Machine Learning and Deep Learning Algorithms for Breast Cancer Diagnosis: A Comprehensive Review

[1]Jooda J. O., [2]Komolafe T. A., [3]Olayinka A. A., [4]Abegunde O., and [5]Ibikunle A. A.
[1,2,3,4,5] Department of Computer Engineering, Redeemer's University, Nigeria.
joodaj@run.edu.ng

Received: 21-MARCH-2026; Reviewed: 25-MAY-2026; Accepted: 30-JUNE-2026
https://dx.doi.org/10.46792/fuoyejet.v11i2.1550

REVIEW PAPER

Abstract —Cancer continue to be among the foremost non-infectious contributor to mortality rare globally. Early detection of breast cancer (BC) is critical to reducing mortality rates and improving patient outcomes. Driven by innovations in machine learning (ML) and deep learning (DL), researchers have developed increasingly precise and reliable predictive models for early-stage BC diagnosis. The study aims to ascertain the most accurate and suitable algorithms for BC prediction. To achieve this, this paper provides a detailed review and critical analysis of various ML and DL algorithms utilized in early-stage BC assessment and detection. Our analysis demonstrates that XGBoost and ensemble/hybrid models currently yield the highest predictive precision. Concurrently, Support Vector Machines (SVM) remain a consistently robust traditional tool, achieving accuracy rates of up to 99%. Furthermore, we identify a growing paradigm shift toward Explainable AI (XAI) to enhance algorithmic transparency and clinical trust. To establish a scalable, robust, and highly accurate framework for modern clinical settings, this study suggests that future developments focus on the hybridization of optimized ensemble models integrated with advanced feature selection techniques.

Keywords—Breast Cancer, Ensemble Learning, Explainable AI, Machine Learning, Prediction, Weka, Random Forest, SVM.

1 INTRODUCTION

Breast cancer (BC) continue to be among the foremost occurring and severe malignancies that affect mostly women and rare men (one out of hundred) worldwide (Centers for Disease Control and Prevention, 2024). According to Global Cancer Observatory (GLOBOCAN), BC recorded 685,000 deaths and 2.3 million new cases worldwide in the year 2020 alone, making it one of the foremost usually diagnosed cancer in the world, this causes cancer-related mortality especially in women (Sung et al., 2021). The disease do starts from the breast tissue and it can forcefully enter in adjacent organs if not detected in early stage thus it can become life threatening diseases. The burden may be disproportionately high due to limited screening infrastructure, late-stage diagnosis and constrained therapeutic resources. It is widely accepted that early and accurate detection is the most important factor in improving the patient's survival. 5-years survival for localized breast cancer is almost 99%, compared with approximately 29% for metastatic disease (American Cancer Society, 2023). The survival probability as a function of the stage can be modeled as follows:

$$S(stage) = 1 / (1 + e^{(\beta_0 + \beta_1 \cdot stage)}) \quad (1)$$

Traditionally, BC diagnosis relies on a merge of clinical imaging modalities, primarily mammography, ultrasound, and magnetic resonance imaging (MRI), supplemented with histopathological biopsy and immunohistochemical profiling. These methods have been the foundation of clinical practice for decades, but have well-documented limitations. Mammographic screening has high false-positive rates of 10% to 15% leading to unnecessary biopsies and patient anxiety Baloyi et al., 2025. The false positive rate is mathematically expressed as:

$$FPR = FP / (FP + TN) \quad (2)$$

Histopathological grading is an inherently subjective process and is sensitive to inter-observer variability with concordance rates as low as 75% reported even among expert pathologists. In addition, the large amount of imaging data produced in big screening programs puts pressure on radiological capacity and raises the chances of perceptual error when cognitive demand is high (Waite et al., 2017).In this sense, data mining (DM) and machine learning (ML) have appeared as transformative tools to solve many limitations existing in the traditional diagnostic pipelines. ML is a large family of statistical and computational techniques that allow systems to learn discriminative patterns from data without explicit rule-based programming. Deep learning (DL) is a subfield of ML using multi-layer artificial neural networks. It has shown powerful ability to learn hierarchical feature representations directly

Corresponding Author
Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Jooda J. O., Komolafe T. A., Olayinka A. A., Abegunde O., and Ibikunle A. A. (2026). Machine Learning and Deep Learning Algorithms for Breast Cancer Diagnosis: A Comprehensive Review. FUOYE Journal of Engineering and Technology (FUOYE**JET), 11(2), 221-240.** https://dx.doi.org/10.46792/fuoyejet.v11i2.1550

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

from raw imaging data without the requirement of handcrafted feature engineering . These methods have been systematically applied to the whole breast cancer diagnosis pipeline, including mass detection, microcalcification identification, lesion segmentation, tumour grading, molecular subtype classification, and prognosis prediction (Shen et al., 2019). They range from classical algorithms such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbours (KNN), to state-of-the-art convolutional neural networks (CNNs), recurrent architectures, and transformer-based models. These automated approaches have helped the oncologists to handle the disease more effectively and economically thereby enhancing the medical decision support. Proposals for identifying BC at an early stage by Computer-aided check-up systems using clinical data (Sharma et al., 2022) are a safer and faster alternative to unnecessary surgical procedures and eliminating overall treatment expenses. The application of ML models are important for the prediction of BC in this modern days. In this paper, some ML and data mining (DM) algorithms applied for BC prediction were reviewed and evaluated, in order to find the most accurate and effective computational approaches. This was done by synthesizing previous work on linear methods (e.g. Logistic Regression), non-linear methods (e.g. Decision Trees (DT), KNN, Naive Bayes and SVM) and ensemble frameworks (e.g. RF and AdaBoost). The growing literature on hybrid models combining these methods was also reviewed. Different sections systematically compare each algorithm in terms of predictive performance and finally identify the best models for clinical application. The organization of this paper is as follows: Section one is the introduction, section 2 has the summary of the basic ML and ensemble concepts used in the BC prediction, and then we briefly introduce the deep learning approaches, the methodology of our study selection was presented in section, and a thorough survey of the current state-of-the-art research. Finally, a detailed discussion was provided and concluding remarks on the future of computational diagnostics in oncology.

2 LITERATURE REVIEW

This section reviewed the synthesis of previous work on linear methods (e.g. Logistic Regression), non-linear methods (e.g. DT, KNN, NB and SVM) and ensemble frameworks (e.g. RF and AdaBoost) for BC prediction.

2.1 Machine Learning Techniques for Predicting Breast Cancer

ML is A self-learning algorithm designed to learn from past datasets. ML model predictions are made for future results by analyzing information from large processed volume of data and utilizing the trained model Kar et al., 2025. For breast cancer prediction, the ML algorithms include:

2.1.1 K-Nearest Neighbors

KNN is amongst the top easy, simplest and the most straightforward techniques used in ML for classification and regression problem. The KNN algorithms classification of new facts points is done utilizing facts by using similarity measures mainly, Classification is obtained as shown in Figure 1 through majority vote to its neighbors (Arora et al, 2020). K-NN performs no learning on the dataset and its attributes provided just like many ML algorithms. K-NN classify the training data to the class which has the first K neighbours closest to it by just utilizing the points from the training data and discovers the K number of nearest neighbours to that data point using Euclidean Distance (Thomas et al., 2020). KNN classification is done based on feature similarity extracted from similar-featured data from huge datasets (Kar et al., 2025).

The KNN algorithm is an effective approach for predicting BC due it usual use for pattern recognition. Each class is given equal importance in this technique when used for recognizing patterns.

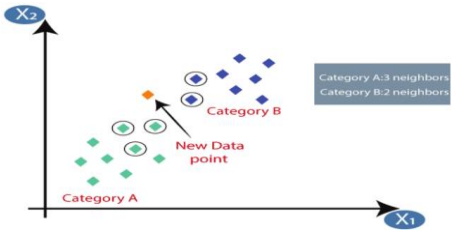

Fig. 1. KNN Algorithm (Arora et al., 2020)

2.1.2 Random Forest

The Random Forest (RF) algorithm is an ensemble method predicated on bagging, functioning as a fundamental component of ML and a supervised learning technique (Kar et al., 2025). RF is utilized to address regression and classification problems by building several decision trees, as shown in figure 2. The final outcomes are based on the majority of the trees and are decided by RF (Arora et al., 2020).

By constructing several decision trees and integrating their results to produce the ultimate forecast in particular, RF makes advantage of decision trees. Due to RF resistant to overfitting and obtaining of excellent accuracy by using variety of decision trees on random subsets of data, makes RF

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

to be extensively used in the prediction of BC (Kar et al., 2025).

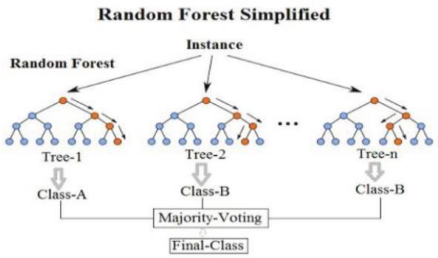

Fig. 2. Random Forest Algorithm (Arora et al., 2020)

2.1.3 Naive Bayes

To inform assumptions made by the NB method, large training datasets are utilized. Bayesian approach is used to compute probabilities. The probabilities of noisy data used as input with excellent accuracy are determined by this algorithm. This analogy classifier is utilized to compare the training datasets and training tuples Kar et al., 2025. The first ML model applied to this dataset for binary classification was the NB classifier. NB is a generative ML technique utilized for classification projects which is based on the well-known Bayes rule. NB assumes that each feature's value is independent of the others and ignores any correlation between the attributes. Training of any NB model is extremely effective because of this. NB performed admirably in many real-world classification projects despite its straightforward and uninformed methodology (Arora et al., 2020).

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (3)$$

Where *P(A)* is the Prior Probability, *P(B)* is the Marginal Likelihood, *P(B|A)* is the Likelihood and *P(A|B)* is the Posterior Probability. To identify a data point's class, Equation (3) is used by the NB algorithm. After the posterior probability is computed based on the vector's location in the feature space, the class with the higher posterior probability is designated as the data point class (Thomas *et al.*, 2020).NB requirement for just a modest amount of training data to generate a classification model is one of it major benefit as supervised machine learning model. Specific class feature that is unconnected to any other feature and has distinct assumptions and properties that may be entirely different from one another is been compared by NB algorithm According to probabilistic approaches. Consequently, the best accuracy is attributed to the resulting class with the highest probability (Arora et al., 2020).

2.1.4 Support Vector Machine

SVM are used for both regression and classification projects as a supervised learning approach. SVM utilizes both theoretical and numerical functions to handle regression computations. SVM is a powerful ML technique that produces highly accurate prediction results for big datasets utilizing 2D and 3D modeling (Kar et al., 2025). Due to the fact that it uses vectors in the feature space to determine a new vector's class, brings about the name SVM (Thomas et al., 2020). SVM is usually used for unlabeled data, classification task that is more relevant to a certain area, hyperplanes that separate two data points as seen in figure 3 (Arora et al., 2020). The new vector is determined to in class one or class two by using Maximum Margin Hyper-plane (MMH) . In a certain scenario when class one and class two are two distinct classes, the data that points to the left of MMH or sits outside the negative hyper-plane, belongs to class one; otherwise, it belongs to class two. When there are more than two classes SVMs can also be utilized (Thomas et al., 2020).

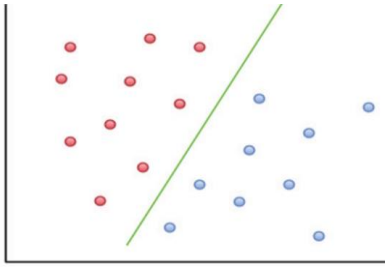

Fig. 3. Support Vector Machine Algorithm (Arora et al., 2020)

2.1.5 Decision Tree

Both regression and classification task can be done utilizing DT, which is a powerful supervised ML technique, even though it is primarily utilized for classification tasks (Thomas et al., 2020, Kar et al., 2025). Discrete input/output and continuous variables can both be utilized with it. The application of DT algorithm on the BC medical dataset was attempted in researches by clustering a dataset into smaller subgroups, allowing for very accurate predictions (Arora et al., 2020). CART, C4.5, and C5.0 which are conditional tree models are instances of the DT approach (Kar et al., 2025).

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

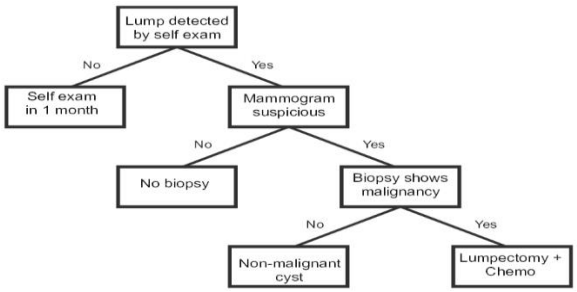

Fig. 4. Decision Tree Algorithm (Arora et al., 2020)

To start with, the terminal nodes indicate the class or prediction value that has to be predicted, displaying a DT structure akin to a tree when each internal node serves as a vector test condition. DT can be expensive to train computationally and no accurate results will be yielded when there are many classes and few training observations. However, for classifying some class labels, it works well (Thomas et al., 2020).

2.1.6 Logistic Regression

Logistic Regression (LR) is a the discriminative supervised ML statistical model that employs Multiple dependent variables used for classification problems with binary output (Kar et al., 2025). Instead of Gaussian distribution found in conventional linear regression, LR employs a Bernoulli distribution. After computing the linear combination of inputs, a sigmoid function is typically employed as the normalization function to guarantee that the model produces a classification result between 0 and 1. Once a threshold is established, each input is classed to identify which class the group of inputs falls. The sigmoid function is shown in Equation (5) (El-Shair et al., 2020).

$$X_{standardized}^{(d)} = \frac{X^{(d)} - \mu(X^{(d)})}{\sigma(X^{(d)})} \quad (4)$$

$$\text{Sigmoid } (W^T X) = \frac{1}{1+e^{W^T X}} \quad (5)$$

The data must be normalized using equation (4) prior to training LR models first, employing only the training set's mean and standard deviation. To guarantee that no data from the test set affects the model, the test set is standardized using the same values (El-*Shair et al.*, 2020).

2.1.7 Artificial Neural Network (Multilayer Perceptron)

The well known ML technique that uses collective solutions, distributed memory, network design and parallel computing is called Artificial Neural Network (ANN). It consists of three layers: input, hidden, and output and very useful for spotting intricate patterns (Kar et al., 2025). ANN may be applied to a wide range of projects like voice recognition, regression, classification, targeted marketing, and more being One of the most powerful and sophisticated ML algorithms, (Thomas et al., 2020).To obtain the best classification results, ANN trains many layers (representing interconnected neurons) using an error function that is minimized over a large number of iterations which is an extension of LR. The number of input nodes equals the number of independent variables which serve as the first input nodes are taken. These inputs are then sent on to the next neuron, where activation functions are applied after the weighted through the synapses. Until the weights have reached the output layer, the process is repeated. The operation of ANNs is easily understood though a little complicated. The expected output is taken as $\hat{y}$ while real output value is taken as y. To determine the Cost Function C, Equation 6 is utilized.

$$C = \frac{1}{2}(\hat{y} - y) \quad (6)$$

The goal of ANNs is to minimize the cost function and this archive using the Back Propagation technique by feeding back the data into the Neural Network. The weights are adjusted if the value of the cost function exceeds a specific threshold, and until the C function is minimal, the process is repeated. For each observation in the dataset,the training of the ANN is performed until the ANN is ready for real-world classification. The architecture of an ANN is shown in Figure 5. Ten neurons were used as an input layer and two hidden layers of six neurons by the authors. To forecast whether a patient would suffer from breast cancer or not, only one neuron is employed at the end of the stated design, (Thomas et al., 2020).

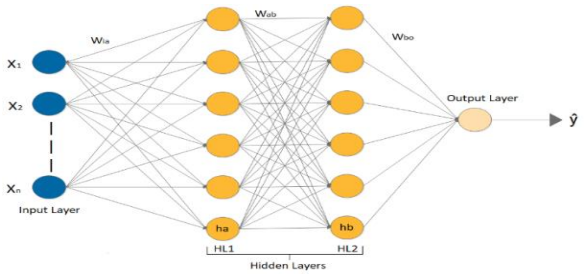

Fig. 5. Architecture of used Artificial Neural Network (Thomas et al., 2020)

2.1.8 Convolutional Neural Networks

A deep learning model that is specifically built to process and evaluate data with a grid-like topology, such as images (2D pixel grids), time-series signals, and audio is called Convolutional Neural Network (CNN). In image-based applications like ultrasound

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

image and mammography classification for detecting BC, CNNs perform particularly well . To extract features (such as edges, textures, and patterns) from images, CNNs use convolutional layers. In BC prediction, where image-based data (such as histological pictures or radiological scans) are common, this is especially effective. CNNs can accurately categorize images (for example, benign vs. malignant) (Kar et al., 2025).

2.1.9 Recurrent Neural Networks

Neurons' outputs substitute as feedback inputs to other neurons in these networks. Due to their feedback feature, These networks are suitable for dynamic information processing, including system identification and optimization, process management, and time series prediction. Since they specialize on sequential data analysis, Recurrent Neural Networks (RNNs) are good for temporal datasets. When the data consists of time-series information, such as changes in patient medical histories or tumor markers over time, RNNs can be used to forecast BC. RNNs can capture the temporal connections between inputs therefore can help forecast future outcomes because (Kar et al., 2025).

2.1.10 K-Means Clustering

The identification of similarities between data points and partitioning them into smaller clusters is done using a clustering approach called K-Means algorithm. to at least one cluster Each data point is assigned, which is appropriate for evaluating large datasets (Kar et al., 2025).Although assigning similar data points to their respective classes is similar-featured to creating a cluster of comparable data points, therefore K-Means approach is widely utilized for classification and data clustering (Thomas et al., 2020). Clustering is a ML technique that groups or clusters data items. Based on a particular dataset, a clustering algorithm can categorize each data point into an explicit cluster. By utilizing similarity and dissimilarity features, each of these cluster is constructed. Therefore, data points with comparable qualities or attributes are combined into the same categories and data points with very divergent attributes or properties are separated into various groups. Clustering is a common technique for data analysis as an unsupervised learning technique (Marne et al., 2020). The first stage in the operation of K-Means algorithm is Elbow Method which is used in determining the number of clusters in a dataset. 'K' is used in denoting the number of clusters. Secondly, the Euclidean distance is utilized to find the nearest data points to the respective centroids after 'K' centroids is randomly initializes in the feature space, which is the algorithm's unique characteristic. The process of randomly adjusting the centroids are then repeated until no more data points are reassigned to any centroid. Finally, when the clusters are identified, the procedure is completed. The only disadvantage of K-Means is incorrect cluster formation caused by Random Initialization Trap. 'K-Means++' is utilized as the random initializer to address this issue (Thomas et al., 2020).

2.1.11 Gaussian mixture algorithm

The Gaussian Mixture technique is a popular approach in unsupervised learning, often known as a soft clustering technique. The probabilities for different forms of clustered data is calculated utilizing this technique and implemented using expectation maximization (Kar et al., 2025).

2.1.12 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is sequentially for building decision trees which is a highly efficient and strong ensemble strategy. In predictive modeling, XGBoost is well-known for its speed and performance, especially for problems such as BC prediction. By focusing on hard-to-classify situations, XGBoost reduces mistakes (Kar et al., 2025).

2.2 Related Works

This section provides an overview of the state-of-the-art literature in five categories that capture the major methodological directions in the field: classical machine learning for feature based classification, CNN and deep feature extraction, transfer learning and domain adaptation, multi-modal and hybrid architectures, and explainability with clinical interpretability. A structure of this kind allows for a more analytical assessment of progress and gaps remaining than a mere chronological survey.

2.2.1 Machine Learning Approaches

Before the era of DL, BC classification was performed using traditional ML algorithms. These algorithms are important benchmarks that are still cited today. SVM was one of the first and most widely used classifiers for this task. A linear programming-based SVM can achieve a classification accuracy of more than 97% on the Wisconsin Diagnostic Breast Cancer (WDBC) data set. This was done by using cytological features extracted from fine-needle aspirate images and SVM was found to be a strong baseline. The SVM optimisation objective is given by:

$$\min_{w,b,\xi} \tfrac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad s.t. \quad y_i(w\cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \qquad (7)$$

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

w is the weight vector defining the separating hyperplane, b is the bias term, $\xi_i \geq 0$ are slack variables permitting soft-margin classification, $C > 0$ is the regularisation parameter that controls the bias, variance trade-off, and $y_i \in \{-1,+1\}$ is the class label.

For non-linearly separable data, a kernel function $k(x_i, x_j)$ maps the inputs to a higher-dimensional feature space. The most commonly used kernel in breast cancer classification is the Radial Basis Function (RBF) kernel. The mathematical expression of the model is:

$$\kappa(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0 \qquad (8)$$

In ensemble methods, particularly with RF and gradient boosting variations, these techniques became strong alternatives to single classifiers like SVM. A RF classifier collects its predictions over DT, the final class label is obtained by majority vote, this can be expressed as:

$$\hat{y} = mode\{ h_t(x) \}_{t=1}^{T} \qquad (9)$$

$h_t(x)$ is the prediction of the t-th tree. Random Forests was applied to mammographic mass data, demonstrating resistance to overfitting in small-sample settings. The out-of-bag (OOB) error estimate, a built-in cross-validation mechanism can be expressed as:

$$OOB\ Error = (1/N) \Sigma_i \mathbb{1}[\hat{y}_i^{OOB} \neq y_i] \qquad (10)$$

While these ML methods provided good interpretability and computational efficiency, they are depended on handcrafted features extracted using statistical descriptors and texture operators such as local binary patterns or wavelet transforms. This reliance limited performance due to the constraints of the feature engineering process (Litjens et al., 2017). This limitation led to a shift towards representation learning using deep neural networks.

2.2.2 Convolutional Neural Networks and Deep Feature Extraction

The use of deep CNNs changed the breast cancer images analysis, allowing an end-to-end learning of hierarchical spatial features from the pixel level. A convolutional layer computes feature maps by convolving an input tensor X with a learned filter bank W, which is illustrated as:

$$F^{(l)(j)} = \sigma(\Sigma_i W^{(l)(ij)} * X^{(l-1)(i)} + b^{(l)(j)}) \qquad (11)$$

where * denotes 2-D convolution, $\sigma(\cdot)$ is a non-linear activation function (typically ReLU: $\sigma(z) = max(0, z)$), $W^{(l)(ij)}$ is the filter connecting input channel i to output channel j in layer l, and $b^{(l)}$(j) is the corresponding bias. The Rectified Linear Unit (ReLU) is illustrated as:

$$ReLU(z) = max(0, z) \qquad (12)$$

Spanhol et al. (2016) proposed the BreaKHis dataset for histopathological image analysis and compared different CNN architectures for magnification-dependent binary classification (benign vs malignant). The cross-entropy loss function minimised during training is mathematically expressed as:

$$L = -(1/N) \Sigma_i [y_i \log(\hat{p_i}) + (1-y_i)\log(1-\hat{p_i})] \qquad (13)$$

where N is the number of training samples, $y_i \in \{0,1\}$ is the ground-truth label, and $\hat{p_i} = \sigma(f\theta(x_i))$ is the predicted malignancy probability from a sigmoid output, with $f\theta$ denoting the network parameterised by θ.

Their results showed that performance of CNNs decreased significantly for higher magnification factors and that scale-aware training strategies are of great importance. Microcalcification detection is one of the main early signs of ductal carcinoma in situ and is particularly challenging due to their small size and low contrast. This was demonstrated by Dhungel et al. (2017) using a multi-stage pipeline of CNN-based candidate detection and a conditional random field for contextual refinement, achieving sensitivity comparable to expert radiologists. Al-masni et al. (2018) also used the You Only Look Once (YOLO) architecture to simultaneously detect and classify breast masses in full-field digital mammograms, that shows the viability of single-stage object detection frameworks for clinical screening tasks. Taken together, the studies show that the selection of CNN architecture, the size of the training data, and the task formulation are strongly correlated predictors of diagnostic performance.

$$L_YOLO = \lambda_coord\ \Sigma[(x-\hat{x})^2+(y-\hat{y})^2] + \Sigma[(\sqrt{w}-\sqrt{\hat{w}})^2+(\sqrt{h}-\sqrt{\hat{h}})^2] + \Sigma(C-\hat{C})^2 + \Sigma_classes(p-\hat{p})^2 \qquad (14)$$

2.2.3 Transfer Learning and Domain Adaptation

One of the common problems in medical imaging is the lack of large and well-annotated datasets. This limitation makes it impossible to train deep networks from scratch. A powerful strategy to overcome this bottleneck has been transfer learning, where models pre-trained on large corpora of natural images are fine-tuned on medical data. The fine-tunning objective minimises:

$$\theta^* = argmin_\theta\ L_target(f\theta(x), y) + \lambda\|\theta - \theta_pretrained\|^2 \qquad (15)$$

where θ_pretrained are weights from the source-domain pre-training stage, λ controls the degree of regularisation towards the pre-trained initialisation, and L_target is the task loss on the target medical dataset.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

A landmark comparative study by Tajbakhsh et al. (2016) demonstrated the superiority of fine-tuned CNN models over training networks from scratch and shallow classifiers for several medical imaging tasks including detection of breast lesions . Crucially, the authors demonstrated that even partial fine-tuning of the pre-trained layers resulted in significant improvements compared to random initialization, validating the transferability of low- and mid-level visual features across imaging domains.

Nawaz et al. (2018) fine-tuned a VGG-19 network on a multi-class histopathology dataset, reaching state-of-the-art results with less than 5,000 labelled training instances. Shen *et al.* (2019) proposed a two-stage transfer learning strategy where a CNN was first pre-trained on a large mammographic dataset and subsequently fine-tuned for mass classification, achieving area under the ROC curve (AUC) values above 0.90 on held-out test sets. We show that domain-specific pre-training is an effective intermediate step between ImageNet initialisation and target-domain fine-tuning. However, in domain shift, the statistical difference between source and target data distributions remains a cause of performance degradation which transfer learning by itself does not fully solve. Guan and Liu (2021) aligned the feature distributions of mammographic datasets from different imaging systems by means of a domain adversarial neural network, leading to an increased cross-site generalisation. This body of work identifies a critical gap in the literature, as most published performance metrics are based on single-site assessments and the external validity of published models is rarely rigorously evaluated.

2.2.4 Multi-Modal and Hybrid Architectures

Breast cancer diagnosis inherently requires multiple modalities including mammography, ultrasound, MRI, pathology slides, and genomic profiles. A late-fusion multi-modal classifier combines modality-specific predictions:

$$\hat{y} = g(\ \alpha_1 \cdot f_1(x_1) + \alpha_2 \cdot f_2(x_2) + \ldots + \alpha_m \cdot f_m(x_m)\ ) \qquad (16)$$

where $f_k(x_k)$ denotes the output of the k-th modality-specific encoder and α_k are learnable fusion weights satisfying $\Sigma_k \alpha_k = 1$.

A growing body of literature has explored architectures that exploit heterogeneous data streams to improve diagnostic accuracy beyond the capabilities of any single modality. Antropova et al. (2017) showed that a late-fusion strategy of radiomic features from dynamic contrast-enhanced MRI and deep CNN features outperformed each modality alone in malignancy prediction with AUC improvements of approximately three to five percentage points in prospective evaluation. Yap et al. (2018) benchmarked three fully convolutional network architectures for lesion detection, providing a systematic comparison of sensitivity-precision trade-offs that is often missing in single-architecture studies. Their work showed that the ensemble combination of the predictions of several architectures always beat any individual model. In terms of genomic data integration , Cheerla and Gevaert (2019) proposed a deep multi-modal model that integrates gene expression, miRNA and clinical variables to classify breast cancer subtypes and predict survival, achieving concordance index values better than single-modality baselines. The present findings suggest that multi-modal fusion is a very promising but still largely unexplored frontier, with most published studies still limited to single imaging modalities. Hybrid architectures that combine CNN-based spatial feature extraction with recurrent neural network or transformer self-attention mechanisms have also been studied. Yala et al. (2019) proposed a hybrid CNN-clinical model that integrated mammographic image features and patient risk factors to generate individualized short-term breast cancer risk scores and outperformed significantly the widely used Tyrer-Cuzick clinical risk model. This work is interesting because it addresses clinical risk stratification rather than just binary malignancy classification, thus extending the applicability of DL beyond the lesion level.

2.2.5 Explainability, Fairness, and Clinical Interpretability

A common criticism of deep learning models for use in the clinic is that they are black boxes, such that high predictive accuracy does not necessarily provide the mechanistic transparency required for clinical trust and regulatory acceptance. Therefore, there is a growing interest in the explainability of ML and DL models for breast cancer diagnosis. Selvaraju et al. (2017) proposed Grad-CAM, a gradient-based visualisation technique to localise discriminative regions in an image that contribute to a model's prediction. For target class *c,* the importance weight of the *k-th* feature map in layer l is:

$$\alpha^c_k = (1/Z)\ \Sigma_i\ \Sigma_j\ \partial y^c / \partial A^{(l)}_{(k,i,j)} \qquad (17)$$

where $Z = H \times W$ is the spatial size of the feature map, and $A^{(l)}_{(k,i,j)}$ is the activation at position (i, j) of the k-th channel. The resulting localisation map is given in the below equation:

$$L^c_Grad\text{-}CAM = ReLU(\ \Sigma_k\ \alpha^c_k \cdot A_k\ ) \qquad (18)$$

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)
http://.doi.org/10.46792/fuoyejet.v11i2.1550 engineering.fuoye.edu.ng/journal

Model-agnostic explanation frameworks including LIME (Ribeiro *et al.*, 2016) and SHAP (Lundberg & Lee, 2017) have been applied to tabular breast cancer datasets. The SHAP value for feature i is grounded in cooperative game theory:

$$\varphi_i(f,x) = \Sigma_{S \subseteq F\{i\}} |S|!(|F|-|S|-1)!/|F|! \cdot [f(S\cup\{i\})-f(S)] \quad (19)$$

Where F is the full feature set, S ranges over all subsets excluding feature i, $f(S)$ is the model output using only features in S, and the summation weights ensure that all features collectively account for the total prediction.These saliency maps have been validated with radiologist annotations in several subsequent studies reporting reasonable spatial agreement for mass detection tasks but weaker agreement for more subtle findings such as microcalcifications (Singh et al., 2022). Tabular breast cancer data has been used to quantify feature importance at the individual prediction level using model-agnostic explanation frameworks such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). Lundberg and Lee (2017) show that SHAP values are more consistent and theoretically sound feature attributions than competing methods for a range of ML models, including gradient boosted trees trained on the WDBC data set. However, Rudin (2019) makes a compelling case that post-hoc explanation of black-box models can never be a perfect substitute for inherently interpretable architectures. Issues of algorithmic fairness are increasingly pressing. Demographic parity requires that the positive prediction rate is equal across subgroups G:

$$P(\hat{y}=1 \mid G=g) = P(\hat{y}=1 \mid G=g') \quad \forall g,g' \in \mathcal{G} \quad (20)$$

Equalised odds further requires equal true positive rates (TPR) and false positive rates (FPR) across groups:

$$P(\hat{y}=1 \mid Y=y, G=g) = P(\hat{y}=1 \mid Y=y, G=g') \quad \forall y \in \{0,1\}, \ \forall g,g' \in \mathcal{G} \quad (21)$$

This perspective has inspired prototype and exemplar models that map internal representations to clinically meaningful constructs. However, these models are in the nascent stage of development for breast imaging applications.Questions of demographic bias and algorithmic fairness which is less pressing, but less often discussed. Larson et al. (2021) called for rigorous regulatory frameworks for the evaluation of AI diagnostic models, while Seyyed-Kalantari et al. (2021) performed a systematic audit and discovered significant performance gaps for under-represented demographic subgroups in chest imaging AI, a pattern with direct implications for breast cancer screening tools. The results show a major deficiency in most published breast cancer ML studies, which do not report subgroup performance analyses, thereby impeding equitable clinical deployment. Given the disproportionate burden of late-stage breast cancer in underserved populations, bias mitigation and fairness should be routine components of model evaluation.

3. METHODOLOGY

The current study make use of a systematic literature review (SLR) technique for identifying, selecting, and analyzing published literature on the use of ML and DL algorithms in diagnosing and classifying BC. This method was selected to ensure the comprehensive nature, repeatability, and rigors of the evidence synthesis process. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed while implementing the systematic literature review to uphold transparency and mitigate, correct unfair slants.The following research questions were used in guiding the review of the literature: the type of algorithms from ML and DL that have been utilized for breast cancer recognition and categorization, what was is the type of classification performance measures used within these algorithms is it recall, F1-score, accuracy, precision,, and AUC, What are the leading databases used for BC research in the realm of ML/DL and the current trends within this field? A planned, ordered and structured investigation was done in a number of database sources in order to discovery current peer-reviewed articles published between 2020 and 2025. The selected period of time is well founded by the fact that it is quite recent and thus indicate the fast development of ML and DL techniques and applications utilized for BC diagnostics.

3.1 Approach for Analysis and Synthesis

Putting into consideration the discrepancies in datasets, algorithm settings, and measures used across the selected literature studies, it became clear that a narrative synthesis methodology should be applied instead of quantitative meta-analysis. The analysis has been conducted across the following narrative lines: algorithm categories comparison: accuracy and other measures acquired by application of traditional machine learning classifiers (SVM, KNN, NB, logistic regression, DT), as well as by using group approaches such as RF, XGBoost, AdaBoost and DL algorithms like ANN, CNN, MLP, RNN. Also, dataset and validation Type analysis: The distribution of commonly used

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)
http://.doi.org/10.46792/fuoyejet.v11i2.1550 engineering.fuoye.edu.ng/journal

benchmark datasets mainly, the WBCD and WDBC, together with the types of validation used in the literature like k-fold cross-validation, train-test splits. Lastly, discovery of trends and gaps: detection of rising trends in the field such as modification of hybrid models, optimization of features, class imbalance, and the use of Explainable Artificial Intelligence (XAI).

4. DISCUSSION

In this section, factors explaining the variability in the performance of the different models are explored, along with a discussion on the suitability of different ML methods under various conditions, and the implications of using these ML methods for breast cancer diagnostics in a clinical setting. A comparative discussion on these models was necessary since the data and figures used in these studies are incomparable as they are extracted from different datasets with different preprocessing and validation techniques and feature selection stages. Table 1 summarizes the type of algorithm used in each study along with the highest reported performance evaluation metric. Table 2 discusses the performance of each algorithm in different scenarios, while Table 3 provides a comparison between all the studies reviewed.

Table 1: Algorithm Category and Performance

Algorithm Category	Top Performers	Notable Metrics
Statistical/Linear	Logistic Regression (LR)	99.26% accuracy (Anyachebelu et al., 2023); 98% (Khan et al., 2022)
Kernel-Based	SVM	99% accuracy (Olorunshola et al., 2025); 97.28% (Mustapha et al., 2022)
Tree / Rule-Based	J-48, Decision Tree, CN2	AUC = 0.922 for J-48 (Shanbehzadeh et al., 2022); CN2 AUC = 0.98 (Devi et al., 2023)
Instance-Based	KStar, KNN, Lazy IBK	98% accuracy – KNN (Singh, 2020); 98% – Lazy IBK (Alshammari & Mezher, 2020)
Probabilistic	NB, Bayes Net	94.7% – NB (Mehak et al., 2025); 97.1% – Bayes Net (Stephen & Adekoya, 2022)
Ensemble / Boosting	XGBoost, RF, AdaBoost	98.25% – XGBoost+FS (Ismaiel, 2024); 99.61% AUC (Prastyo et al., 2020)
Hybrid / Optimization	LR+Ant Search, MLP+RF, NN+Extra Trees	92.98% accuracy (Singh et al., 2025); 99.41% accuracy (Sharma et al., 2022)
Deep Learning / Neural	MLP, ANN, CNN, RNN	AUC-ROC = 0.9959 – MLP (Devi et al., 2023); 99.41% – NN+Extra Trees (Sharma et al., 2022)

4.1 Dominance and Limitations of Traditional Supervised Learning

Another frequently sighted observation from the studies under review is the generally high performance of the SVM classifier. For example, SVM was used as a benchmark classifier in Rashid et al. (2022), Qodri (2025), Shubham and Kamalraj (2022), and Olorunshola et al. (2025), providing amongst the highest accuracies, ranging from 96% to 99%, under varying experimental circumstances. SVM models can maintain an edge in classification tasks due to their ability to find the most optimally separated hyperplanes in feature spaces, while the 'maximum margin' principle helps to prevent overfitting when performing breast cancer prediction using moderate-sized datasets. Also, SVM can be applied with linear or non-linear kernel functions in accordance with the data distribution. The SVM classifier's proficiency in handling multi-dimensional cytological data is a significant advantage compared to other simple classifiers commonly used in clinical prediction studies.

But, it is crucial to examine the conditions under which SVM outperforms other algorithms. First, Olorunshola et al. (2025) reported their SVM results with 99% accuracy on the WBCD, a well-structured, balanced, low-dimensional dataset consisting of tabular data from fine needle aspirate (FNA) biopsies. But, when the identical class of SVM models is applied to other clinical datasets that are more diverse, feature higher dimensions, or exhibit pronounced class imbalances—such as the mammographic feature dataset utilized by Baloyi et al. (2025), or the NHS Abuja clinical dataset used by Ishola et al. (2025)—the performance metrics decline drastically.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

Similarly, in the case of logistic regression (LR), Anyachebelu et al. (2023) achieved a verified test accuracy of 99.26% using an optimised LR model on the WDBC dataset, and it outperformed all ensemble and kernel-based models in their study. Khan et al. (2022) also achieved 98% accuracy using LR. Yet LR accuracy drastically drops down to 37.43% in Mehak et al.'s study (2025) on the very same WBCD dataset. This is not an issue with LR itself but rather a consequence of inadequate experimental design: in Mehak et al.'s study, critical preprocessing steps, particularly feature selection, were omitted, and most likely the class distribution was not properly balanced.

Regarding the ensemble classifiers, XGBoost demonstrated superior performance compared to other algorithms in the cited papers (Ismaiel, 2024; Prastyo et al., 2020). Ismaiel (2024) reported an accuracy of 98.25% for XGBoost when feature selection methods were applied, while Prastyo et al. (2020) reported an AUC of 99.61%. Likewise, Random Forest (RF) exhibited exceptional performance in several cited papers (Tachicart et al., 2025; Stephen & Adekoya, 2022). In summary, different studies applied diverse data validation methods and feature selection techniques, which may have influenced the results. Ultimately, the choice of classification method depends on the dataset and the prioritization of accuracy, sensitivity, or specificity in a clinical setting.

4.1.1 Context-Dependency of Algorithm Performance: Dataset, Validation, and Environment

One interesting point of discussion drawn from the existing studies is an analysis of why the same algorithm performed differently across studies. There are three main reasons for differences that can be identified from the studies reviewed. Firstly, the influence of dataset properties is a critical moderator. Breast Cancer Wisconsin Datasets WBCD and WDBC, which are used collectively in over 80% of total studies, depict only a limited perspective of the real clinical landscape. Both are derived from FNA biopsy cytology, with small sample sizes (569–699), moderate level of class imbalance rather than extreme imbalance, and no missing values, which are not uncommon properties in real clinical datasets. Some of these studies are conducted by using real clinical datasets, such as Shanbehzadeh et al. (2022), where 1,052 patient samples were used; Arravalli et al. (2025), where data from the University of Calabar Teaching Hospital were used; and Ishola et al. (2025), which used data from NHS Abuja. These studies have revealed lower accuracy compared to those conducted on WBCD datasets, even though the same algorithms were used, which is reasonable given the relatively more complex, heterogeneous, and missing-value-included nature of real clinical data.

Secondly, the validation method used affects the reported model performance. Studies that relied on a single data split (e.g., 70% training and 30% testing) without a validation set to tune hyperparameters reported more optimistic accuracy than those employing proper cross-validation. For instance, a comparison conducted by Olorunshola et al. (2025) showed that SVM accuracy dropped from 99% using a 70% data split in a Jupyter Notebook to 97.7% with WEKA's 70% split and 98.1% with WEKA's 10-fold cross-validation. The drastically different accuracies were based on the same dataset and algorithm. For studies reporting extremely high accuracy (greater than 98%), the lack of cross-validation should be considered a limitation. Thirdly, the different implementations are yet another factor to consider. WEKA uses certain default hyperparameter settings, feature handling, and normalisation steps that are not replicated in the Python implementations used in scikit-learn or XGBoost. Thus, a comparison of algorithm performance that draws results from WEKA and Python-based implementations without accounting for these differences is inherently a comparison of somewhat different conditions. Table 2 provides an overview of this cross-contextual performance variation.

Table 2: Context-Dependent Performance Variability Across Algorithms, Datasets, and Validation Methods

Algorithm	Dataset	Validation Method	Reported Accuracy	Key Contextual Factor
SVM	WBCD	70% split (Jupyter)	99.0%	Optimised hyperparameters; clean benchmark data
SVM	WBCD	10-fold CV (WEKA)	98.1%	Cross-validation reduces optimism bias vs. 70%-split
SVM	WBCD	70%-split	97.7%	Same

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

		(WEKA)		tool/data but holdout split — lower than 10-fold CV
SVM	BCCD (clinical)	Train-test split	~96–97%	Greater feature heterogeneity
XGBoost	WDBC	CV + feature selection	98.25%	InfoGain feature selection; cross-validated
XGBoost	NHS Abuja dataset	Train-test split	88.0%	Real-world clinical data; higher complexity
Logistic Regression	WDBC	Optimised (10-fold)	99.26%	Feature selection + hyperparameter tuning
Logistic Regression	WBCD	Unspecified split	37.43%	Poor preprocessing; no feature engineering
J-48 Decision Tree	Clinical (1,052 pts)	Unspecified	AUC = 0.922	Real clinical dataset; best among 6 classifiers
RF + XAI	Calabar Teaching Hospital	Train-test split	F1 = 84%	Real clinical data; lower than WBCD benchmarks
NN + Extra Trees (Ensemble)	WBCD	Unspecified	99.41%	Ensemble with high model complexity on small dataset

4.2 Feature Selection, Dimensionality Reduction, and Optimisation

The factor that emerged most frequently across the papers was the impact of data quality on model performance, with data preprocessing and feature engineering often affecting results more than the choice of machine learning algorithm itself. This finding shakes some of the conventional assumptions in breast cancer ML research and suggests data curation to be a primary driver for classification performance rather than the specific algorithm employed.

Research that systematically selects features demonstrates that this process adds value regardless of the class of algorithm used. For example, Gupta and Ahmed (2025) show that when Logistic Regression—a relatively simple algorithm—was trained on biomarkers filtered through Weka's InfoGain and Ranker algorithms, it produced an accuracy of 98.25%. This result is extremely significant because it shows that the slight increases in accuracy attributed to sophisticated deep learning models may be the result of careful feature selection, rather than the models themselves. Baloyi et al. (2025) provided the most rigorous examination of dimensionality reduction effects in this review, reporting that Correlation-based Feature Selection (CFS) successfully reduced WBCD features from 30 to 11 while Neural Network accuracy remained at 97.2%, and noted the improvement in model training time and computational resources. Hoque et al. (2024) determined that the reduced set of top 5–11 nuclear features (concave points and perimeter were specifically highlighted) contained the discriminative structure necessary from the WBCD dataset for XGBoost to achieve 94.74% accuracy and 95.24% recall 10-fold cross-validation reduced the risk of overfitting and ensured the robustness of their findings. Din et al. (2025) further showed that in order to achieve optimum performance, feature selections were needed as the SMO algorithm produced the highest accuracy of 96.996% when all relevant features were retained after feature

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

selection. These similar results in different studies strongly recommend feature selection or dimensionality reduction not as an optimisation step, but as a mandatory step in the machine learning process for breast cancer. Any models tested without feature engineering should be considered as producing lower performance results, rather than being used as a clinical benchmark. In addition to reducing dimensionality, feature selection can implicitly regularise the model by selecting important features and removing those that are redundant or noisy, which can lead to overfitting. Olorunshola et al. (2025) corroborated these results, finding 98.1% accuracy for SVM with WEKA's 10-fold cross-validation. They tested SVM with the entire WBCD feature set. Hoque *et al.* (2024), who reported 94.74% accuracy and 95.24% recall with XGBoost on the WBCD, found that selecting 5–11 top nuclear features was adequate to capture the important structure of the WBCD while reducing computational costs. This dual consideration of feature engineering elevates performance and generalisability, and it is perhaps the most beneficial endeavour in developing ML breast cancer classification models.

4.3 Interpretability, Explainability, and Clinical Trust

The explainability of the models is one of the most impactful factors from a clinical perspective yet was barely discussed in any of the papers. The utilizaton of ML models for BC diagnosis in a clinical setting would not only need to be accurate but also have interpretable reasons behind the diagnosis that clinicians, patients, and authorities can verify and trust. The algorithms discussed above cover a large spectrum of interpretability. Models such as Logistic Regression, DT, NB, or rule-based classifiers like CN2 are highly interpretable. Predictions made by these models can be directly traced to the underlying coefficients, decisions, or posterior probabilities, making them intrinsically transparent without the need for additional explanation. Importantly, the above transparent models were not trounced by more complex models in the tested literature: optimised LR achieved 99.26% accuracy (Anyachebelu et al., 2023), and CN2 achieved 0.98 AUC (Devi et al., 2023), which is above the reported performance of ensemble and deep learning models in the same studies.

In contrast, the top performing algorithms in terms of metrics, XGBoost, RF, deep neural networks, and hybrid ensemble models, are black-box models. These models outperform the other models in predictability but not in interpretability, which poses a dilemma when used in a clinical setting. If a clinician is using a black-box model that has predicted a malignancy, there is no way for the clinician to assess the reasoning behind the model's prediction, understand where the model may have failed, or explain the recommendation to the patient. Post-hoc explainability methods offer interpretability while retaining the high performance of complex models. SHapley Additive exPlanations (SHAP) calculates the contribution of each feature to a particular prediction using game theory, allowing for feature-level attribution in individual predictions. Local Interpretable Model-agnostic Explanations (LIME) generates linear approximations that are locally faithful to the complex model, enabling the interpretation of individual predictions irrespective of underlying model class. Attention mechanisms in deep learning models indicate the areas or features that influenced the model's decision, providing built-in interpretability for neural network models utilized in imaging applications.

Arravalli et al. (2025) represents the most comprehensive integration of XAI in the reviewed literature, combining Random Forest classification (achieving 84% F1-score) with XAI-based feature attribution on a real clinical dataset from the University of Calabar Teaching Hospital. They show that interpretability along with good predictive performance can coexist in ML and I suggest future ML clinical studies should follow their example. Their relatively low F1-score of 84% compared to studies done on WBCD data both illuminates the performance challenges of real clinical data and the need to supplement accuracy metrics with explainability.On top of that, from a legal , there are increasing regulations that mandate explainability for machine learning-driven diagnostic tools. For instance, the Food and Drug Administration (FDA) in the U.S. has issued guidelines detailing expectations for AI/ML applications software (Artificial Intelligence/Machine Learning) that function as medical devices, and the European Union's AI Act designates clinical decision support systems as high-risk artificial intelligence systems subject to conformity assessments that include explainability requirements. These regulatory factors make XAI not only a field of interest but a critical criterion for the clinical applicability of ML models. The scant attention to XAI in the studies under review, barring Arravalli et al. (2025), reflects a stark disjunction with the clinical potential of these machine learning models.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

4.4 Transition Toward Deep Learning and Hybrid Models

A clear trend in the literature is that the models have evolved from one-algorithm-for-tabular-data to hybrid models comprising multiple components and deep learning models for more complex, multi-modal, and high-dimensional data. This reflects the increasing availability of imaging data in BC diagnosis and the lack of a single traditional algorithm that performs adequately well in all cases.

In distinguishing between benign and malignant cases using standardised datasets, deep learning models have outperformed traditional models. For example, Devi et al. (2023) found that their Multilayer Perceptron (MLP) deep learning model had the highest overall performance in their comparative study, achieving an AUC-ROC of 0.9959, a precision of 96.57%, and an accuracy of 96.49% on the WDBC dataset, surpassing all traditional classifiers including SVM, Gradient Boosting, and CN2. Sharma et al. (2022) achieved the single highest reported accuracy in this review (99.41%) through an ensemble combining Neural Networks with Extra Trees classifiers, further demonstrating the performance ceiling reachable through neural-ensemble integration.

Hybrid modelling approaches represent a particularly promising direction. Singh et al. (2025) demonstrated that integrating Logistic Regression with Ant Colony Search Optimisation, a metaheuristic optimisation technique that explores the feature and hyperparameter space, substantially reduced false negatives while achieving 92.98% accuracy on the WBCD dataset. Sharma et al. (2022) achieved the single highest reported accuracy in this review (99.41%) through an ensemble combining Neural Networks with Extra Trees classifiers, yielding a specificity of 1.0 and MCC of 98.76 which demonstrates that principled combination of complementary algorithmic strengths can produce results substantially superior to any single model. These hybrid results are significant because they demonstrate that performance improvements are achievable not only through model complexity but through principled integration of diverse learning paradigms which is a strategy more likely to translate to sustained real-world performance than simply adding neural network layers.

However, the transition to deep learning and hybrid models must be approached with methodological caution, particularly in the context of small clinical datasets. Standard deep learning architectures require large training datasets to learn generalisable representations; on datasets of the size typical in breast cancer ML research (several hundred to low thousands of instances), complex neural architectures are at substantially elevated risk of overfitting. The consistently strong performance of deep learning models reported in this review almost exclusively on WBCD dataset, may therefore reflect overfitting to dataset-specific patterns rather than genuine generalisation. This concern is directly addressed in Section 4.6.

4.5 Class Imbalance: Clinical Implications and Mitigation Strategies

Class imbalance constitutes one of the most clinically consequential methodological challenges in ML-based breast cancer classification, yet it received inconsistent and often insufficient treatment in the reviewed literature. Understanding and addressing class imbalance is not merely a statistical concern , it is a patient safety imperative with direct implications for diagnostic outcomes. In the context of breast cancer screening, the asymmetric cost structure of classification errors must drive model development priorities. A false negative which is a malignant case classified as benign results in a missed diagnosis, delayed treatment, and potentially fatal disease progression. A false positive (benign case predicted as malignant) can lead to unnecessary stress, additional tests, and costs (and in extreme cases, unnecessary treatment), but rarely to irreversible harm. This means that sensitivity (recall for the malignant class) and the F1-score are far more clinically relevant metrics compared to accuracy (which measures overall correctness), particularly when benign cases form the majority class in the dataset. For instance, a classifier that always predicts “benign” in a dataset with 90% benign cases would achieve 90% accuracy but would have no clinical value, which is not apparent if one looks only at accuracy. Explicitly reporting the negative effects of class imbalance on clinically relevant performance measures, Tachicart et al. (2025) found their RF and ANN models, while having strong overall accuracy (84.72% and similar), struggled to detect positive (malignant) cases (ie: false negatives) in class-imbalanced medical datasets. This is concerning — their models performed well by aggregate measures but missed the most important cases. Addressing this issue directly, Singh et al. (2025) proposed a hybrid model with class balancing that significantly reduced false negatives and achieved 92.98% accuracy.

Various techniques that can mitigate class imbalance are well supported in the general ML literature, but

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

are relatively underutilized in the studies surveyed here. Oversampling methods like SMOTE, or Synthetic Minority Oversampling Technique, create artificial instances of minority-class data points, eliminating bias against the majority class, and importantly, without sacrificing information. Undersampling the majority class is a simpler alternative but loses training information. Cost-sensitive learning increases misclassification costs for the minority (malignant) class, embedding the medical cost differential in the model's objective function. Ensembles-based resampling of RF and AdaBoost like BalancedRandomForest and EasyEnsemble build the model by including re-sampling as part of the ensemble creation itself.

A takeaway from this analysis is for future breast cancer classification studies to report sensitivity, specificity, F1-score, and AUC-ROC as primary metrics along with—or instead of—accuracy. Accuracy is only an acceptable metric if class prevalence is roughly similar, which cannot be relied on for clinical breast cancer datasets. Future reviews or meta-analyses of clinical applicability of breast cancer ML tools should attempt to exclude single study accuracy reporting as the apparent excellent performance in studies will not be generalisable to clinical practice.

4.6 Overfitting, Generalisability, and the Importance of Validation

The term 'overfitting' refers to a model that has 'learned' the particular noise and peculiar characteristics of a particular training dataset, rather than learning diagnostic features which would be more generalizable, the major danger for an ML technique. While a standard problem with machine learning, in a clinical setting, the impacts of over-fitting are disastrous as it may appear that the model performs exceedingly well for the described parameters, but fails when presented with actual, novel patients, thereby misinforming clinical diagnosis.There are a variety of factors in the reviewed articles which would cause there to be an increased risk of over-fitting. The WBCD, used in over 80% of reviewed studies, contains only 569 instances which is a dataset size at which complex models with many parameters (deep neural networks, highly-tuned XGBoost models, large RF ensembles) are prone to memorising training examples rather than learning generalisable class boundaries. When models such as the NN + Extra Trees ensemble (Sharma et al., 2022) achieve 99.41% accuracy on this dataset, the question of whether this reflects genuine generalisation or dataset-specific overfitting is clinically critical and remains unanswered in the absence of external validation. The fact that equivalent algorithms applied to independent clinical datasets (Ishola et al., 2025; Arravalli et al., 2025) achieve substantially lower performance (84–88%) is consistent with overfitting to WBCD's specific statistical properties.

Validation is the primary safeguard against overfitting in reported metrics. K-fold cross-validation is a technique in which the dataset is partitioned into k subsets, with k rounds of training and evaluation each using a different held-out fold. This offers a more reliable estimation of generalization performance than just split-based approaches (hold-out and split/validation set) especially for smaller data sets. Stratified k-fold cross-validation also maintains the distribution of classes in each fold and it is thus recommended in the cases of imbalanced data sets. The papers in the survey that used 10-fold cross-validation (Ismaiel, 2024; Olorunshola et al., 2025; Hoque et al., 2024) all return slightly less accuracy but also more trustworthy values of this measure than using split-based methods and this again is evidence for optimistic bias. A final important aspect is the conduct of hyper-parameter tuning not on the test set, but on separate validation set, which is very rarely reported in the studies considered in this review. This brings forth the potential for a bias, as results of test-set performances may have been optimized (unknowingly) through implicit data leakage. Regularization methods prevent overfitting in algorithmic level. L2 regularization in Logistic Regression (Ridge) or regularization parameter C in SVM restrict the complexity of the model by penalizing large magnitude of coefficient, dropout regularization in neural networks randomizing disabling of neuron connections prevents their co-adaptation and therefore excessive reliance on combinations of features, max-depth parameter and min-samples-per-leaf parameter in decision trees and random forests are specific settings aimed at restricting the model complexity. XGBoost's L1/L2 regularization and also subsampling and column sampling features are specifically designed to prevent overfitting of the model whilst retaining high predictive power, the former partially explaining why this models shows high performance generalization in comparison to other complex models. Possibly the most compelling conclusion to come out of this review concerning over-fitting is the repeated finding that a simpler,

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

regularised model, with well chosen features can match or improve upon the performance of very complex ones. Both Gupta and Ahmed (2025) managed to reach 98.25% accuracy with Logistic Regression and InfoGain feature selection, matching the performance of XGBoost in other trials. Baloyi et al. (2025) revealed that neural network training on 11 selected features reached an accuracy of 97.2% whilst remaining efficient and producing performance only slightly inferior to the original model using 30 features. Both these examples concur with the principle of Occam's Razor in statistical learning whereby with equivalent prediction, a simpler model is to be favored as less likely to over-fit to specific data, and therefore more capable of accurately representing future, previously unseen clinical data. Future work should strive to not only focus on building increasingly complex models, but also to attempt to beat simple, well-trained baseline models and provide evidence that such complex models actually justify their performance.

4.7 Clinical Applicability: From Benchmark Performance to Real-World Deployment

The most damaging rift between research and practice is the enormous distance that exists between best performance on curated benchmarks and actual utility in a clinical diagnostic pipeline. This section addresses the obstacles that stand in the way of translating even the highest-performing ML algorithms to clinical practice, and outlines areas of research needed to bridge this divide. The benchmark to bedside gap is well exemplified by the dataset landscape of this review. WBCD is a 1991 data set from FNA biopsy cytology that is still the de facto evaluation benchmark in this space, in spite of over thirty years of dramatic advancements in clinical BC diagnosis. The dominant modalities in current practice now are digital mammography, automated breast ultrasound, MRI, and digital pathology whole-slide imaging, all distinct from FNA cytology. Machine learning algorithms trained and evaluated solely on WBCD tabular data features may be optimized on a diagnostic modality and data collection process that has become an increasingly fringe one in current practice, meaning that the reported high accuracies of 97-99% do not provide evidence that an algorithm is ready for clinical deployment; instead, it provides evidence that the algorithm is capable of excellent performance on a 1991 data set that has little resemblance to current clinical data. Real-world data have qualitatively distinct problems to curated benchmarks, such as missing data from incompleteness of clinical records, variable data from site- and device-specific acquisition and measurement variation, patient heterogeneity (race and comorbidity) and variable data collection protocols, and high class imbalance typical of actual prevalence in screening populations. None of these challenges is adequately simulated on WBCD.

A number of non-algorithmic barriers to clinical adoption are introduced as well, rarely explored in literature review: the requirement of the model to integrate within existing EMR, PACS and LIS requires wrapping of the ML models into API or SaMD standards to meet healthcare standards like HL7 FHIR, DICOM. Real-time performance limitations might necessitate specific hardware resources. Continued model monitoring is required in case the model performance decreases due to changes in patient population and clinical practice (dataset shift or concept drift), and need for continuous model retraining and versioning are also not part of the literature review.

Lastly, most studies were performed on single-modal data in tabular form. However BC diagnosis involves multi-modal data and is composed of mammography images, Ultrasound images, MRI images, histopathology information, genomics and patient information. ML architectures like combining CNN-based image feature extraction, and tabular ML with clinical parameters represent the latest development trends in clinical breast cancer AI, which are mostly missing from reviewed literature. Integration of CNN-based image analysis, which achieved strong results with mammography and histopathology (Adam et al., 2023; Nasser & Yusof, 2023), and structured clinical data ML is currently the highest research direction.

4.8 Comparative Effectiveness of Implementation Environments and Tools

The choice of implementation environment (either a graphical data mining software such as WEKA, or a programmatic language environment such as Python (Jupyter Notebook / scikit-learn)) has long been a underemphasized and confounding source of variability across studies that makes direct cross-study comparison difficult. This choice has multiple implications, as default parameter values, preprocessing techniques, the use of random state and availability of optimisation routines vary widely between platforms. It has been the workhorse of data mining for decades (Arora et al., 2024; Qodri, 2025; Stephen & Adekoya, 2022; Olorunshola et al., 2024, 2025; Alshammari & Mezher, 2020), due to its user-friendly graphical interface, standardized

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

algorithms, and its relatively easy accessibility for users who do not have substantial programming skills. Standardized implementations of commonly used algorithms such as J48 and KStar within WEKA make for cross-study replicability, and its built-in facilities for cross-validation offer ease of robust model evaluation. Its default hyperparameter settings are not universally optimal for every task, and custom preprocessing steps and features may be more easily expressed within a programmatic environment than on the graphical interface.

Python-based implementations (scikit-learn, XGBoost, TensorFlow, and PyTorch) give significantly more room to maneuver in custom optimization, creative feature engineering, and deep-learning model design, and permit relatively direct interfacing with XAI models such as SHAP and LIME. Of the cross-environment tests in this review, one was about as direct a comparison as could exist: Olorunshola et al. (2025) used identical algorithms on identical datasets in WEKA and the Jupyter Notebook; using 70%-holdout, SVM reached 99%-accuracy in Jupyter Notebook, but WEKA reported only 97.7% with the same 70%-split and 98.1% with 10-fold cross validation. These three values – the performance on the same algorithm on the same dataset under two platform/validation-method pairs – suggest that these two factors alone are not insignificant. Table 3 presents a comparative Analysis of ML Techniques for Prediction of BC.

Table 3. Comparative Analysis of ML Techniques in Predicting Breast Cancer

Author(s) & Year	Datasets Used	Algorithms Evaluated	Key Findings / Best Model
Arora *et al.* (2024)	WBCD (WEKA)	KStar, Decision Table, REPTree	KStar highest accuracy: 0.980 (Node-Caps); superior recall and F-Measure
Rashid *et al.* (2022)	WBCD & BCCD	SVM, NB, LR, DT, RF, k-NN	SVM outstanding; recommended for medical applications
Qodri (2025)	WDBC (WEKA)	NB, C4.5, SVM	SVM: 96.9957% accuracy — highest of three
Ismaiel (2024)	WDBC	RF, XGBoost, NB, KNN, DT, LR, SVM	XGBoost + feature selection: 98.25%; cross-validation: 97.89%
Tachicart *et al.* (2025)	Medical datasets	SVM, LR, RF, ANN	RF: 84.72% acc, 88.41% F-score; ANN: 83.33% precision. Both limited by class imbalance
Shanbehzadeh *et al.* (2022)	Clinical dataset (1,052 samples)	J-48, RF, MLP, NB, AdaBoost, LR	J-48 best: AUC = 0.922; NB lowest: 0.669
Harshitha & Priya (2023)	WBCD	SVM, Naïve Bayes	SVM best precision; recommended for cancer analysis
Din *et al.* (2025)	Mammographic images	NB, LR, SMO, DT	SMO + NB: 96.996% with all features; feature selection improves accuracy
Prastyo et al. (2020)	WBCD (5-fold CV)	KNN, SVM, GNB, RF, AdaBoost, GB, XGBoost, MLP	XGBoost best: 97.19% acc, 99.61% AUC, 96.99% F1
Shubham & Kamalraj	WBCD	KNN, SVM, DT	SVM: 97% accuracy

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://.doi.org/10.46792/fuoyejet.v11i2.1550 engineering.fuoye.edu.ng/journal

(2022)			
Singh (2020)	WBCD	KNN, SVM, LR, NB	KNN: 98% accuracy
Rawal (2020)	WBCD	SVM, C4.5	SVM: fewer false positives; high true-positive rates
Khan *et al.* (2022)	WBCD	Logistic Regression	LR: 98% accuracy
Anyachebelu *et al.* (2023)	WDBC	RF, SVM, KNN, LR	Optimised LR: 99.26%; lowest FPR (1) and FNR (4)
Devi *et al.* (2023)	WDBC (UCI)	KNN, GB, SVM, NB, SGD, CN2, Tree, MLP	MLP best overall: 96.49% acc, AUC-ROC = 0.9959; CN2 AUC = 0.98
Singh *et al.* (2025)	WBCD (Jupyter)	Hybrid LR+Ant Search, SVM, RF, KNN	Hybrid LR: 92.98% acc; significantly reduced false negatives
Baloyi *et al.* (2025)	Tertiary hospital dataset	10 classifiers incl. SVM, LR, NN	SVM & LR: 97.7% (full features); NN: 97.2% on reduced 11-feature set (CFS)
Arachchi *et al.* (2024)	WBCD	SVM, KNN, NB, LR, AdaBoost, DT	AdaBoost: 96% acc; LR: 96% ROC
Mehak *et al.* (2025)	WBCD	KNN, LR, DT, SVM, ANN, NB	NB: 94.7% best; KNN: 91.81%; LR: 37.43% (poor preprocessing)
Stephen & Adekoya (2022)	WBCD (WEKA)	RF, Random Tree, Bayes Net, NB, J48	RF: 97.1%
Hoque *et al.* (2024)	WBCD	XGBoost	XGBoost: 94.74% acc, 95.24% recall
Olorunshola *et al.* (2025)	WBCD (WEKA + Jupyter)	SVM, NB, RF, MLP, DT	SVM: 99% (Jupyter 70% split); 98.1% (WEKA 10-fold CV); 97.7% (WEKA 70%-split)
Alshammari & Mezher (2020)	WBCD & WDBC (WEKA)	10 classifiers incl. NB, LR, IBK, Kstar	Lazy IBK: 98% accuracy — best of ten
Arravalli *et al.* (2025)	Univ. of Calabar Teaching Hospital	RF, Stacked Ensemble + XAI	RF: 84% F1-score with XAI feature attribution
Ishola *et al.* (2025)	NHS Abuja clinical dataset	DT, KNN, XGBoost, LR	XGBoost: 88% acc — best; real-world dataset performance lower than WBCD benchmarks
Sharma et al. (2022)	WBCD	Ensemble (NN + Extra Trees)	99.41% acc, specificity = 1.0, MCC = 98.76

5.0 CONCLUSION AND RECOMMENDATION

Machine learning has shown promising results in BC prediction and aiding early diagnosis. The literature survey showed that conventional supervised learning algorithms such as SVM and LR are still dependable and clinically feasible as they consistently produce state-of-the-art results over benchmark datasets. Further, ensemble/hybrid approaches like XGBoost, RF outperformed other

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

http://.doi.org/10.46792/fuoyejet.v11i2.1550 engineering.fuoye.edu.ng/journal

models and reported the lowest false-negative rates along with Neural Network-based methods in some studies. DL also improved classification accuracy on image/text data along with feature extraction and representation learning.

However, there are still a few limitations to the reported works. Several studies use widely available and publicly used benchmark datasets such as WBCD and WDBC to develop their methods. Real-world clinical use cases are much more diverse and challenging than these publicly available datasets can portray. Also, most of the studies optimize for accuracy without focusing on the clinical interpretability/fairness/robustness/deployment complexity of models in real healthcare settings. Further, many ML models (especially deep learning and ensemble models) are "black-box" in nature and provide little to no insights into medical interpretations. Lastly, only a few studies tried to validate their proposed algorithms in real-time hospital setups and most of them report highly optimized models for predictive performance under experimental setups. For example, there is no consideration for class imbalance, model scalability and integration into electronic health record systems, high inference costs, or lack of clinical adoption.

To overcome some of these limitations, future work can be done to make ML models more interpretable by using XAI. Studies that implement their approaches in real clinical settings with patient-level data and model scalability should be conducted. Future work can also be done to report more clinically relevant metrics such as recall, sensitivity, and false-negative rates to truly improve diagnosis.

In conclusion, ML can be used to its full potential for diagnosing breast cancer. Nevertheless, balancing predictive accuracy with interpretability, fairness, robustness and real word application will be required to achieve reliable clinical adoption.

REFERENCES

Adam R., Dell'Aquila K., Hodges L., Maldjian T., and Duong T. Q. (2023). Deep learning applications to breast cancer detection by Magnetic Resonance Imaging: A literature review, *Breast Cancer Research*, 25(1),

Al-masni, M. A., Al-antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., Valarezo, E., Choi, M.-T., Han, S.-M., & Kim, T.-S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Computer Methods and Programs in Biomedicine, 157, Pp. 85–94. https://doi.org/10.1016/j.cmpb.2018.01.017

Alshammari M., and Mezher M. (2020). A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox, *International Journal of Advanced Computer Science and Applications*, 11(8), Pp. 224 - 229

Amethiya Y., Pipariya P., Patel S., and Shah M. (2022). Comparative analysis of breast cancer detection using machine learning and Biosensors, *Intelligent Medicine*, 2(2), Pp. 69–81,

Antropova, N., Huynh, B. Q., & Giger, M. L. (2017). A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical Physics, 44(10), 5162–5171. https://doi.org/10.1002/mp.12453

Anyachebelu K. T., Abdullahi M. U., Abdullahi M. (2023). Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prediction Wisconsin Breast Cancer, *International Journal of Engineering & Technology*, 7(4), Pp. 160– 166.

Arachchi H. A., Pathirana H. P., Samarasinghe A., Samarasinghe G. D. (2024). Comparative Performance Analysis of machine learning models for Breast Cancer Prediction, *8th SLAAI International Conference on Artificial Intelligence* (SLAAI-ICAI), IEEE

Arravalli T., Chadaga K., Muralikrishna H., Sampathila N., Cenitta D., Chadaga R. & Swathi K. S. (2025). Detection of breast cancer using machine learning and explainable artificial intelligence, *Scientific Reports*, 15(1), Pp. 1 -10

Arora N., Garg A. and Baggan V. (2024). Machine Learning Approaches for Breast Cancer Prediction: A WEKA-Based Performance Analysis, *IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, pp. 1-6,

Baloyi V., Mattiev J., Mokwena S. (2025). Importance of Data Preprocessing for Accurate and Effective Prediction of Breast Cancer: Evaluation of Model Performance in Novel Data, *Big Data Cogn. Comput.*, 9(1), pp. 1 - 20,

Centers for Disease Control and Prevention. (2024, September 16). About breast cancer in men. U.S. Department of Health and Human Services. https://www.cdc.gov/breast-cancer/about/men.html

Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics, 35(14), i446–i454. https://doi.org/10.1093/bioinformatics/btz342

Devi S., Ghanekar R. K., Pande J. A., Dumbre D., Chavan R., Gupta H. (2023) Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models, *Asian Pacific Journal of Cancer Prevention*, 25(1), Pp. 1077-1085,

Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Medical Image Analysis, 37, 114–128. https://doi.org/10.1016/j.media.2017.01.009

Din M. B., Daud P., Ismail N. A., Ismail N. L., Nadi F. (2025). A Comparative Study of Features Selections in Breast Cancer Classification Using Machine Learning Algorithms, *International Journal Of Academic Research In Business And Social Sciences*,15(2), pp. 956 – 967

Ebrahim M., Sedky A. A., and Mesbah S. (2023). Accuracy assessment of machine learning algorithms used to predict breast cancer, *Data*, 8(2), Pp. 35,

Elmore, J. G., Armstrong, K., Lehman, C. D., & Fletcher, S. W. (2005). Screening for breast cancer. JAMA, 293(10), 1245–1256. https://doi.org/10.1001/jama.293.10.1245

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.

This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

El-Shair Z. A., Sánchez-Pérez L. A., Rawashdeh S. A. (2020). Comparative Study of Machine Learning Algorithms using a Breast Cancer Dataset, *IEEE Xplore*, Pp. 500-508

Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering, 69(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3117407

Gupta S. and Ahmed F. (2025). A Comparative Performance Analysis of Machine Learning Models for Breast Cancer Prediction, *25th International Conference on Computer and Information Technology (ICCIT)*,

Harshitha R. and Priya S. M. (2023). Prediction of Breast Cancer Using Classification Algorithms, *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics*, Pp. 276-280

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Hoque R., Das S., Hoque M. and Haque E. (2024). Breast Cancer Classification using XGBoost, *World Journal of Advanced Research and Reviews*, 21(2), Pp. 1985–1994

Ishola O. B., Hammawa M. B., Ogwueleka F. N. & Owolabi O. (2025). A Comparative Evaluation Of Some Machine Learning Techniques For Predicting Breast Cancer, *Journal of Health Systems Research*, 7(3), Pp. 91-101

Ismaiel W. (2024). Analysis Performance of Machine Learning Classifiers for Breast Cancer Detection Based on Features Selection and Cross-Validation, *International Conference on Open Innovation and Digital Transformation (OIDT)*, Pp. 1-9

Kar S. K., Pandey A. and Gautam C. S. (2025). A Review of Machine Learning Techniques for Breast Cancer Prediction, *International Journal of Current Engineering and Technology*, 15(3), pp. 219-228

Khan M., Islam S., Sarkar S., Ayaz F. I., Kabir M., Tazin T., Albraikan A. A., & Almalki F. A. (2022). Machine Learning Based Comparative Analysis for Breast Cancer Prediction, *Journal of Healthcare Engineering*, Pp. 1 – 8,

Kirola M., Memoria M., Dumka A., and Joshi K., (2022). A Comprehensive Review Study on: Optimized Data Mining, Machine Learning and Deep Learning Techniques for Breast Cancer Prediction in Big Data Context, *Biomedical and Pharmacology Journal*, 15(1)

Larson, D. B., Harvey, H., Rubin, D. L., Irani, N., Justin, R. T., & Lungren, M. P. (2021). Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms. Journal of the American College of Radiology, 18(3), 413–423. https://doi.org/10.1016/j.jacr.2020.09.030

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liyanage U. A., Sirisena N. D., Deshapriya P. C., and Dissanayake V. H. (2023). Breast cancer surveillance in BRCA positive Sri Lankan women: Health equity for a high-risk group at a limited resource setting, *BMC Women's Health*, 23(1),

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, Pp. 4765–4774.

Manikandan P., Durga U., and Ponnuraja C. (2023). An integrative machine learning framework for classifying Seer breast cancer, *Scientific Reports*, 13(1),

Martínez-Ramírez J. M., Carmona C., Ramírez-Expósito M. J., and Martínez-Martos J. M. (2025). Extracting Knowledge from Machine Learning Models to Diagnose Breast Cancer, *Life*, 15(2), Pp. 211,

Mehak A., Ali T. M., Nawaz A., Parwati L. S. (2025). Breast Cancer Predictions Using Machine Learning Algorithms, *Engineering Proceeding*, Pp. 107, 129.

Mustapha M. T. (2022). Breast Cancer Screening Based on Supervised Learning and Multi-Criteria Decision-Making, *Diagnostics*, 12(6), Pp. 13 – 26

Naji M. A., El Filali S., Aarika K., Benlarmar E. H., Abdelouhahid R. A., & Debauche O. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis, *Procedia Computer Science*, 191(1), Pp. 487–492.

Nasser M. and Yusof U. K. (2023). Deep learning based methods for breast cancer diagnosis: A systematic review and future direction, *Diagnostics*, 13(1), Pp. 161.

Nawaz, W., Ahmed, S., Tahir, A., & Khan, H. A. (2018). Classification of breast cancer histology images using ALEXNET. In A. Campilho, F. Karray, & B. ter Haar Romeny (Eds.), Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science 10882(1), Pp. 869–876). Springer. https://doi.org/10.1007/978-3-319-93000-8_99

Olorunshola O. E., Ebuka O. D. and Ademuwagun A. K. (2025). Evaluating the Generalizability of Support Vector Machine for Breast Cancer Detection, *Journal of Computing and Social Informatics*, 4(1), Pp. 1-10

Olorunshola O. E., Ebuka O. D., Ademuwagun A. K. (2024). Evaluating the Generalizability of Support Vector Machine for Breast Cancer Detection, *Trends Artificial Intelligent*, 7(1), Pp.106-112

Prastyo P. H., Paramartha G. Y., Pakpahan M. S. M., Ardiyanto I. (2020). Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms, *Proceeding International Conference Science Engineering*, 3(1), Pp. 455-459

Qodri K. N. (2025). Breast Cancer Classification Using Data Mining Tool (WEKA). (2025). *IJITECH : Indonesian Journal of Information Technology*, 3(1), Pp.25-29.

Radak M., Lafta H. Y., and Fallahi H. (2023). Machine learning and deep learning techniques for breast cancer diagnosis and classification: A Comprehensive Review of Medical Imaging Studies, *Journal of Cancer Research and Clinical Oncology*, 149(12), Pp. 10473–10491.

RamakrishnaM. T. , Venkatesan V. K., Izonin I., Havryliuk M., and Bhat C. R. (2023). Homogeneous Adaboost ensemble machine learning algorithms with reduced entropy on Balanced Data, *Entropy*, 25(2), Pp. 245,

Rashid A., Binta Farhad S. S., Bhuyian A., Yeasmin N., Azim M. A., Alom Z. (2022). A Comparative Analysis of Machine Learning techniques on Breast Cancer diagnosis using WEKA, *25th International Conference on Computer and Information Technology (ICCIT)*, Pp. 663-668,

Rawal, R. (2020). Breast Cancer Prediction Using Machine Learning, *Journal of Emerging Technologies and Innovative Research (JETIR)*, 7(5), Pp. 13–24.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rufai M. A., Muhammad A. S., Garba S., & Audu L. (2020). Machine Learning Model for Breast Cancer Detection, *FUDMA Journal of Sciences (FJS)*, 4(1), Pp. 55–61.

Samiksha Marne, Shweta Churi and Maheshwari Marne (2020). Predicting Breast Cancer using effective Classification with Decision Tree and K Means , *Clustering technique,IEEE Xplore*, Pp. 39 – 42

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

Sevinç E. (2022). An empowered AdaBoost algorithm implementation: A COVID-19 dataset study, *Computers & Industrial Engineering*, 165(1), Pp. 107912

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Sharma, D., Kumar, R., & Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning, *Measurement: Sensors*, 24(1), Pp. 1-9, Article 100560. https://doi.org/10.1016/j.measen.2022.100560

Shanbehzadeh M., Nopour R., Erfannia L., Amraei M., Mehrabi N. (2022). Comparing data mining algorithms for breast cancer diagnosis, *Shiraz E-Medical Journal*, 23(7),

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. Scientific Reports, 9, Article 12495. https://doi.org/10.1038/s41598-019-48995-4

Shubham K., & Kamalraj R. (2022). Breast Cancer Detection Using Machine Learning Algorithms. *International Journal of Advances in Engineering and Management (IJAEM)*, 4(3), Pp. 987–994

Singh G. (2020). Breast Cancer Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, *Engineering and Information Technology*, 8(4), Pp. 1 – 8

Singh J., Raahat S., Sharma A., Dellu S., Arora S. (2025). Early Detection of Breast Cancer: Comparative Analysis of Machine Learning and Deep Learning Algorithms, *International Journal of Preventive Medicine and Health (IJPMH)*, 5(2), Pp. 1-8,

Singh L. K., Khanna M., and SinghR. (2023). Artificial intelligence based medical decision support system for early and accurate breast cancer prediction, *Advances in Engineering Software*, 175(1), Pp. 103338

Singh, V., Bhatt, D., Devi, M., Bhatt, J., & Sharma, A. (2022). Explainability of deep learning models for breast cancer detection: A survey. Artificial Intelligence in Medicine, 132, Article 102372. https://doi.org/10.1016/j.artmed.2022.102372

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

Stephen A. and Adekoya A. F. (2022). Performance Analysis of Data Mining Algorithms: Breast Cancer Predictive Models, *Journal of Health & Medical Informatics*, 13(7), Pp. 1-7

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: *A Cancer Journal for Clinicians*, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Tachicart R., Rezki H., Korchi A., Abatal A. (2025). A comparative study of machine learning algorithms for breast cancer diagnosis, *Journal of Medical Artificial Intelligence*, pp 1-14, doi: 10.21037/jmai-24-368

Thomas T., Pradha N. and Dhaka V. S. (2020). Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey, *Proceedings of the Fifth International Conference on Inventive Computation Technologies IEEE Xplore*, Pp. 192-196

Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., & Reede, D. (2017). Interpretive error in radiology. American Journal of Roentgenology, 208(4), 739–749. https://doi.org/10.2214/AJR.16.16963

[68] Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. Radiology, 292(1), 60–66. https://doi.org/10.1148/radiol.2019182716

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A. K., & Martí, R. (2018). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE Journal of Biomedical and Health Informatics, 22(4), 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873

Yee W. S., Ng H., Tzen T., Yap V., Goh V. T., Ng K. H., & Cher D. T. (2022). An Evaluation Study on the Predictive Models of Breast Cancer Risk Factor Classification, *Journal of Logistics, Informatics and Service Science*, 9(3), Pp. 129–145.

© 2026 The Author(s). Published by Faculty of Engineering, Federal University Oye-Ekiti.
This is an open access article under the CC BY NC license. (https://creativecommons.org/licenses/by-nc/4.0/)