  │
│ ("prevention" OR "diagnosis"│         │ ALL("prevention") OR         │
│ OR "risk analysis") AND "AUC"│        │ ALL("diagnosis") OR ALL("risk│
│   AND "accuracy" AND        │         │ analysis"))) AND ALL("auc") AND│
│   PUBYEAR>2016              │         │   ALL("accuracy")           │
│                             │         │   PUBYEAR > 2016            │
│          242                │         │          203                │
└─────────────────────────────┘         └─────────────────────────────┘
```

```
┌─────────────────────────────┐         ┌─────────────────────────────┐
│  "Model name" AND ("machine │         │ ALL("model name") AND TITLE-ABS-│
│ learning") AND "breast cancer"│       │ KEY("breast cancer") AND ALL("validation")│
│ AND  ("prevention" OR        │        │ AND ALL("prevention") OR ALL("diagnosis")│
│ "diagnosis" OR "risk analysis")│      │ OR ALL("risk analysis"))) AND ALL("auc") AND│
│ AND "AUC" AND "accuracy"     │        │   ALL("accuracy")           │
│ AND PUBYEAR > 2016          │         │   PUBYEAR > 2016            │
│                             │         │                             │
│     173 (unique 98)         │         │     176 (unique 86)         │
└─────────────────────────────┘         └─────────────────────────────┘
```

```
        ┌─────────────────────────┐
        │     PubMed: 98          │
        │     Scopus: 86          │        ┌──────────────────────────────┐
        │                         │        │ Excluded papers (106):        │
        │  Total: 184 (unique 160)│        │ - not full text articles: 11  │
        └─────────────────────────┘        │ - Chinese-written papers: 1   │
                                           │ - Deep Learning analysis: 31  │
                                           │ - Incomparable results: 49    │
                                           │ - Not focused on BC or        │
                                           │ prevention: 14                │
                                           │                               │
                                           │    Selected papers: 54        │
                                           └──────────────────────────────┘
```
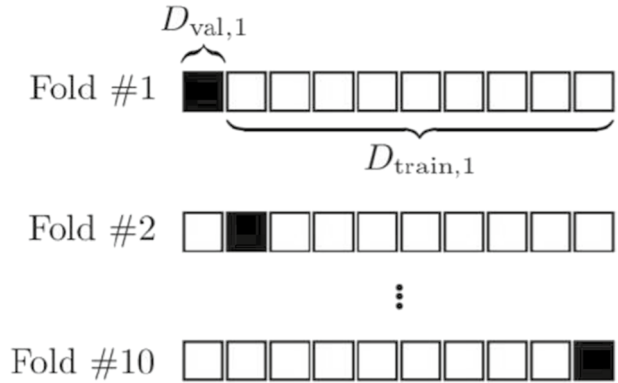
Fig. 1 Papers selection flow

accuracy even if it fails to identify the minority class adequately. This is the most obvious limitation in depending solely on accuracy as a performance metric.

To solve this constraint, more measures must be incorporated into the review process.

Fig. 2 10-fold-cross-validation, modified from [31]

The Area Under the Receiver Operating Characteristic Curve (AUC) being visually appealing and providing an overview of a classifier's performance across a wide range of specificities, is the performance measure most frequently used within the ML studies working with imbalanced datasets [32].

Additionally, by incorporating the AUC alongside accuracy, practitioners can acquire a more nuanced and trustworthy picture of classifier performance, which is especially important in cases when class imbalances are widespread.

The AUC is a measure of the classifier's ability to distinguish between different classes, and it provides a summary of the Receiver Operator Characteristic (ROC) curve (see Fig. 3). The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds. The AUC represents the area under this curve corresponding to the probability that the model will rate a random positive case higher than a random negative example.

Fig. 3 Example of ROC curve, with False Positive Rate (FPR) on x-axis and True Positive Rate (TPR) on y-axis
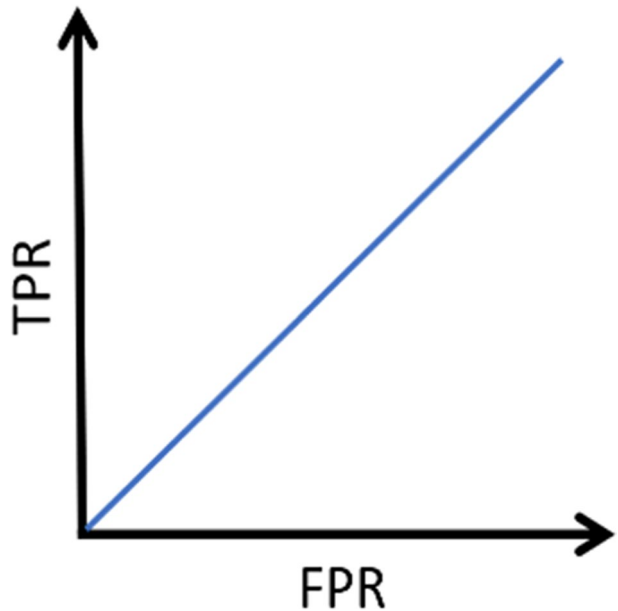
A higher AUC indicates a better performance of the model in distinguishing between the positive and negative classes. When the AUC is equal to one, the classifier achieves perfect discrimination between all the positive and negative class instances. Conversely, an AUC value of zero suggests that the classifier incorrectly predicts all negatives as positives and all positives as negatives. Higher AUC values indicate better performance.

It is worth noting that AUC, as well as accuracy, is affected by distortions due to the problem of imbalanced training data occurring frequently in bioinformatics. When ML methods are trained on very imbalanced data sets, they often tend to produce majority classifiers – over-predicting the presence of the majority class being mild levels of imbalance – at 30–40% of the data in the minority class – sufficient to alter the values of the measures commonly used to assess models performance. When large amounts of data in the minority class are easy to obtain, some authors suggested to undersample the majority class and effectively balance the data sets.The same authors also suggested when these data are sparse, then bioinformatics researchers would do well to consider the oversampling and cost-sensitive learning techniques, developed in machine learning in recent years [33, 34]. Furthermore, Saitto and colleagues, focusing on the performance evaluation of the final ML models, proposed some ROC alternatives as the Concentrated ROC (CROC), the Cost Curves (CC), and the Precision/Recall (PRC) plots. The authors concluded that being the PRC the only visual analysis tool that changes with the ratio of positives and negatives it represents the most informative one [32].

3 Results

A total of 184 articles were initially selected, with 86 retrieved from Scopus and 98 from PubMed. After removing duplicates, the remaining 160 papers underwent screening and evaluation. Among these, 106 articles were excluded based on specific criteria. This included 11 articles that were not available in full-text, 1 article written in Chinese, 31 articles that focused on Deep Learning (DL) or other techniques instead of Machine Learning (ML) techniques, 49 articles with incomparable results, and 14 articles that either focused on the prevention of morbidities other than BC or were not specifically focused on prevention.

After applying these exclusion criteria, a total of 54 papers were included in the review. The majority of these papers utilised the LR model, accounting for 22.4% of the included studies. The SVM model followed with 18.3% representation, while the RF model accounted for 13.8%. The remaining papers, approximately 45.5%, employed various other ML techniques. The distribution of these works across different methods and years of publication is depicted in Fig. 4.

Although the majority of the analysed studies were conducted in the USA, the patient populations in the included papers encompassed individuals from other countries such as China, Japan, Africa, and Iran, indicating a broader geographical representation in the research.

This information highlights the selection process, the distribution of ML techniques used in the included papers, and the geographical diversity of the studied patient populations.

The temporal distribution of papers based on the ML models provides insights into the growing interest in ML methods, particularly the notable increase in publications focusing on Logistic Regression (LR) since 2021. The high number of papers utilising the SVM model could be attributed to its frequent use as a benchmark when evaluating the
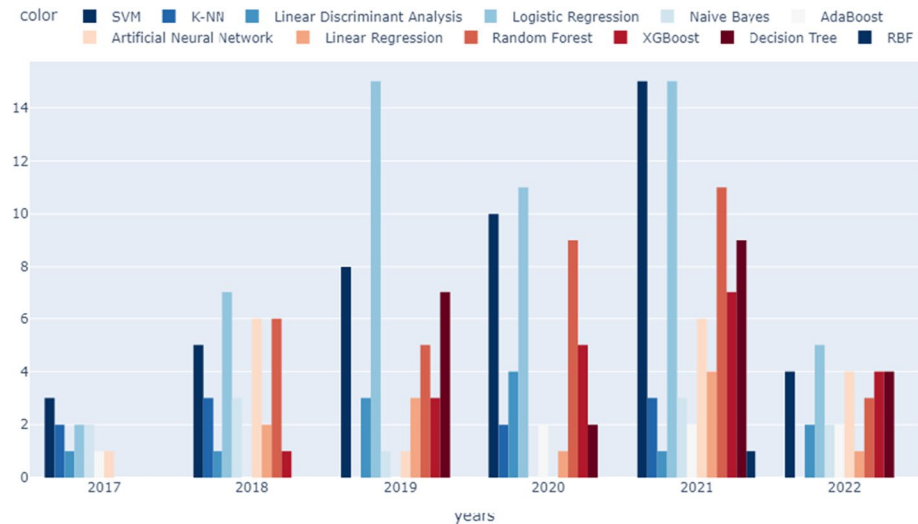
Fig. 4 Distribution of different methods used in selected related works in years

performance of other algorithms. Additionally, the number of studies utilising the RF technique has also increased in recent years. RF is often compared to DT to showcase the differences in results when using a single DT versus combining multiple DTs to obtain the final outcome.

The results of the 59 selected articles are organised by ML methods in Tables [2–15]. Each table provides the main characteristics of the study and presents the accuracy and AUC results. These metrics serve as indicators of the performance achieved by the respective ML models.

By examining these tables, readers can gain an understanding of the characteristics and outcomes of studies conducted using different ML techniques, allowing for comparisons and evaluations based on accuracy and AUC performance measures.

3.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning algorithms commonly used for binary classification tasks. In SVMs, a hyperplane is established to separate the sample items into two classes, as illustrated in Fig. 5. The hyperplane established by SVMs is determined by a subset of data points called support vectors, which lie closest to the decision boundary. These support vectors play a crucial role in defining the hyperplane and determining the classification boundaries.

The goal is to find the optimal hyperplane that maximises the margin between the two classes, ensuring the best generalisation ability.

By maximising the margin, SVMs aim to achieve high accuracy not only on the training set but also on new, unseen data that may be added to the dataset in the future [35]. This ability to generalise well to new data is a key advantage of SVMs.

Overall, SVMs are powerful tools for binary classification, providing an effective means of separating data into distinct classes by finding an optimal hyperplane that maximises the margin between them.

Fig. 5 Optimal hyperplane in SVM, for a binary classification. Modified from [35]
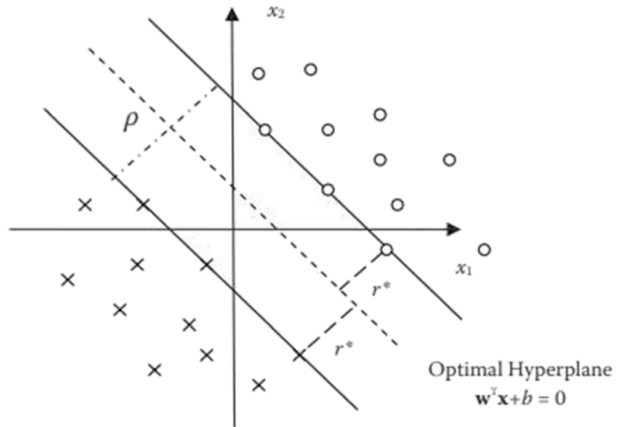
Articles concerning the application of SVM models in the BC prevention field are summarised in Table 2, sorted by dataset cardinality.

SVM models in some cases are combined with Fast Fourier Transform (FFT), in other cases with Discrete Cosine Transform (DCT), with Structural Similarity Index (SSIM) or with Sequential Forward Feature Selection (SFSS) as feature selection techniques. Paper [45] combines SVM with PET (Positron Emission Tomography) features and CT (Computed Tomography) features. Paper [56] instead applies SVM to Quantitative UltraSound (QUS) features. In other cases, SVM is combined with Semi-Supervised Learning (SSL) or Supervised Learning (SL) techniques.

The selected papers in the field of BC prevention propose various approaches to achieve optimal performance. Some common strategies include feature selection, modifying the algorithm to fit the data, and selecting data to suit the specific model being used.

In study [36], the SVM model is combined with three different feature selection techniques: the **SSIM**, which quantifies image quality degradation due to compression or data transmission losses, the **DCT**, which transforms pixel information from spatial domain to frequency domain, and the **FFT**, which computes the discrete Fourier transform of input sequences. These techniques aim to identify important patterns and information in the input images.

Another approach is presented in a different paper [37], where the dataset is divided into a modelling dataset and an external verification dataset. The authors selected 75% of the samples from the modelling dataset as the training set. They employed variable selection, one-hot encoding, and a basic model, which were assembled into a pipeline. This pipeline was then entered into grid search using the tenfold cross-validation technique, allowing for thorough evaluation and optimization of the model's performance.

The selected papers [40, 42, 44, 54] introduce specific variations of the SVM model, different from the standard version, and investigate their impact on performance. For instance, paper [61] achieved the best accuracy and AUC values by using SVM with the quadratic kernel function (SVM_Q), while the worst performance was observed when using SVM with the linear kernel function (SVM_L). This indicates that different SVM models applied to the same data can have varying effects on performance outcomes.

In contrast, paper [39] employed the linear version of the SVM and combined it with both SL and SSL techniques. The performance achieved in this case was significantly

Table 2 Support vector machine results

Authors	Paper	Model	Accuracy	AUC	Dataset
Heidari et al. 2019	[36]	SVM + FFT features without chest removal	72%	0.70	1959 FFDM
Heidari et al. 2019	[36]	SVM + FFT features with chest removal	77%	0.84	1959 FFDM
Heidari et al. 2019	[36]	SVM + DCT features without chest removal	69%	0.68	1959 FFDM
Heidari et al. 2019	[36]	SVM + DCT features with chest removal	83%	0.89	1959 FFDM
Heidari et al. 2019	[36]	SVM + SSIM features without chest removal	61%	0.62	1959 FFDM
Heidari et al. 2019	[36]	SVM + SSIM features with chest removal	71%	0.78	1959 FFDM
Heidari et al. 2019	[36]	SVM + fusion without chest removal	83%	0.79	1959 FFDM
Heidari et al. 2019	[36]	SVM + fusion with chest removal	89%	0.94	1959 FFDM
Huo et al. 2021	[37]	SVM + test set	71%	0.64	1125 ABUS
Huo et al. 2021	[37]	SVM + validation set	77%	0.85	1125 ABUS
Park et al. 2019	[38]	SVM + lymph node	57%	0.34	723 CT
Park et al. 2019	[38]	SVM + Tumor grade	63%	0.35	723 CT
Park et al. 2019	[38]	SVM + Tumor size	48%	0.42	723 CT
Park et al. 2019	[38]	SVM + ER	67%	0.44	723 CT
Park et al. 2019	[38]	SVM + PR	64%	0.39	723 CT
Park et al. 2019	[38]	SVM + Ki67	53%	0.35	723 CT
Park et al. 2019	[38]	SVM + Molecular subtype	41%	0.65	723 CT
Al-Azzam et al. 2021	[39]	SVM LINEAR + SL	97%	0.97	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	SVM LINEAR + SSL	97%	0.96	569 patients from WDBC dataset
Jiang et al. 2021	[40]	SVM	70%	0.68	243 Tumor marker measurements
Sun et al. 2018	[41]	SVM + EN	60%	0.64	392 FFDM
Sun et al. 2018	[41]	SVM + RF	66%	0.71	392 FFDM
Sun et al. 2018	[41]	SVM + SFFS	66%	0.69	392 FFDM
Sun et al. 2018	[41]	EnSVM	69%	0.74	392 FFDM
Whitney et al. 2018	[42]	SVM + Low-Low vs. High-High	86%	0.81	178 H&E stained images

Table 2 (continued)

Authors	Paper	Model	Accuracy	AUC	Dataset
Whitney et al. 2018	[42]	SVM + Low vs. High	76%	0.72	178 H&E stained images
Whitney et al. 2018	[42]	SVM + Low vs. Intermediate and High	64%	0.58	178 H&E stained images
Whitney et al. 2018	[42]	SVM + Low and Intermediate vs. High	74%	0.65	178 H&E stained images
Sathipati et al. 2018	[43]	SVM-BRC	80%	0.87	386 miRNA profiles
Guo et al. 2021	[44]	SVM + training cohort	83%	0.83	259 circRNA profiles
Guo et al. 2021	[44]	SVM + testing cohort	79%	0.80	259 circRNA profiles
Guo et al. 2021	[44]	SVM + all cohort	82%	0.82	259 circRNA profiles
Chen et al. 2022	[45]	SVM + CT feature	59%	0.60	217 PET/CT images
Chen et al. 2022	[45]	SVM + PET feature	51%	0.61	217 PET/CT images
Chen et al. 2022	[45]	SVM + PET/CT concat feature	58%	0.66	217 PET/CT images
Chen et al. 2022	[45]	SVM + PET/CT- mean feature	62%	0.49	217 PET/CT images
Lei et al. 2019	[46]	SVM + primary cohort	75%	0.83	212 mammography images
Lei et al. 2019	[46]	SVM + validation cohort	76%	0.78	212 mammography images
Nanglia et al. 2022	[47]	SVM + twofold cross validation	68%	0.74	192 from CBC dataset
Nanglia et al. 2022	[47]	SVM + threefold cross validation	67%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	SVM + fivefold cross validation	67%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	SVM + tenfold cross validation	71%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	SVM + 20-fold cross validation	76%	0.78	192 patients from CBC dataset
Hao et al. 2020	[48]	SVM + T1 + C features	67%	0.71	178 MRI
Hao et al. 2020	[48]	SVM + T2 features	59%	0.69	178 MRI
Hao et al. 2020	[48]	SVM + Fusion features	74%	0.77	178 MRI
Zhu et al. 2021	[49]	SVM + training set	83%	0.91	177 DCE-MRI
Zhu et al. 2021	[49]	SVM + validation set	78%	0.86	177 DCE-MRI
Tomas et al. 2022	[50]	SVM	90%	0.96	166 FFPE breast biopsies
Arefan et al. 2020	[51]	SVM + 2D	77%	0.80	154 MRI

Table 2 (continued)

Authors	Paper	Model	Accuracy	AUC	Dataset
Arefan et al. 2020	[51]	SVM + 3D	76%	0.76	154 MRI
Vamvakas et al. 2022	[52]	SVM	84%	0.88	140 mpMRI
Harrington et al. 2018	[53]	Radial SVM	78%	0.57	128 ADH found on core needle biopsies
Song et al. 2020	[54]	SVM + Precontrast images	68%	0.67	98 DCE-MRI
Song et al. 2020	[54]	SVM + Postcontrast images	71%	0.74	98 DCE-MRI
Song et al. 2020	[54]	SVM + Subtraction images	83%	0.89	98 DCE-MRI
Daimiel Naranjo et al. 2021	[55]	SVM in DWI dataset	73%	0.79	93 MRI
Daimiel Naranjo et al. 2021	[55]	SVM in DCE dataset	77%	0.83	93 MRI
Daimiel Naranjo et al. 2021	[55]	SVM in Multiparametric dataset	82%	0.85	93 MRI
Bhardwaj et al. 2022	[56]	SVM + First and second-order (QUS + QUS-Tex^1)	84%	0.78	83 QUS
Bhardwaj et al. 2022	[56]	SVM + All features (QUS + QUS-Tex^1 + QUS-Tex^1-Tex^2)	85%	0.78	83 QUS
Jiang et al. 2019	[57]	SVM	81%	0.87	73 DCE-MRI
MoghadasDastjerdi et al. 2020	[58]	SVM	75%	0.75	72 qCT
Liu et al. 2019	[59]	SVM + training set	76%	0.82	62 DCE-MRI
Liu et al. 2019	[59]	SVM + validation set	85%	0.83	62 DCE-MRI
Fusco et al. 2021	[60]	SVM + BOLD & DWI extracted features	80%	0.81	59 MRI
Rana et al. 2019	[61]	SVM$_L$	62%	0.23	54 UWB
Rana et al. 2019	[61]	SVM$_Q$	98%	0.94	54 UWB
Tahmassebi et al. 2019	[62]	Linear SVM + RCB	~ 80%	0.85	38 mpMRI
Tahmassebi et al. 2019	[62]	Linear SVM + RFS	~ 73%	0.78	38 mpMRI
Tahmassebi et al. 2019	[62]	Linear SVM + DSS	~ 73%	0.78	38 mpMRI

better than that of paper [61]. The authors attribute this improvement to the utilisation of labelled data in training the SL algorithm and the availability of a larger training dataset. SSL, on the other hand, is a hybrid approach that combines SL and unsupervised learning, utilising unlabeled data and unsupervised information. It achieves competitive results with a smaller amount of data compared to SL methods.

Other papers in the selection combine the standard SVM model with specific features or additional learning models. Among these, paper [50] achieved the highest values for both accuracy and AUC, followed by papers [42, 49, 52]. These studies demonstrate the potential of incorporating additional features or employing hybrid models to enhance the performance of SVM in breast cancer prevention.

The paper [38] achieved the worst result by combining the SVM model with Molecular subtype features. The authors reviewed histopathological reports to identify prognostic biomarkers (such as Lymph node status, tumor grade, ER, PR, HER2, and Ki67) that were strongly associated with molecular subtypes of BC. Despite the inclusion of these features, the performance of the SVM model in this study was low.

Similarly, paper [54] obtained low performance values when using Precontrast and Post-contrast images. Precontrast images are acquired before contrast material injection, while postcontrast images are acquired during the fifth phase after contrast material injection. Subtraction images are obtained by subtracting the Precontrast and Postcontrast images. The SVM model performed better when applied to the subtraction images compared to the other two types of images.

These findings suggest that the inclusion of certain features or imaging modalities does not always lead to improved performance in breast cancer prevention using SVM models. It highlights the importance of carefully selecting relevant features and considering the specific characteristics of the data to optimise the performance of the SVM model.

3.2 Naive Bayes

The Bayesian classifier, as demonstrated in the paper [63], is capable of predicting the probabilities of class membership, which represent the likelihood that a given sample belongs to a specific class. This classifier is built on the foundation of Bayes' theorem, which is expressed by the following formula:

$$P(h|d) = \frac{P(d|h) * P(h)}{P(d)} \tag{2}$$

The Bayesian classifier uses probabilities to estimate the likelihood of a sample belonging to different classes and assigns it to the class with the highest probability. It considers both the prior probability of each class and the likelihood of observing the input given each class, providing a probabilistic framework for classification.

The naive Bayesian classifier makes the assumption of class conditional independence, meaning it assumes that the effect of an attribute value on a class is independent of the values of other attributes. This assumption simplifies the computation and is the reason behind the "naive" characteristic of the classifier.

Table 3 collects all the performance results obtained by the articles using the NB algorithms, sorted by descending cardinality of the dataset.

Table 3 Naïve Bayes results

Authors	Paper	Model	Accuracy	AUC	Dataset
Esmaeili et al. 2020	[64]	Naïve Bayes	81%	0.87	2441 mammography images
Fanizzi et al. 2021	[65]	Naïve Bayes	66%	0.71	907 US
Park et al. 2019	[38]	Naïve Bayes + lymph node	53%	0.58	723 CT
Park et al. 2019	[38]	Naïve Bayes + Tumor grade	63%	0.70	723 CT
Park et al. 2019	[38]	Naïve Bayes + Tumor size	58%	0.63	723 CT
Park et al. 2019	[38]	Naïve Bayes + ER	68%	0.71	723 CT
Park et al. 2019	[38]	Naïve Bayes + PR	65%	0.65	723 CT
Park et al. 2019	[38]	Naïve Bayes + HER2	72%	0.69	723 CT
Park et al. 2019	[38]	Naïve Bayes + Ki67	59%	0.69	723 CT
Park et al. 2019	[38]	Naïve Bayes + Molecular subtype	49%	0.70	723 CT
Al-Azzam et al. 2021	[39]	Gaussian Naïve Bayes + SL	95%	0.94	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	Gaussian Naïve Bayes + SSL	90%	0.89	569 patients from WDBC dataset
Nanglia et al. 2022	[47]	Naïve Bayes + twofold cross validation	65.9%	0.73	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Naïve Bayes + threefold cross validation	68.3%	0.76	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Naïve Bayes + fivefold cross validation	68.29%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Naïve Bayes + tenfold cross validation	69.5%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Naïve Bayes + 20 fold cross validation	69.5%	0.78	192 patients from CBC dataset
Tomas et al. 2022	[50]	Naïve Bayes	64.69%	0.67	166 FFPE breast biopsies
Arefan et al. 2020	[51]	Naïve Bayes + 2D	78%	0.78	154 MRI
Arefan et al. 2020	[51]	Naïve Bayes + 3D	73%	0.75	154 MRI

The paper [39] achieved the best results in terms of NB application, similar to the findings for SVM models. Additionally, the NB model showed different performance values when combined with Supervised Learning (SL) and Semi-Supervised Learning (SSL) techniques. The authors concluded that better performance can be obtained using a fuzzy version of the algorithm instead of the Gaussian one.

On the other hand, the worst performance was observed in the paper [38], where Naive Bayes was combined with Lymph node features and Molecular subtype features.

3.3 Linear and logistic regression

In the paper [66], the performance values of different types of regression models were described. Regression models are used to estimate the impact of independent predictors on a single dependent variable. Specifically, the Linear Regression model assumes a linear relationship between the predicted continuous variable and the predictor variables (Fig. 6). On the other hand, the LR model assumes that the predicted variable represents the logarithmic probability of an event occurrence, based on the predictor variables. The predicted variable in LR is dichotomous, ranging from 0 to 1, representing the probability of the event happening.

As well as in the previous sections, the data presented in Table 4 are sorted by decreasing dataset cardinality.

All of the articles in Table 4 use logistic regression (LR) on their datasets, combining it with various feature selection approaches [62], applying it to training and testing subsets [68], to different types of pictures [54], or to different forms of analysis [69]. The study [65] proposes a variant of the conventional LR that incorporates the LASSO (Least Absolute Shrinkage and Selection Operator) technique, whereas the paper [53] employs a Logistic elastic net to analyse its data.

The paper [39] demonstrated the best performance in terms of the LR model, consistent with the findings for other models in previous tables. In [69], the authors aimed to highlight the importance of early detection of lymphedema in BC survivors. They identified

Fig. 6 Scatter plot showing a linear relation between the two variables, from [66]
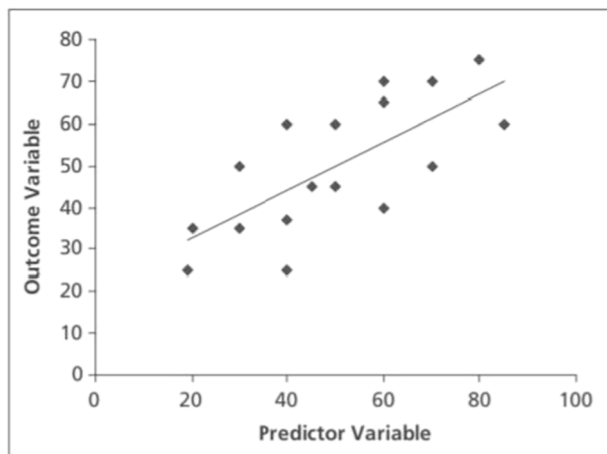
Table 4 Logistic regression results

Authors	Paper	Model	Accuracy	AUC	Dataset
Hou et al. 2020	[67]	LR	58%	0.63	7127 patients from BCIMS
Xu et al. 2019	[68]	LR + training group	74%	0.72	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	LR + testing group	74%	0.72	1570 patients from Sun Yat-sen Memorial Hospital database
Huo et al. 2021	[37]	LR + test set	72%	0.77	1125 ABUS
Huo et al. 2021	[37]	LR + validation set	77%	0.91	1125 ABUS
Fanizzi et al. 2021	[65]	LR	68%	0.72	907 US
Park et al. 2019	[38]	LR + lymph node	62%	0.66	723 CT
Park et al. 2019	[38]	LR + Tumor grade	67%	0.71	723 CT
Park et al. 2019	[38]	LR + Tumor size	64%	0.69	723 CT
Park et al. 2019	[38]	LR + ER	70%	0.73	723 CT
Park et al. 2019	[38]	LR + PR	66%	0.68	723 CT
Park et al. 2019	[38]	LR + HER2	78%	0.69	723 CT
Park et al. 2019	[38]	LR + Ki67	65%	0.70	723 CT
Park et al. 2019	[38]	LR + Molecular subtype	48%	0.69	723 CT
Al-Azzam et al. 2021	[39]	LR + SL	97%	0.96	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	LR + SSL	98%	0.98	569 patients from WDBC dataset
Wei et al. 2021	[69]	LR + Early detection of lymphedema	83%	0.89	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	LR + Late detection of lymphedema	86%	0.93	533 patients from a tertiary hospital in Beijing
Guo et al. 2021	[44]	LR + training cohort	81%	0.80	259 circRNA profiles
Guo et al. 2021	[44]	LR + testing cohort	81%	0.81	259 circRNA profiles
Guo et al. 2021	[44]	LR + all cohort	81%	0.80	259 circRNA profiles
Jiang et al. 2021	[40]	LR	72%	0.79	243 Tumor marker measurements
Nanglia et al. 2022	[47]	LR + twofold cross validation	68%	0.76	192 patients from CBC dataset
Nanglia et al. 2022	[47]	LR + threefold cross validation	69%	0.75	192 patients from CBC dataset
Nanglia et al. 2022	[47]	LR + fivefold cross validation	69%	0.75	192 patients from CBC dataset

Table 4 (continued)

Authors	Paper	Model	Accuracy	AUC	Dataset
Nanglia et al. 2022	[47]	LR + tenfold cross validation	72%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	LR + 20-fold cross validation	71%	0.77	192 patients from CBC dataset
Zhu et al. 2021	[49]	LR + training set	84%	0.92	177 DCE-MRI
Zhu et al. 2021	[49]	LR + validation set	78%	0.84	177 DCE-MRI
Tomas et al. 2022	[50]	LR	65%	0.86	166 FFPE breast biopsies
Perre et al. 2021	[70]	LASSO LR + BI-RADS	63%	0.83	174 MRI
Perre et al. 2021	[65]	LASSO LR + Malignant probability score	78%	0.88	174 MRI
Harrington et al. 2018	[53]	Logistic elastic net	76%	0.57	128 ADH found on core needle biopsies
Harrington et al. 2018	[53]	LR	73%	0.56	128 ADH found on core needle biopsies
Song et al. 2020	[54]	LR + Precontrast images	67%	0.62	98 DCE-MRI
Son et al. 2020	[54]	LR + Postcontrast images	67%	0.73	98 DCE-MRI
Song et al. 2020	[54]	LR + Subtraction images	80%	0.88	98 DCE-MRI
Liu et al. 2019	[59]	LR + training set	71%	0.82	62 DCE-MRI
Liu et al. 2019	[59]	LR + validation set	77%	0.88	62 DCE-MRI
Jiang et al. 2019	[57]	LR	81%	0.85	73 DCE-MRI
Tahmassebi et al. 2019	[62]	LR + RCB	~ 79%	0.82	38 mpMRI
Tahmassebi et al. 2019	[62]	LR + RFS	~ 79%	0.83	38 mpMRI
Tahmassebi et al. 2019	[62]	LR + DSS	~ 75%	0.81	38 mpMRI

24 lymphedema-associated symptoms as potential predictors and found that Logistic Regression achieved the best performance for early detection. Conversely, in [49], all studied models showed no significant differences in performance. In this study, features were extracted using LASSO from DCE-MRI images.

In [57], excellent results were achieved by utilising 279 textual features for each case. These features were analysed using the MaZda software, publicly accessible through [71, 72]. To reduce the complexity of subsequent ML analysis, a feature selection analysis was performed using SPSS, resulting in the reduction of weak features.

Interestingly, in [54], the LR model performed well when applied to subtraction images, but yielded poorer results when applied to Precontrast and Postcontrast images.

In [67], the LR model was applied to data collected through a questionnaire to investigate the impact of demographic and other risk factors on BC onset. The authors selected a total of 10 variables, including 3 demographic factors, 6 reproductive history factors, and family history of BC. However, the LR model yielded poor performance results in this study.

Similarly, in [38], when the LR model was combined with the molecular subtype features of BC, low performance results were obtained.

3.4 K-nearest neighbour

In the paper [73], the K-Nearest Neighbors (K-NN) algorithm is described as one of the most fundamental and straightforward classification methods. It involves associating new data with the most common class among its k nearest neighbours, where k is a predefined parameter that influences the final result. The accuracy of the K-NN classifier is influenced by both the choice of k and the distance metric used to compute distances between data points. Different distance measures can yield varying levels of accuracy depending on the presence of noise in the data, as discussed in [74].

In Fig. 7, from paper [74], a visual comparison of 10 distance measures is presented, including the Average (L_1, L_{inf}) distance, Canberra distance, Clark distance, Correlation

Fig. 7 Average Accuracy of KNN classifier using top 10 distance measures with different level of noise from paper [74]. AvgD, Average (L_1, L_{inf}) distance; CanD, Canberra distance; ClaD, Clark distance; CorD, Correlation distance; CosD, Cosine distance; DicD, Dice distance; DivD, Divergence distance; LD, Lorentzian distance; MD, Manhattan distance; SCSD, Squared Chi-Squared; WIAD, Whittaker's index of association distance
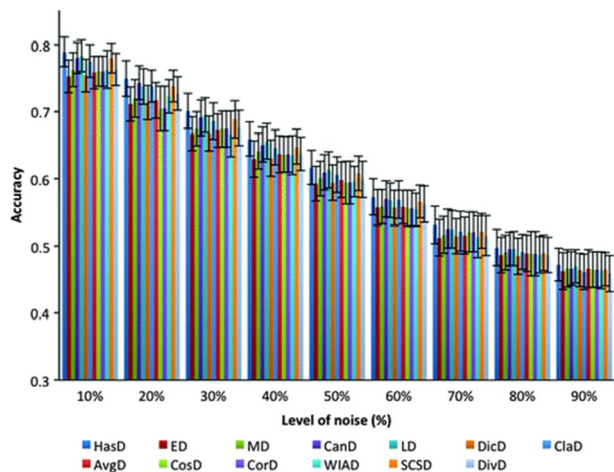
distance, Cosine distance, Dice distance, Divergence distance, Lorentzian distance, Manhattan distance, Squared Chi-Squared distance, and Whittaker's index of association distance. These distance measures are used to compute the distances between data points in the K-Nearest Neighbors (K-NN) algorithm.

In Table 5, all the papers listed combine the standard K-NN model with a set of features. However, the paper [61] is excluded from the table as its results for the K-NN model are not comparable with the other papers.

The paper [53] proposes a different version of the K-NN model called Weighted KNN. In this version, the distance of the nearest neighbours is incorporated, and the observations of the nearest neighbours are upweighted compared to those of more distant neighbours. The authors conclude that this weighted version improves the classifier's performance [75].

In the context of supervised and semi-supervised learning, the paper [39] achieved the best results among the studies examined. On the other hand, the paper [47] obtained low results when combining the K-NN model with twofold and threefold cross-validation. This can be attributed to the fact that with small training sets, increasing the number of folds in cross-validation helps reduce bias in generalisation error estimation by utilising more training data in each iteration.

3.5 Decision tree

According to the paper [76], a DT is a formal representation of classification flow within a given set of instances. In a DT, each leaf node represents one of the possible classes, and the intermediate nodes correspond to the tests performed on the data. Each branch originating from a node represents one of the possible outcomes of the test conducted at that node.

Table 6 contains papers that utilise a DT model, while Table 7 includes papers that focus on other tree-based classification approaches.

The paper [39] stands out as one of the best-performing articles in Table 6. On the other hand, the results of the paper [49] cannot be considered the best due to overfitting, even though it achieved 100% accuracy on the training set. Therefore, the paper [39] still holds the best results.

Additionally, the paper [38] has already been mentioned in previous sections for its poor performance when combining the DT model with molecular subtype features. In the case of the paper [62], the authors extracted 23 features from the dataset for each lesion but only used three features as reference standards: histopathologic Residual Cancer Burden (RCB) class, Recurrence-Free Survival (RFS), and Disease-Specific Survival (DSS). The DT model combined with RFS yielded low results. Table 6 presents the median of the four-fold cross-validation results.

While Table 6 compiles all publications that show DTs as a model, Table 7 compiles all research that presents another form of tree. These models can be utilised in various forms of analysis, as in article [69], or can be used for training and testing datasets, as in papers [37, 77, 79]. Paper [77] includes a Linear Discriminant Analysis (LDA) step to maximise the ratio of between-class variation to within-class variance in the dataset, assuring maximum separability. The LDA model will be presented in more detail in Section 3.6.

Study [49] experienced overfitting when the model was applied to the training set, and therefore, its result cannot be considered the best in Table 7.

Table 5 K-nearest neighbour results

Authors	Paper	Model	Accuracy	AUC	Dataset
Esmaeili et al. 2020	[53]	K-NN	84%	0.73	2441 mammography images
Al-Azzam et al. 2021	[39]	K-NN+SL	98%	0.98	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	K-NN+SSL	97%	0.97	569 patients from WDBC dataset
Nanglia et al. 2022	[47]	K-NN+twofold cross validation	77%	0.80	192 patients from CBC dataset
Nanglia et al. 2022	[47]	K-NN+threefold cross validation	69%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	K-NN+fivefold cross validation	69%	0.77	192 patients from CBC dataset
Nanglia et al. 2022	[47]	K-NN+tenfold cross validation	74%	0.78	192 patients from CBC dataset
Nanglia et al. 2022	[47]	K-NN+20-fold cross validation	72%	0.76	192 patients from CBC dataset
Arefan et al. 2020	[51]	K-NN+2D	77%	0.77	154 MRI
Arefan et al. 2020	[51]	K-NN+3D	75%	0.75	154 MRI
Harrington et al. 2018	[53]	Weighted K-NN	74%	0.63	128 ADH found on core needle biopsies
Bhardwaj et al. 2022	[56]	K-NN+First and second-order (QUS+QUS-Tex^1)	74%	0.70	83 QUS
Bhardwaj et al. 2022	[56]	K-NN+All features (QUS+QUS-Tex^1+QUS-Tex^1-Tex^2)	81%	0.83	83 QUS
Fusco et al. 2021	[60]	K-NN+BOLD extracted features	73%	0.81	59 MRI

Table 6 Decision tree results

Authors	Paper	Model	Accuracy	AUC	Dataset
Xu et al. 2019	[68]	DT + training group	73%	0.70	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	DT + testing group	75%	0.69	1570 patients from Sun Yat-sen Memorial Hospital database
Al-Azzam et al. 2021	[39]	DT + SL	91%	0.89	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	DT + SSL	93%	0.9	569 patients from WDBC dataset
Jiang et al. 2021	[40]	DT	65%	0.70	243 Tumor marker measurements
Park et al. 2019	[38]	Decision + Tree lymph node	66%	0.65	723 CT
Park et al. 2019	[38]	DT + Tumor grade	71%	0.69	723 CT
Park et al. 2019	[38]	DT + Tumor size	65%	0.65	723 CT
Park et al. 2019	[38]	DT + ER	72%	0.69	723 CT
Park et al. 2019	[38]	DT + PR	69%	0.67	723 CT
Park et al. 2019	[38]	DT + HER2	77%	0.67	723 CT
Park et al. 2019	[38]	DT + Ki67	66%	0.66	723 CT
Park et al. 2019	[38]	DT + Molecular subtype	50%	0.63	723 CT
Nanglia et al. 2021	[47]	DT + twofold cross validation	73%	0.75	192 patients from CBC dataset
Nanglia et al. 2021	[47]	DT + threefold cross validation	66%	0.67	192 patients from CBC dataset
Nanglia et al. 2021	[47]	DT + fivefold cross validation	66%	0.67	192 patients from CBC dataset
Nanglia et al. 2021	[47]	DT + tenfold cross validation	73%	0.74	192 patients from CBC dataset
Nanglia et al. 2021	[47]	DT + 20-fold cross validation	76%	0.78	192 patients from CBC dataset
Zhu et al. 2021	[49]	Decision + Tree training set	100%	1.00	177 DCE-MRI
Zhu et al. 2021	[49]	Decision + Tree validation set	74%	0.74	177 DCE-MRI
Tomas et al. 2022	[50]	DT	69%	0.69	166 FFPE breast biopsies
Tahmassebi et al. 2019	[62]	DT + RCB	~ 73%	0.77	38 mpMRI
Tahmassebi. et al. 2019	[62]	DT + RFS	~ 52%	0.57	38 mpMRI
Tahmassebi et al. 2019	[62]	DT + DSS	~ 75%	0.80	38 mpMRI

In the study [69], the authors aimed to determine the most effective model for stratifying the risk of breast cancer survivors and excluding potential patients with lymphedema. When the two proposed models were applied to the late detection of lymphedema, they achieved some of the best results, as shown in Table 7.

Additionally, it is noted that the paper [49] obtained the worst result when the Gradient Boosting Decision Tree model was applied to the testing set. This confirms that the

Table 7 Other tree results

Authors	Paper	Model	Accuracy	AUC	Dataset
Hou et al. 2021	[37]	Extra tree + test set	72%	0.75	1125 ABUS
Hou et al. 2021	[37]	Extra tree + validation set	81%	0.86	1125 ABUS
Wei et al. 2021	[69]	Classification and Regression Tree + Early detection of lymphedema	82%	0.85	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	Classification and Regression Tree + Late detection of lymphedema	87%	0.87	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	C5.0 DT + Early detection of lymphedema	82%	0.88	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	C5.0 DT + Late detection of lymphedema	86%	0.93	533 patients from a tertiary hospital in Beijing
Zhao et al. 2021	[77]	Gradient Boosting DT + LDA + Training set	87%	0.92	265 MRI
Zhao et al. 2021	[77]	Gradient Boosting DT + LDA + Testing set	87%	0.91	265 MRI
Zhu et al. 2021	[49]	Gradient Boosting DT + training set	100%	1.00	177 DCE-MRI
Zhu et al. 2021	[49]	Gradient Boosting DT + validation set	72%	0.82	177 DCE-MRI
Harrington et al. 2018	[53]	Gradient Boosting tree	78%	0.68	128 ADH found on core needle biopsies

high result achieved when applying the same model to the training set was solely due to overfitting.

3.6 Discriminant analysis

Linear Discriminant Analysis (LDA) [78] is a method that handles cases where within-class frequencies are uneven. It has been tested using randomly generated test data, and its performance has been evaluated. The goal of LDA is to maximise the ratio of between-class variance to within-class variance in a given dataset, ensuring maximal separability.

The use of Linear Discriminant Analysis for data classification is particularly applied to classification problems in speech recognition. In comparison to Principal Component Analysis (PCA) [79], literature suggests that LDA performs better. PCA is another dimensionality reduction technique commonly used in machine learning, but LDA has shown to provide improved results in terms of separability and classification accuracy.

Among the papers listed in Table 8, the paper [60] achieved the best results. The study aimed to assess the differentiation of benign and malignant breast lesions using Blood Oxygenation Level Dependent Magnetic Resonance Imaging (BOLD-MRI) and Diffusion Weighted Magnetic Resonance Imaging (DW-MRI). The combination of LDA with robust BOLD and DWI features, extracted using the Least Absolute Shrinkage and Selection Operator (LASSO), yielded the best performance.

Another paper, [51], also obtained high results. The study compared the effects of radiomic analysis on 2D and 3D tumor segmentation using different machine learning (ML) models. An independent testing technique was employed, where a training set of 103 patients and a testing set of 51 patients were used. The LASSO method, followed by a ten-fold cross-validation, was used to select the best subset of features based on mean square error. The performance of the models was evaluated using these selected features, comparing their distributions in 2D and 3D analysis.

On the other hand, the papers [50, 62] achieved the worst outcomes in Table 8. In the first paper, the authors mainly focused on comparing their results with artificial neural network (ANN) results and did not extensively investigate the reasons behind the performance of other machine learning models. The second paper, [50], extracted 23 features for each lesion but only utilised three of them: histopathologic Residual Cancer Burden (RCB) class, Recurrence-Free Survival (RFS), and Disease-Specific Survival (DSS). The results presented in Table 8 represent the median of a fourfold cross-validation.

3.7 Artificial neural network

As defined in the paper [81], Artificial Neural Networks (ANN) are an intelligent system inspired by biological neural networks.

ANN are characterised by the activation function used by their artificial neurons (see Fig. 8) and by the links between artificial neurons in different layers of the networks, as presented in Fig. 9.

In Table 9, the paper [69] achieved the best results, in addition to the previously discussed paper [39]. The study described in [69] aimed to find an effective approach to stratify the risk of BC survivors while excluding potential lymphedema patients. The DT model was used to achieve this goal, and it yielded impressive results.

Table 8 Discriminant analysis results

Authors	Paper	Model	Accuracy	AUC	Dataset
Guo et al. 2021	[44]	LDA + training cohort	78%	0.77	259 circRNA profiles
Guo et al. 2021	[44]	LDA + testing cohort	79%	0.80	259 circRNA profiles
Guo et al. 2021	[44]	LDA + all cohort	78%	0.78	259 circRNA profiles
Tomas et al. 2022	[50]	LDA	66%	0.72	166 FFPE breast biopsies
Arefan et al. 2020	[51]	LDA + 2D	79%	0.81	154 MRI
Arefan et al. 2020	[51]	LDA + 3D	80%	0.82	154 MRI
Braman et al. 2017	[80]	Diagonal LDA	76%	0.78	117 DCE-MRI
Song et al. 2020	[54]	Quadratic Discriminant analysis + Precontrast images	66%	0.57	98 DCE-MRI
Song et al. 2020	[54]	Quadratic Discriminant analysis + Postcontrast images	69%	0.73	98 DCE-MRI
Song et al. 2020	[54]	Quadratic Discriminant analysis + Subtraction images	73%	0.83	98 DCE-MRI
Jiang et al. 2019	[57]	Quadratic Discriminant Analysis	73%	0.81	73 DCE-MRI
Fusco et al. 2021	[60]	LDA + DWI extracted features	71%	0.74	59 MRI
Fusco et al. 2021	[60]	LDA + All robust (by LASSO) BOLD and DWI extracted features	83%	0.83	59 MRI
Tahmassebi et al. 2019	[62]	LDA + RCB	~ 61%	0.66	38 mpMRI
Tahmassebi et al. 2019	[62]	LDA + RFS	~ 71%	0.76	38 mpMRI
Tahmassebi et al. 2019	[62]	LDA + DSS	~ 76%	0.90	38 mpMRI
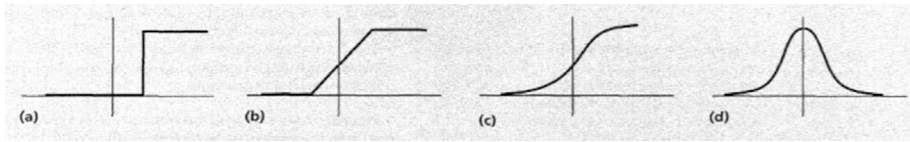
Fig. 8 Different types of activation functions from [81]: **a** Threshold, **b** Piecewise linear, **c** Singmoid and (**d**) Gaussian
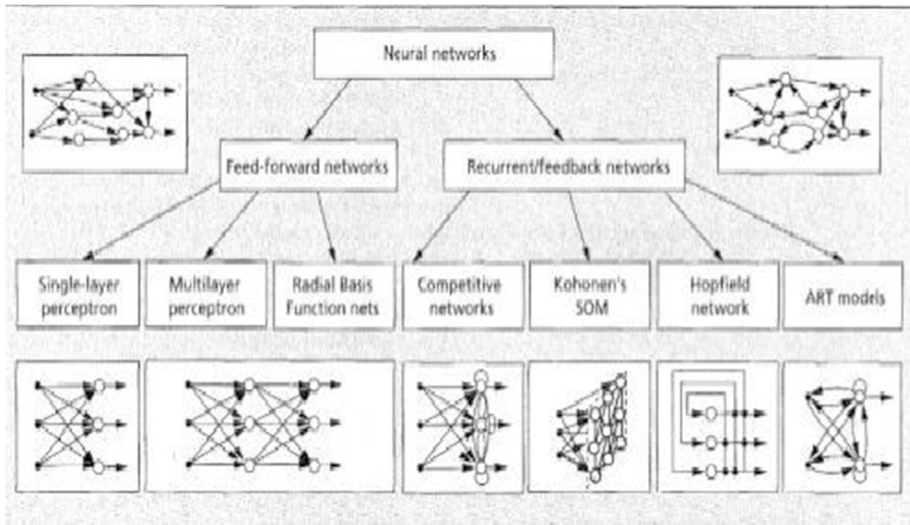
Fig. 9 Different types of artificial neural networks from [81]

On the other hand, the paper [38] obtained low results when applying the Artificial Neural Network (ANN) model to the molecular subtype features. This suggests that the chosen model was not suitable for the molecular subtype features, but it performed well with the other types of features.

It's important to note that the performance of a model can vary depending on the specific features and characteristics of the dataset being analysed. The success or failure of a model in one context does not necessarily guarantee the same outcome in a different context or with different features.

Table 10 collects results from papers that describe other types of Neural Networks (NN), like MultiLayer Perceptron Neural Network (MLP—NN) or Feed Forward Neural Network (FNN). The studies in Table 10 use the MLP—NN to specific feature sets, PET (Positron Emission Tomography) features or CT (Computed Tomography) features, or to certain types of images, DES (Dual-energy subtracted) or LE (Low-Energy), with associated segmentation.

In Table 10, the paper [50] achieved the best results. This study focused on designing and testing different types of Feed Forward Neural Networks (FNN) with varying layer

Table 9 Artificial neural network results

Authors	Paper	Model	Accuracy	AUC	Dataset
Park et al. 2019	[38]	ANN + lymph node	64%	0.68	723 CT
Park et al. 2019	[38]	ANN + Tumor grade	68%	0.73	723 CT
Park et al. 2019	[38]	ANN + Tumor size	65%	0.71	723 CT
Park et al. 2019	[38]	ANN + ER	75%	0.77	723 CT
Park et al. 2019	[38]	ANN + PR	69%	0.72	723 CT
Park et al. 2019	[38]	ANN + HER2	76%	0.73	723 CT
Park et al. 2019	[38]	ANN + Ki67	66%	0.71	723 CT
Park et al. 2019	[38]	ANN + Molecular subtype	35%	0.72	723 CT
Wei et al. 2021	[69]	ANN + Early detection of lymphedema	84%	0.90	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	ANN + Late detection of lymphedema	86%	0.93	533 patients from a tertiary hospital in Beijing
Al-Azzam et al. 2021	[39]	ANN + SL	97%	0.94	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	ANN + SSL	97%	0.89	569 patients from WDBC dataset
Sun et al. 2018	[41]	ANN + EN	60%	0.64	392 FFDM
Sun et al. 2018	[41]	ANN + RF	65%	0.71	392 FFDM
Sun et al. 2018	[41]	ANN + SFSS	63%	0.68	392 FFDM
Nanglia et al. 2021	[47]	ANN + twofold cross validation	73%	0.78	192 patients from CBC dataset
Nanglia et al. 2021	[47]	ANN + threefold cross validation	72%	0.79	192 patients from CBC dataset
Nanglia et al. 2021	[47]	ANN + fivefold cross validation	72%	0.79	192 patients from CBC dataset
Nanglia et al. 2021	[47]	ANN + tenfold cross validation	71%	0.80	192 patients from CBC dataset
Nanglia et al. 2021	[47]	ANN + 20-fold cross validation	71%	0.80	192 patients from CBC dataset

Table 10 Others neural network results

Authors	Paper	Model	Accuracy	AUC	Dataset
Chen et al. 2022	[45]	MLP—NN + CT feature	63%	0.61	217 PET/CT images
Chen et al. 2022	[45]	MLP—NN + PET feature	58%	0.56	217 PET/CT images
Chen et al. 2022	[45]	MLP—NN + PET/CT-concat feature	67%	0.67	217 PET/CT images
Chen et al. 2022	[45]	MLP—NN + PET/CT-mean feature	65%	0.66	217 PET/CT images
Tomas et al. 2022	[50]	FNN—2	96%	0.90	166 FFPE breast biopsies
Tomas et al. 2022	[50]	FNN—4	95%	0.90	166 FFPE breast biopsies
Tomas et al. 2022	[50]	FNN—8	95%	0.90	166 FFPE breast biopsies
Danala et al. 2018	[82]	MLP Neural Network + DES images + Original segmentation	68%	0.77	111 CEDM
Danala et al. 2018	[82]	MLP—NN + LE images + Original segmentation	72%	0.75	111 CEDM
Danala et al. 2018	[82]	MLP—NN + DES images + Optimally mapped segmentation	68%	0.74	111 CEDM
Danala et al. 2018	[82]	MLP—NN + LE images + Optimally mapped segmentation	78%	0.85	111 CEDM
Song et al. 2020	[54]	MLP—NN	73%	0.74	98 DCE-MRI

sizes. The authors found that the optimised learning rate for each model was 0.01, and they determined the specific number of nodes per layer to maximise performance. The best-performing models were FNN2 with 350 nodes, FNN4 with 400 nodes, and FNN8 with 300 nodes.

On the other hand, the paper [45] obtained the worst result among the papers in Table 10. This study applied Artificial Neural Networks (ANN) to PET features. The PET and CT images were processed using different methods, including Exponential, Gradient, Laplacian of Gaussian (LoG), Logarithm, Square, Square root, and Wavelet filtering. The PET/CTconcat features represented an integration of all the PET and CT radiomic features, while the PET/CTmean radiomic feature was the average value of the individual CT and PET radiomic features.

The performance of a model can be influenced by various factors, such as the choice of features, data preprocessing techniques, and model configuration. The best and worst results in Table 10 reflect the outcomes specific to the approaches taken in each respective paper.

3.8 Random forest

The RF algorithm combines several decision trees and aggregates their prediction by averaging [83]. Some authors think RF aggregates random decision trees without considering how the trees are obtained. Other authors instead claim that RF refers to Breiman's [84] original algorithm.

In Table 11, it is important to consider the issue of overfitting and not solely rely on accuracy results obtained on the training set. The paper [49] achieving 100% accuracy on the training set but lower accuracy on the validation set indicates a potential case of overfitting, where the model has memorised the training data instead of learning general patterns.

As mentioned, the paper [39] consistently achieved the best results among the studies reported in Table 11. This indicates the effectiveness of the approach presented in that paper across multiple evaluation metrics or datasets.

The paper [77] also achieved excellent results by manually segmenting scanner images and extracting texture features for classification. This indicates the importance of careful pre-processing and feature extraction techniques in achieving good performance.

On the other hand, the paper [62, 67] obtained poor results in Table 11, which were also mentioned in previous sections. It suggests that the chosen models or feature representations might not have been suitable for the respective datasets or for classification tasks.

In summary, it is crucial to consider the impact of overfitting and the generalizability of results when evaluating the performance of classification models. The best results are often achieved by papers that effectively address these considerations and demonstrate good performance on validation or independent test sets.

3.9 Boosting

Boosting is a particular ML approach based on a combination of a highly accurate rule with other weaker or less accurate rules.

Table 11 Random forest results

Authors	Paper	Model	Accuracy	AUC	Dataset
Hou et al. 2020	[67]	RF	66%	0.72	7127 patients from BCIMS
Esmaeili et al. 2020	[64]	RF	82%	0.90	2441 mammography images
Xu et al. 2019	[68]	RF + training group	77%	0.90	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	RF + testing group	74%	0.71	1570 patients from Sun Yat-sen Memorial Hospital database
Huo et al. 2021	[37]	RF + test set	73%	0.75	1125 ABUS
Huo et al. 2021	[37]	RF + validation set	81%	0.86	1125 ABUS
Fanizzi et al. 2021	[65]	RF	68%	0.68	907 US
Park et al. 2019	[38]	RF + lymph node	78%	0.86	723 CT
Park et al. 2019	[38]	RF + Tumor grade	80%	0.88	723 CT
Park et al. 2019	[38]	RF + Tumor size	77%	0.85	723 CT
Park et al. 2019	[38]	RF + ER	82%	0.88	723 CT
Park et al. 2019	[38]	RF + PR	78%	0.85	723 CT
Park et al. 2019	[38]	RF + HER2	83%	0.88	723 CT
Park et al. 2019	[38]	RF + Ki67	77%	0.85	723 CT
Park et al. 2019	[38]	RF + Molecular subtype	66%	0.82	723 CT
Al-Azzam et al. 2021	[39]	RF + SL	96%	0.96	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	RF + SSL	96%	0.96	569 patients from WDBC dataset
Wei et al. 2021	[69]	RF + Early detection of lymphedema	82%	0.87	533 patients from a tertiary hospital in Beijing
Wei et al. 2021	[69]	RF + Late detection of lymphedema	87%	0.89	533 patients from a tertiary hospital in Beijing
Zhao et al. 2021	[77]	RF + LDA + Training set	84%	0.91	265 MRI
Zhao et al. 2021	[77]	RF + LDA + Testing set	82%	0.88	265 MRI
Jiang et al. 2021	[49]	RF	72%	0.77	243 Tumor marker measurements
Chen et al. 2022	[45]	RF + CT feature	68%	0.61	217 PET/CT images
Chen et al. 2022	[45]	RF + PET feature	65%	0.60	217 PET/CT images
Chen et al. 2022	[45]	RF + PET/CT concat feature	71%	0.69	217 PET/CT images

Table 11 (continued)

Authors	Paper	Model	Accuracy	AUC	Dataset
Chen et al. 2022	[45]	RF + PET/CT mean feature	66%	0.68	217 PET/CT images
Romeo et al. 2021	[85]	RF + training set	82%	0.90	201 US
Romeo et al. 2021	[85]	RF + test set	82%	0.82	201 US
Nanglia et al. 2022	[47]	Ensemble of RF + twofold cross validation	74%	0.82	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Ensemble of RF + threefold cross validation	66%	0.73	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Ensemble of RF + fivefold cross validation	66%	0.74	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Ensemble of RF + tenfold cross validation	73%	0.76	192 patients from CBC dataset
Nanglia et al. 2022	[47]	Ensemble of RF + 20-fold cross validation	68%	0.78	192 patients from CBC dataset
Zhu et al. 2021	[49]	RF + training set	100%	1.00	177 DCE-MRI
Zhu et al. 2021	[49]	RF + validation set	81%	0.85	177 DCE-MRI
Tomas et al. 2022	[50]	RF	77%	0.85	166 FFPE breast biopsies
Arefan et al. 2020	[51]	RF + 2D	79%	0.78	154 MRI
Arefan et al. 2020	[51]	RF + 3D	80%	0.80	154 MRI
Harrington et al. 2018	[53]	RF	77%	0.67	128 ADH found on core needle biopsies
Tahmassebi et al. 2019	[62]	RF + RCB	~ 75%	0.84	38 mpMRI
Tahmassebi et al. 2019	[62]	RF + RFS	~ 62%	0.67	38 mpMRI
Tahmassebi et al. 2019	[62]	RF + DSS	~ 76%	0.86	38 mpMRI

As explained in [86], the Adaptive Boosting (AdaBoost) is the first practical boosting algorithm and it is still one of the most used and studied. AdaBoost pseudocode is presented in Algorithm 1.

Algorithm 1 AdaBoost pseudocode from paper [86]

Given: m labelled training examples $(x_1,y_1),...,(x_m,y_m)$ where $x_i \in X_i$
  and the labels $y_i \in \{-1,+1\}$
Given: ϵ - Stopping criterion threshold

1: Initialise distribution

$$D_1(i) = \frac{1}{m} \; for \; i = 1,...,m$$

(3)

2: For each round $t = 1,...,T$:
3: Train weak learner using distribution D_t
4: Get weak hypothesis

$$h_t : X \leftarrow \{-1,+1\}$$

(4)

5: Aim: select h_t with low weighted error:

$$\epsilon_t = P_{r_i \sim D_i}[h_t(x_i) \neq y_i$$

(5)

6: Choose

$$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$$

(6)

7: Update, for

$$i = 1,...,m: D_{t+1}(i) = \frac{D_t(i) - \alpha_i y_i h_i(x_i)}{Z_t}$$

(7)

  where Z_t is a normalisation factor (chosen so that D_{t+1} will be a distribution).
8: Output of the final hypothesis:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

(8)

XGBoost is short for eXtreme Gradient Boosting package. As described in the paper [87], XGBoost is an efficient and scalable implementation of a Gradient Boosting Machine, that is defined in [88].

For each round $t = 1,...,T$, a distribution D_t is computed over the m training instances, and a particular weak learner or weak learning algorithm is performed to find a weak hypothesis, as defined in Eq. 4. The weak learner's goal is to find a weak hypothesis with a low weighted error ϵ_t relative to D_t. Equation 8 is the final or combined hypothesis $H(x)$. $H(x)$ is calculated as a weighted majority vote of the weak hypothesis h_t, with weight α_t assigned to each hypothesis.

The latter paper presents a general gradient descent "boosting" model developed for any of the fitting criteria. Specific algorithms for Least Squares, Least Absolute

Table 12 AdaBoost results

Authors	Paper	Model	Accuracy	AUC	Dataset
Vamvakas et al. 2022	[52]	Adaboost	83%	0.90	140 mpMRI
MoghadasDastjerdi et al. 2020	[58]	AdaBoost—SVM	78%	0.78	72 qCT
MoghadasDastjerdi et al. 2020	[58]	AdaBoost—DT	84%	0.84	72 qCT
Tahmassebi et al. 2019	[62]	AdaBoost + RCB	~ 77%	0.81	38 mpMRI
Tahmassebi et al. 2019	[62]	AdaBoost + RFS	~ 65%	0.70	38 mpMRI
Tahmassebi et al. 2019	[62]	AdaBoost + DSS	~ 79%	0.83	38 mpMRI

Deviation, Huber-M loss functions for regression, and multiclass logistic likelihood for classification are also described.

To describe the Boosting algorithms' results, papers are divided in three Tables 12, 13, and 14.

Table 12 focuses on papers that describe the results of AdaBoost [80] in various machine learning tasks. Among the papers listed, the best result was achieved by the paper [58], which applied AdaBoost with two different classifiers: SVM and DT. The AdaBoost-DT classifier performed the best among the two, surpassing the performance of the hybrid classifier proposed in the same paper, which combined SVM, RF, and DT classifiers. On the other hand, the worst result in Table 12 was obtained by the paper [62], which was already mentioned in the previous sections.

AdaBoost is known for its ability to boost the performance of weak classifiers by focusing on misclassified instances and iteratively updating the weights of the training data. It has been successfully applied to various classification and regression problems, and the papers in Table 12 provide insights into its effectiveness when combined with different base classifiers and applied to specific datasets.

Table 13 focuses on papers that present results of Gradient Boosting algorithms. Among the papers listed, the best result was achieved by the paper [39], which has already been discussed extensively in previous sections. Another paper, [52], introduced a modified version of the standard Gradient Boosting algorithm [89] called LightGBM, which showed excellent results. LightGBM is considered a new addition to the collection of boosting models and is known for its efficiency and performance advantages over XGBoost in certain aspects. The principles of LightGBM, Gradient Boosting Decision Trees (GBDT), and XGBoost are similar, with all three methods utilising the negative gradient of the loss function to approximate the residuals and fit decision trees.

In the paper [52], radiomic feature extraction was performed on the original parametric images without any filtering, and features related to shape, grey-level, and grey-tone were calculated. A total of 293 features for each subject's imaging set were extracted in this study.

On the other hand, the worst result in Table 13 was obtained by the paper [40], where only eight features were selected because they were detected in all patients in the dataset. Among these features, five markers were identified for differentiating between breast cancers and benign tumors.

Table 14 presents the results of papers that utilised the XGBoost model. Apart from the paper [39], the best results among the papers in the table were achieved by the papers [52, 59]. In the paper [59], a large number of quantitative imaging features (1,409 features) were automatically extracted from each VOI (Volume of Interest).

Table 13 Boosting results

Authors	Paper	Model	Accuracy	AUC	Dataset
Xu et al. 2019	[68]	Gradient Boosting + training group	75%	0.78	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	Gradient Boosting + testing group	73%	0.73	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	LightGBM + training group	76%	0.89	1570 patients from Sun Yat-sen Memorial Hospital database
Xu et al. 2019	[68]	LightGBM + testing group	73%	0.71	1570 patients from Sun Yat-sen Memorial Hospital database
Al-Azzam et al. 2021	[39]	Gradient Boosting + SL	94%	0.98	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	Gradient Boosting + SSL	94%	0.92	569 patients from WDBC dataset
Jiang et al. 2021	[40]	Gradient Boosting machine	68%	0.68	243 Tumor marker measurements
Vamvakas et al. 2022	[52]	Light Gradient Boosting Machine (LGBM)	87%	0.94	140 mpMRI
Vamvakas et al. 2022	[52]	Gradient Boosting	83%	0.89	140 mpMRI

Table 14 XGBoost results

Authors	Paper	Model	Accuracy	AUC	Dataset
Hou et al. 2020	[67]	XGBoost	67%	0.74	7127 patients from BCIMS
Hou et al. 2020	[37]	XGBoost + test set	71%	0.77	1125 ABUS
Hou et al. 2020	[37]	XGBoost + validation set	78%	0.90	1125 ABUS
Al-Azzam et al. 2021	[39]	XGBoost + SL	97%	0.97	569 patients from WDBC dataset
Al-Azzam, N. et al., 2021	[39]	XGBoost + SSL	93%	0.91	569 patients from WDBC dataset
Zhao et al. 2021	[77]	XGBoost + LDA + Training set	79%	0.90	265 MRI
Zhao et al. 2021	[77]	XGBoost + LDA + Testing set	81%	0.90	265 MRI
Chen et al. 2022	[45]	XGBoost + CT feature	64%	0.64	217 PET/CT images
Chen et al. 2022	[45]	XGBoost + PET feature	66%	0.62	217 PET/CT images
Chen et al. 2022	[45]	XGBoost + PET/CT -concat feature	72%	0.72	217 PET/CT images
Chen et al. 2022	[45]	XGBoost + PET/CT -mean feature	74%	0.76	217 PET/CT images
Vamvakas et al. 2022	[52]	XGBoost	88%	0.95	140 mpMRI
Liu et al. 2019	[59]	XGBoost + training set	84%	0.92	62 DCE-MRI
Liu et al. 2019	[59]	XGBoost + validation set	85%	0.83	62 DCE-MRI
Tahmassebi et al. 2019	[62]	XGBoost + RCB	~ 81%	0.86	38 mpMRI
Tahmassebi et al. 2019	[62]	XGboost + RFS	~ 73%	0.77	38 mpMRI
Tahmassebi et al. 2019	[62]	XGBoost + DSS	~ 73%	0.90	38 mpMRI

These features were then categorised into four groups. Data preprocessing and a tenfold cross-validation were performed, followed by the use of the LASSO function to select relevant features. Radiomic features with non-zero coefficients were identified as the final set of features.

On the other hand, the paper [45] obtained the worst results when the XGBoost model was applied to CT and PET features. In this study, a semi-automatic segmentation algorithm was used to determine the VOI of the three-dimensional gross tumor volume (GTV), and manual adjustment was performed for accuracy. Subsequently, PET and CT images underwent various processing methods, including Exponential, Gradient, Laplacian of Gaussian (LoG), Logarithm, Square, Square root, and Wavelet filtering. A total of 1,218 CT and 1,218 PET radiomic features were extracted from the segmented tumor region of each patient.

3.10 RBF network

The Radial Basis Function (RBF) Networks, as described in the paper [90], are a type of machine learning model commonly used for prediction and forecasting tasks. The structure of an RBF Network typically consists of three layers: an input layer, a hidden layer, and an output layer, as shown in Fig. 10.

Unlike traditional neural networks with multiple intermediate layers, RBF Networks have only a single hidden layer. However, they are still capable of solving complex problems. This is achieved through the use of a Gaussian function applied in the hidden layer, which allows the network to transform nonlinear inputs into linear outputs. The hidden layer computes the nonlinear output based on the input, utilising the Gaussian function centred around specific radial basis functions.

The linear output of the RBF Network is obtained by summing the weighted nonlinear outputs from the hidden layer. This combination of nonlinear transformations and linear aggregation enables RBF Networks to effectively capture and model complex relationships in the data.

The description of paper [39], the only one in Table 15, has been thoroughly explored in the preceding sections, being often one of the best results in the tables previously discussed.

In order to understand why the RBF+SSL model outperforms the RBF+SL model in paper [39], one first needs to outline the dataset on which the two models are applied. The

Fig. 10 RBF Network structure, from [90]
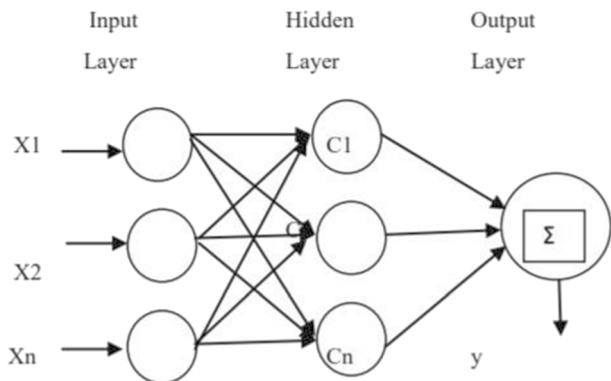
Table 15 RBF network results

Authors	Paper	Model	Accuracy	AUC	Dataset
Al-Azzam et al. 2021	[39]	RBF SVM + SL	96%	0.96	569 patients from WDBC dataset
Al-Azzam et al. 2021	[39]	RBF SVM + SSL	97%	0.96	569 patients from WDBC dataset

Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains 569 samples, 357 of which are benign and 212 of which are malignant. Although it is not a small dataset, it is also not excessively enormous. The RBF+SSL model can improve its knowledge of the data by making better use of the limited labelled samples. Furthermore, where there is a class imbalance, such as more benign than malignant instances in the WDBC, RBF+SSL is able to balance the model's exposure to both classes, increasing classification performance.

4 Discussion

AI techniques have advanced quickly in recent years allowing consistent improvements in medical image processing, computer-aided diagnosis, image interpretation, fusion, registration, segmentation, image-guided therapy, image retrieval, and image analysis. To date, understanding the associated data structures and statistics to the aim of converting the ML algorithm into a product working consistently in broad clinical use is still complex and prone to ethical issues [91]. This makes the scientific research about the ML algorithms performance, as well as the most recent development of DL techniques, a focus of the discussion about early pathology detection and medical imaging interpretation times and health costs saving, to the final goal of address to society's expectations towards innovative health solutions based on concrete and health safe models and policies.

This study selected and analysed the recent literature on the application of ML techniques in the field of preventive BC diagnosis along three dimensions: Accuracy, AUC and Dataset cardinality. Papers that applied the same models to different datasets and used different feature selection methods have been compared. The selected models included LR, RF, DT, Boosting algorithms (such as XGBoost), and Artificial Neural Networks. To the aim of reducing the heterogeneity in terms of cardinality among the reviewed studies, the ones led by applying DL models only, have been excluded as they are generally based on substantially larger datasets [92]. Additionally, our choice of excluding DL studies on large amounts of data is motivated by the most realistic condition of managing limited data sets of medical imaging. This is the fundamental issue affecting the creation of ML models that simultaneously learns from its surroundings and it largely depends on the time-consuming procedures of medical picture segmentation and annotation, which greatly limit medical imaging data collection [92].

Based on the analysis of the reviewed papers, the performances of the selected models showed to be generally comparable. Figure 11 plots the performance results across different models. It's important to note that the apparent superiority of the RBF model's performance may be skewed by the fact that only one paper meeting the selection criteria was found for this particular model.

This suggests that while different models may exhibit varying degrees of performance in different scenarios, no single model emerged as consistently superior across all studies. The choice of the most appropriate model may depend on factors such as the specific
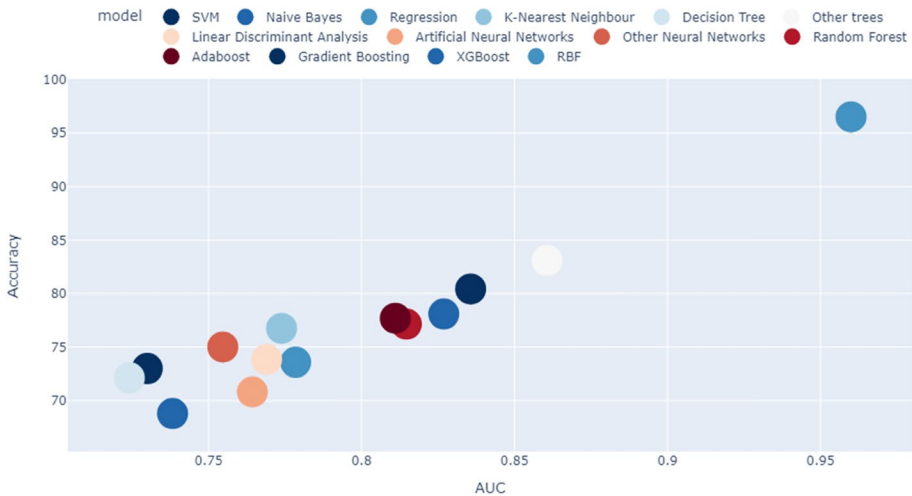
Fig. 11 Mean results for analysed papers, grouped by ML model

dataset, feature selection methods, and other considerations relevant to the particular application of preventive breast cancer diagnosis.

The paper by Al-Azzam et al. [39] demonstrated the best performance among the papers reviewed in terms of tumor type diagnosis. The authors explored various algorithms and combined them with both SL and SSL approaches to achieve high specificity in their diagnosis.

In SL, the use of labelled data is required for training the algorithm. However, labelling data can be a time-consuming task. The advantage of SSL is that it can achieve competitive results with less labelled data, reducing the overall cost of diagnosis. The authors found that SSL algorithms yielded very similar accuracies to SL algorithms, ranging from 91 to 98%.

These findings indicate that SSL is a promising and competitive solution for tumor type diagnosis, particularly when dealing with a small sample of labelled data and limited computational resources. SSL algorithms can provide accurate results while optimising the use of labelled data, making them an efficient approach for tumor type classification.

Our results agree to the ones of a recent review concentrated on the most widely used ML approaches (SVM, DT, Nearest Neighbour, Naive Bayesian Networks, ANN, and Convolutional Neural Networks) that highlighted the need of using labelled images throughout the training of SL methods application [92].

5 Conclusion

The analysis of the reviewed literature enables us to draw conclusions regarding the factors that influence the performance of the selected models. One notable finding is that the feature selection process has a greater impact on model performance compared to the size of the dataset. This highlights the importance of selecting relevant and informative features and the need to have large quantities of diagnostic labelled images available to achieve accurate predictions.

This review confirms previous findings: Semi-Supervised Learning is a promising approach, with respect to Supervised Learning, since by exploiting a smaller set of labelled data it is able to achieve similar results, in terms of accuracy [93]. This makes the diagnosis process for breast cancer screening faster and cheaper. The future research challenge is to enhance ML algorithm efficiency and accuracy for fueling precision medicine, in a holistic patient-centric approach that integrates personal, clinical, genetic and environmental data in order to improve both diagnosis (faster, more accurate, cheaper) and therapy (reducing side effects and improving efficacy) [94]. In this perspective, ML is the base for fueling progress over time. However ethical, social, and legal implications of using Artificial Intelligence in healthcare need to be investigated in depth [95].

To further enhance the accuracy of predictive models in BC risk assessment, future work should focus on standardising image acquisition scanners, lighting and enlargement factors configurations, sizes, as well as incorporating large volumes of personal and behavioural health data. This additional data can provide valuable insights and improve the model's ability to accurately predict an individual's risk of developing BC. Continued research and advancements in this area can contribute to more effective and personalised BC risk prediction models in the future.

Data Availability The data that support the findings of this study are openly available in the reviewed manuscripts.

Declarations

References

1. European Commission (2024) Horizon Europe [Internet]. European Commission. Available from: https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/fundingprogrammes-and-open-calls/horizon-europe_en

2. Alqahtani B, Alnajrani B, Alhaidari F (2021) Machine learning for predicting cancer disease: comparative analysis. In: Enabling machine learning applications in data science. Springer, pp 237–248

3. Mathappan N, Soundariya R, Natarajan A, Gopalan SK (2020) Biomedical analysis of breast cancer risk detection based on deep neural network. Int J Med Eng Inf 12(6):529–541

4. Dafni U, Tsourti Z, Alatsathianos I (2019) Breast cancer statistics in the European union: incidence and survival across European countries. Breast Care 14(6):344–353

5. Kalafi E, Nor N, Taib N, Ganggayah M, Town C, Dhillon S (2019) Machine learning and deep learning approaches in breast cancer survival prediction using clinical data. Folia Biol 65(5/6):212–220

6. Zielonke N, Kregting LM, Heijnsdijk EA, Veerus P, Heinävaara S, McKee M, de Kok IM, de Koning HJ, van Ravesteyn NT, collaborators, E.-T. (2021) The potential of breast cancer screening in europe. Int J Cancer 148(2):406–418

7. Kaklamanis MM, Filippakis M, Touloupos M, Christodoulou K (2020) An experimental comparison of machine learning classification algorithms for breast cancer diagnosis. In: 16th European, Mediterranean, and Middle Eastern Conference on Information System, EMCIS 2019. Springer India, pp 18–30

8. Yassin NI, Omran S, El Houby EM, Allam H (2018) Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. Comput Methods Programs Biomed 156:25–45

9. World Health Organization. World Health Organization Releases AI Guidelines for Health [Internet]. GovTech; [cited 2024 March 7]. Available from: https://www.govtech.com/products/world-health-organization-releases-ai-guidelines-for-health

10. EBCP-EC (2021) EU cancer plan. European commission. https://ec.europa.eu/health/sites/default/files/noncommunicablediseases/docs/eucancer-planen.pdf

11. Waks AG, Winer EP (2019) Breast cancer treatment: a review. JAMA 321(3):288–300

12. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A (2017) Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin 67(6):439–448

13. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, Jemal A, Siegel RL (2019) Breast cancer statistics, 2019. CA Cancer J Clin 69(6):438–451

14. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics, 2021. CA Cancer J Clin 71(1):7–33

15. Azamjah N, Soltan-Zadeh Y, Zayeri F (2019) Global trend of breast cancer mortality rate: a 25-year study. Asian Pac J Cancer Prev 20(7):2015

16. Hortobagyi GN, de la Garza Salazar J, Pritchard K, Amadori D, Haidinger R, Hudis CA, Khaled H, Liu M-C, Martin M, Namer M et al (2005) The global breast cancer burden: variations in epidemiology and survival. Clin Breast Cancer 6(5):391–401

17. Sun Y-S, Zhao Z, Yang Z-N, Xu F, Lu H-J, Zhu Z-Y, Shi W, Jiang J, Yao P-P, Zhu H-P (2017) Risk factors and preventions of breast cancer. Int J Biol Sci 13(11):1387

18. Britt KL, Cuzick J, Phillips K-A (2020) Key steps for effective breast cancer prevention. Nat Rev Cancer 20(8):417–436

19. Carlson RW, Allred DC, Anderson BO, Burstein HJ, Carter WB, Edge SB, Erban JK, Farrar WB, Forero A, Giordano SH et al (2010) Breast cancer: noninvasive and special situations. J Natl Compr Canc Netw 8(10):1182–1207

20. El Naqa I, Murphy MJ (2015) What is machine learning? Springer International Publishing

21. Bonaccorso G (2017) Machine learning algorithms. Packt Publishing Ltd, Birmingham. ISBN 978–1–78588–962–2

22. Kelleher JD (2019) Deep learning. The MIT Press

23. Bhardwaj R, Nambiar AR, Dutta D (2017) A study of machine learning in healthcare. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol 2. pp 236–241. https://doi.org/10.1109/COMPSAC.2017.164

24. Bener A, Barışık CC, Acar A, Özdenkaya Y (2019) Assessment of the Gail Model in estimating the risk of breast cancer: effect of cancer worry and risk in healthy women. Asian Pac J Cancer Prev 20(6):1765–1771. https://doi.org/10.31557/APJCP.2019.20.6.1765

25. Stark GF, Hart GR, Nartowt BJ, Deng J (2019) Predicting breast cancer risk using personal health data and machine learning models. PLoS ONE 14(12):0226765

26. Sharma S, Aggarwal A, Choudhury T (2018) Breast cancer detection using machine learning algorithms. In: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE, pp 114–118

27. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I (2018) Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. Asian Pac J Cancer Prev 19(7):1747

28. Osareh A, Shadgar B (2010) Machine learning techniques to diagnose breast cancer. In: 2010 5th International Symposium on Health Informatics and Bioinformatics. IEEE, pp 114–120

29. Nassif AB, Talib MA, Nasir Q, Afadar Y, Elgendy O (2022) Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artif Intell Med 102276

30. (2023) Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. Computers 12(5):91. https://doi.org/10.3390/computers12050091
31. Berrar D (2019) Cross-validation. Encycl Bioinforma Comput Biol 1–3:542–545
32. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
33. Wei Q, Dunbrack RL Jr (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE 8(7):e67863. https://doi.org/10.1371/journal.pone.0067863
34. Auriemma Citarella A, Di Biasi L, Risi M et al (2022) SNARER: new molecular descriptors for SNARE proteins classification. BMC Bioinformatics 23:148. https://doi.org/10.1186/s12859-022-04677-z
35. Xue H, Yang Q, Chen S (2009) Svm: support vector machines. In: The top ten algorithms in data mining. Taylor & Francis Group, LLC, pp 51–74
36. Heidari M, Mirniaharikandehei S, Liu W, Hollingsworth AB, Liu H, Zheng B (2019) Development and assessment of a new global mammographic image feature analysis scheme to predict likelihood of malignant cases. IEEE Trans Med Imaging 39(4):1235–1244
37. Huo L, Tan Y, Wang S, Geng C, Li Y, Ma X, Wang B, He Y, Yao C, Ouyang T (2021) Machine learning models to improve the differentiation between benign and malignant breast lesions on ultrasound: a multicenter external validation study. Cancer Manag Res 13:3367
38. Park EK, Lee K-S, Seo BK, Cho KR, Woo OH, Son GS, Lee HY, Chang YW (2019) Machine learning approaches to radiogenomics of breast cancer using low-dose perfusion computed tomography: predicting prognostic biomarkers and molecular subtypes. Sci Rep 9(1):1–11
39. Al-Azzam N, Shatnawi I (2021) Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. Ann Med Surg 62:53–64
40. Jiang N, Tian T, Chen X, Zhang G, Pan L, Yan C, Yang G, Wang L, Cao X, Wang X (2021) A diagnostic analysis workflow to optimal multiple tumor markers to predict the nonmetastatic breast cancer from breast lumps. J Oncol 2021
41. Sun W, Tseng T-LB, Qian W, Saltzstein EC, Zheng B, Yu H, Zhou S (2018) A new near-term breast cancer risk prediction scheme based on the quantitative analysis of ipsilateral view mammograms. Comput Methods Programs Biomed 155:29–38
42. Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, Feldman M, Gilmore H, Madabhushi A (2018) Quantitative nuclear histomorphometry predicts oncotype dx risk categories for early stage er+ breast cancer. BMC Cancer 18(1):1–15
43. Sathipati SY, Ho S-Y (2018) Identifying a mirna signature for predicting the stage of breast cancer. Sci Rep 8(1):1–11
44. Guo Z-W, Cai G-X, Zhai X-M, Lin L, Yang X-X, Li M, Li K, Zhou C-L, Liu T, Han B-W et al (2021) Plasma-derived extracellular vesicles circular rnas serve as biomarkers for breast cancer diagnosis. Front Oncol 4575
45. Chen Y, Wang Z, Yin G, Sui C, Liu Z, Li X, Chen W (2022) Prediction of her2 expression in breast cancer by combining pet/ct radiomic analysis and machine learning. Ann Nucl Med 36(2):172–182
46. Lei C, Wei W, Liu Z, Xiong Q, Yang C, Yang M, Zhang L, Zhu T, Zhuang X, Liu C et al (2019) Mammography-based radiomic analysis for predicting benign bi-rads category 4 calcifications. Eur J Radiol 121:108711
47. Nanglia S, Ahmad M, Ali Khan F, Jhanjhi NZ (2022) An enhanced predictive heterogeneous ensemble model for breast cancer prediction. Biomed Signal Process Control 72:103279. https://doi.org/10.1016/j.bspc.2021.103279
48. Hao W, Gong J, Wang S, Zhu H, Zhao B, Peng W (2020) Application of mri radiomics-based machine learning model to improve contralateral bi-rads 4 lesion assessment. Front Oncol 2265
49. Zhu Y, Yang L, Shen H (2021) Value of the application of ce-mri radiomics and machine learning in preoperative prediction of sentinel lymph node metastasis in breast cancer. Front Oncol 11:757111–757111
50. Tomas RC, Sayat AJ, Atienza AN, Danganan JL, Ramos MR, Fellizar A, Notarte KI, Angeles LM, Bangaoil R, Santillan A et al (2022) Detection of breast cancer by atr-ftir spectroscopy using artificial neural networks. PLoS ONE 17(1):0262489
51. Arefan D, Chai R, Sun M, Zuley ML, Wu S (2020) Machine learning prediction of axillary lymph node metastasis in breast cancer: 2d versus 3d radiomic features. Med Phys 47(12):6334–6342
52. Vamvakas A, Tsivaka D, Logothetis A, Vassiou K, Tsougos I (2022) Breast cancer classification on multiparametric mri–increased performance of boosting ensemble methods. Technol Cancer Res Treat 21:15330338221087828

53. Harrington L, diFlorio-Alexander R, Trinh K, MacKenzie T, Suriawinata A, Hassanpour S (2018) Prediction of atypical ductal hyperplasia upgrades through a machine learning approach to reduce unnecessary surgical excisions. JCO Clin Cancer Inform 2:1–11

54. Song L, Lu H, Yin J (2020) Preliminary study on discriminating her2 2+ amplification status of breast cancers based on texture features semiautomatically derived from pre-, post-contrast, and subtraction images of dce-mri. PLoS ONE 15(6):0234800

55. Daimiel Naranjo I, Gibbs P, Reiner JS, Lo Gullo R, Sooknanan C, Thakur SB, Jochelson MS, Sevilimedu V, Morris EA, Baltzer PA et al (2021) Radiomics and machine learning with multiparametric breast mri for improved diagnostic accuracy in breast cancer diagnosis. Diagnostics 11(6):919

56. Bhardwaj D, Dasgupta A, DiCenzo D, Brade S, Fatima K, Quiaoit K, Trudeau M, Gandhi S, Eisen A, Wright F et al (2022) Early changes in quantitative ultrasound imaging parameters during neoadjuvant chemotherapy to predict recurrence in patients with locally advanced breast cancer. Cancers 14(5):1247

57. Jiang Z, Song L, Lu H, Yin J (2019) The potential use of dce-mri texture analysis to predict her2 2+ status. Front Oncol 9:242

58. Moghadas-Dastjerdi H, Sha-E-Tallat HR, Sannachi L, Sadeghi-Naini A, Czarnota GJ (2020) A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative ct and machine learning. Sci Rep 10(1):1–11

59. Liu J, Sun D, Chen L, Fang Z, Song W, Guo D, Ni T, Liu C, Feng L, Xia Y et al (2019) Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer. Front Oncol 980

60. Fusco R, Granata V, Pariante P, Cerciello V, Siani C, Di Bonito M, Valentino M, Sansone M, Botti G, Petrillo A (2021) Blood oxygenation level dependent magnetic resonance imaging and diffusion weighted mri imaging for benign and malignant breast cancer discrimination. Magn Reson Imaging 75:51–59

61. Rana SP, Dey M, Tiberi G, Sani L, Vispa A, Raspa G, Duranti M, Ghavami M, Dudley S (2019) Machine learning approaches for automated lesion detection in microwave breast imaging clinical data. Sci Rep 9(1):1–12

62. Tahmassebi A, Wengert GJ, Helbich TH, Bago-Horvath Z, Alaei S, Bartsch R, Dubsky P, Baltzer P, Clauser P, Kapetas P et al (2019) Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. Invest Radiol 54(2):110

63. Leung KM (2007) Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007, 123–156

64. Esmaeili M, Ayyoubzadeh SM, Ahmadinejad N, Ghazisaeedi M, Nahvijou A, Maghooli K (2020) A decision support system for mammography reports interpretation. Health Inf Sci Syst 8(1):1–8

65. Fanizzi A, Lorusso V, Biafora A, Bove S, Comes MC, Cristofaro C, Digennaro M, Didonna V, Forgia DL, Nardone A et al (2021) Sentinel lymph node metastasis on clinically negative patients: preliminary results of a machine learning model based on histopathological features. Appl Sci 11(21):10372

66. Worster A, Fan J, Ismaila A (2007) Understanding linear and logistic regression analyses. Can J Emerg Med 9(2):111–113

67. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J et al (2020) Predicting breast cancer in chinese women using machine learning techniques: algorithm development. JMIR Med Inform 8(6):17364

68. Xu Y, Ju L, Tong J, Zhou C, Yang J (2019) Supervised machine learning predictive analytics for triple-negative breast cancer death outcomes. Onco Targets Ther 12:9059

69. Wei X, Lu Q, Jin S, Li F, Zhao Q, Cui Y, Jin S, Cao Y, Fu MR (2021) Developing and validating a prediction model for lymphedema detection in breast cancer survivors. Eur J Oncol Nurs 54:102023

70. Perre SV, Duron L, Milon A, Bekhouche A, Balvay D, Cornelis FH, Fournier L, Thomassin-Naggara I (2021) Radiomic analysis of htr-dce mr sequences improves diagnostic performance compared to birads analysis of breast mr lesions. Eur Radiol 31(7):4848–4859

71. Strzelecki M, Szczypinski P, Materka A, Klepaczko A (2013) A software tool for automatic classification and segmentation of 2d/3d medical images. Nucl Instrum Methods Phys Res, Sect A 702:137–140

72. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A (2009) Mazda—a software package for image texture analysis. Comput Methods Programs Biomed 94(1):66–76

73. Peterson LE (2009) K-nearest neighbor. Scholarpedia 4(2):1883

74. Abu Alfeilat HA, Hassanat AB, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, Prasath VS (2019) Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big Data 7(4):221–248

75. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp 785–794
76. Quinlan JR (1996) Learning decision tree classifiers. ACM Comput Surv (CSUR) 28(1):71–72
77. Zhao Y, Chen R, Zhang T, Chen C, Muhelisa M, Huang J, Xu Y, Ma X (2021) Mri-based machine learning in differentiation between benign and malignant breast lesions. Front Oncol 11
78. Xanthopoulos P, Pardalos PM, Trafalis TB, Xanthopoulos P, Pardalos PM, Trafalis TB (2013) Linear discriminant analysis. Robust Data Mining 27–33
79. Khan A, Farooq H (2012) Principal component analysis-linear discriminant analysis feature extractor for pattern recognition. ArXiv, abs/1204.1177
80. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, Plecha D, Madabhushi A (2017) Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. Breast Cancer Res 19(1):1–14
81. Hassoun MH et al (1995) Fundamentals of artificial neural networks. IEEE transactions on information theory, vol 42, N°.4. MIT Press, Cambridge
82. Danala G, Patel B, Aghaei F, Heidari M, Li J, Wu T, Zheng B (2018) Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms. Ann Biomed Eng 46(9):1419–1431
83. Biau G, Scornet E (2016) A random forest guided tour. TEST 25(2):197–227
84. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
85. Romeo V, Cuocolo R, Apolito R, Stanzione A, Ventimiglia A, Vitale A, Verde F, Accurso A, Amitrano M, Insabato L et al (2021) Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions. Eur Radiol 31(12):9511–9519
86. Schapire RE (2013) Explaining adaboost. In: Empirical inference. Springer, Springer, Berlin, Heidelberg, pp 37–52. https://doi.org/10.1007/978-3-642-41136-6_5
87. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al (2015) Xgboost: extreme gradient boosting. R package version 0.4–2 1(4), 1–4
88. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat:1189–1232
89. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30
90. Arora Y, Singhal A, Bansal A (2014) A study of applications of RBF network. Int J Comput Appl 94.2
91. Dreyer KJ, Geis JR (2017) when machines think: radiology's next frontier. Radiology 285(3):713–718. https://doi.org/10.1148/radiol.2017171183
92. Jalloul R, Chethan HK, Alkhatib R (2023) A review of machine learning techniques for the classification and detection of breast cancer from medical images. Diagnostics (Basel) 13(14):2460. https://doi.org/10.3390/diagnostics13142460
93. Eckardt JN, Bornhäuser M, Wendt K, Middeke JM (2022) Semi-supervised learning in cancer diagnostics. Front Oncol 14(12):960984. https://doi.org/10.3389/fonc.2022.960984.]
94. Yan J, Liu Z, Du S, Li J, Ma L, Li L (2020) Diagnosis and treatment of breast cancer in the precision medicine era. Methods Mol Biol 2204:53–61. https://doi.org/10.1007/978-1-0716-0904-0_5
95. Dileep G, Gianchandani Gyani SG (2022) Artificial intelligence in breast cancer screening and diagnosis. Cureus 14(10):e30318. https://doi.org/10.7759/cureus.30318

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Giada Anastasi[1,4] ⬤ **· Michela Franchini**[1] **· Stefania Pieroni**[1] **· Marina Buzzi**[2] **·
Maria Claudia Buzzi**[2] **· Barbara Leporini**[3] **· Sabrina Molinaro**[1]

✉ Giada Anastasi
   giadaanastasi@cnr.it

   Michela Franchini
   michela.franchini@cnr.it

   Stefania Pieroni
   stefania.pieroni@cnr.it

   Marina Buzzi
   marina.buzzi@iit.cnr.it

   Maria Claudia Buzzi
   claudia.buzzi@iit.cnr.it

   Barbara Leporini
   barbara.leporini@isti.cnr.it

   Sabrina Molinaro
   sabrina.molinaro@ifc.cnr.it

[1]   IFC, National Council of Research, Pisa, Italy

[2]   IIT, National Council of Research, Pisa, Italy

[3]   ISTI, National Council of Research, Pisa, Italy

[4]   Department of Computer Science, University of Pisa, Pisa, Italy