Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Machine Learning Techniques for Breast Cancer Prediction: A Comprehensive Review on Techniques and Datasets

Archana Singh[1*], Kuldeep Singh Kaswan[2], Rajani[3]

[1,2]Department of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.

[3]Department of Computer Science, Kalindi College, University of Delhi, New Delhi, India.

Article History:

Received: 12-11-2024

Revised: 17-12-2024

Accepted: 06-01-2025

Abstract:

Breast cancer (BC) is a key health concern worldwide; early detection and accurate diagnosis are crucial for improving patient outcomes. Machine learning techniques have shown promise in revolutionizing breast cancer diagnosis. This review covers various machine learning techniques, ranging from classic algorithms like decision trees and k-nearest neighbor (KNN) to advanced methodologies such as ensemble learning and deep learning. The variation in accuracy metrics and the lack of standardized evaluation methodologies make it challenging to directly compare the performance of diverse algorithms. This study discusses the data sources and methodology employed in the examined studies, as well as comparing various machine learning approaches. The findings of this work indicate that machine learning approaches may greatly enhance the diagnosis of breast cancer. The comparison analysis clarified that ensemble learning provided better results on the Wisconsin breast cancer dataset (WDBC), attaining the highest metrics with an accuracy, precision, recall, and an F1-score. Furthermore, the optimized framework demonstrated highest accuracy on ultrasound image data, underscoring its efficacy and robustness in medical diagnostics. This review provides a unique and critical analysis of the machine learning techniques and data sources used in breast cancer diagnosis and highlights the need for further research in this area.

Keywords: breast cancer, diagnosis, ml techniques, dataset, wdbc.

1. Introduction

Breast cancer is a multifaceted and widespread ailment that predominantly impacts the cells and tissues of the breast. This condition is identified by the unregulated proliferation and division of irregular cells in the breast tissue, resulting in the development of tumors, which may be categorized as either benign or malignant. Breast cancer affects individuals of all genders, though it is more commonly diagnosed in women. A new Global Breast Cancer Initiative Framework was announced by the World Health Organization (WHO) in 2023 with the objective of offering a strategic route to reach the lofty target of aims to prevent 2.5 million breast cancer-related deaths by 2040 [1]. The framework strongly recommends that nations adopt three fundamental pillars of action centred on health promotion, early discovery, and appropriate diagnosis of breast cancer. These strategic pillars are envisioned as the driving forces to meet the set targets. This review paper seeks to give a summary of the various studies that have been conducted in this area of different ML algorithms. In this context, 38 studies published between 2016 and 2024 have been reviewed, covering a range of approaches, including “logistic regression, decision trees, K-nearest neighbors, artificial neural networks, support vector machines, and deep learning. The Wisconsin Diagnostic Breast Cancer (WDBC) is the reference database used

in the majority of references, and it serves as a standard for comparing the outcomes of various methods.

Breast cancer is the most frequent type of cancer in people, with over 2.3 million new cases reported each year. Surprisingly, it is either the leading or secondary cause of female cancer-related mortality in 95% of countries. Despite these alarming statistics, survival rates for breast cancer exhibit substantial disparities both between countries and within them. Unfortunately, around eighty percent of deaths from cervical and breast cancer happen in countries of low and middle income. Age, a family history of the condition, certain genetic mutations, hormonal changes, and environmental variables are among the risk factors for BC. Early detection and correct treatment play a crucial role in improving the survival rates for the diagnosis of BC [2]. Machine learning allows computers to learn from data and expand performance without explicit programming. They cover various methods, including unsupervised learning (which finds patterns in unlabelled data) and supervised learning (which teaches models from labelled examples). Reinforcement learning enables systems to make decisions through trial and error. Deep learning employs neural networks to handle complex tasks like image and speech recognition.

A hybrid technique for BC prediction based on machine learning was given [3]. The ability of DL models to accurately identify metastatic lymph nodes, a critical component of staging and therapy planning was highlighted by their work [4]. The SVM method yielded the best result, suggesting its potential for usage in clinical applications with 97.14% accuracy [5]. It is found that the SVM got the highest 94.44% accuracy [6]. [7] evaluated the effectiveness of ML algorithms and discovered that SVM produces the best result with 97.5% accuracy. [8] focused on diagnostic accuracy as well as predicting the prognosis of BC.

In a study, the KNN approach found an accuracy of 95.61% using the Wisconsin Breast Cancer dataset [9]. [10] designed an MLP model that incorporated feature selection and data balancing techniques, achieving an accuracy of 97.70% and demonstrating its potential for clinical applications. [11] offered an analysis of several ML methods applied to the diagnosis of BC, emphasizing the high classification accuracy attained by DTs, SVMs, and ANNs. In a study comparing ML algorithms discovered that the Random Forest algorithm outperformed in the context of accuracy and F1-score [12]. In [13] different models and found that SVM and RF were the best models, achieving an accuracy of 95.1739%. [14] utilized ensemble learning for breast cancer detection with an accuracy of 99.30%. [15] compared the SVM and ANN and identified that the SVM was more accurate in detecting breast cancer. [16] evaluated the performance of ANN and decision trees and observed that ANN provides the best result.[17] employed feature selection techniques and achieved 97.45% accuracy by using a using a random forest algorithm. [18] proposed an ensemble ML approach and got an accuracy of 97% by combining multiple ML models. [19] compared MLP and CNN models and found that CNN has higher accuracy than MLP in breast cancer detection. [20] developed an improved predictive model using ML. In this study, the polynomial SVM achieved an accuracy of 99.12%.[21] proposed an optimized framework for BC classification, achieving a high accuracy of 99.86%. [22] investigated BC detection with machine learning, with the XGboost algorithm achieving an accuracy of 98.24%.

[23] designed a breast cancer detection model using thermographic images, and CNN outperformed ML models with 99.65% accuracy. [24] developed a BC diagnosis method based on image processing

and segmentation techniques, achieving a high accuracy rate of 96.8%. [25] developed a deep NN-based computer-aided diagnosis system in which transfer learning was used and achieved an accuracy of 99.70%.[26] designed a deep learning approach for early-stage BC diagnosis, achieving an accuracy rate of 97.2% with a hybrid optimization technique. [27] classified breast cancer metastases, and their decision tree method had an 83% accuracy rate. [28] found that the SVM and logistic regression algorithms had the greatest accuracy rate of 99.12%.

A decision tree was used to find the category of BC C [29]and got an accuracy 97.4%. [30] discovered that the SVM had a 97.36% accuracy rate. The accuracy of predictions can be increased yet more by using ensemble models and optimized frameworks. The application of ML algorithms in breast cancer diagnosis has increased day by day because they have become extremely useful tools in the diagnosis of BC. Mammography, MRI scans, histopathological images all are clarified using different algorithms to help clinicians spot the early-stage tumors at the right instance providing a better diagnostic result. [31] proposed an ensemble framework for BC prediction with 97.66% accuracy value.

2. ML techniques for breast cancer diagnosis

The most effective way to categorize breast cancer trends and make decisions is through ML algorithms, which take significant components out of massive datasets. Positive outcomes have been obtained from the classification of data utilizing the KNN, SVM, and DT algorithms. Additionally, these methods significantly support clinical diagnosis and decision-making. In this section, ML techniques for breast cancer diagnosis are presented. It covers a range of techniques, such as decision trees, random forests, SVM, ANN, logistic regression, and KNN. Each technique is described, including its underlying principles, training process, and specific applications in breast cancer diagnosis. The significance of these techniques is discussed, providing insights into their suitability for clinical implementation [32]. There are various ML technologies deployed for breast cancer diagnosis. The major ones among these are presented in Fig. 1.

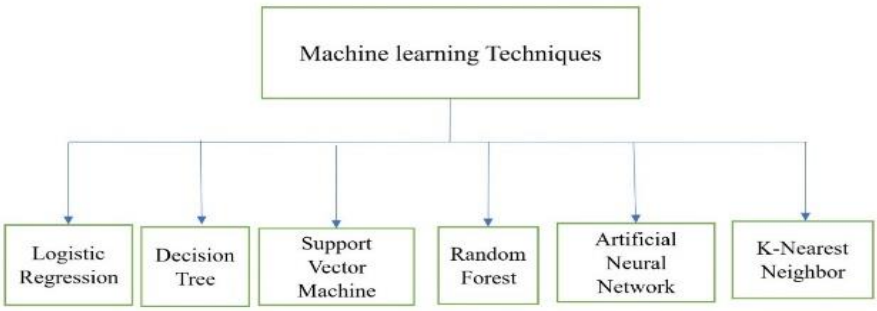

Fig. 1 Machine learning techniques

2.1 Logistic regression (LR)

Logistic regression may be used to detect if a tumor is benign or malignant. The algorithm learns from a labeled dataset, where each instance represents a tumor and is associated with a binary class label indicating its malignancy.The logistic regression model uses a special function known as “the sigmoid function” to map the linear relationship (if one feature grows, the output also grows and when one feature reduces, the output reduces.) of input features to a probability (between 0 and 1). This

probability represents the chance of the tumor being malignant. A threshold is then applied to the predicted probability to make the final classification decision. It has low computational cost and is easy to implement and design but the performance of this algorithm decreases with large data sets [35].

2.2 Random Forest (RF)

Multiple DTs are used in the random forest ensemble learning technique to generate predictions. It is used to build a set of decision trees for diagnosing breast cancer, with each tree being trained using a different subset of the available information and characteristics [11]. Either a majority vote or an average of the estimates made by each individual tree is used to determine the final prediction. RF can handle high-dimensional data, capture relationships, and minimize overfitting.

2.3 Decision tree (DT)

A DT is a common ML technique for classification tasks such as diagnosing breast cancer. By recursively dividing the data according to the values of input features, it creates a model that resembles a tree [36]. In this tree structure, every leaf node represents a class (benign or malignant), and every internal node represents a feature. The DT algorithm learns from the data to determine the optimal splits that best separate the different classes. DT has moderate computational complexity together with high readability; however, this methodology might overfit. Fig. 2 is a representation of a decision tree for BC diagnosis [11].

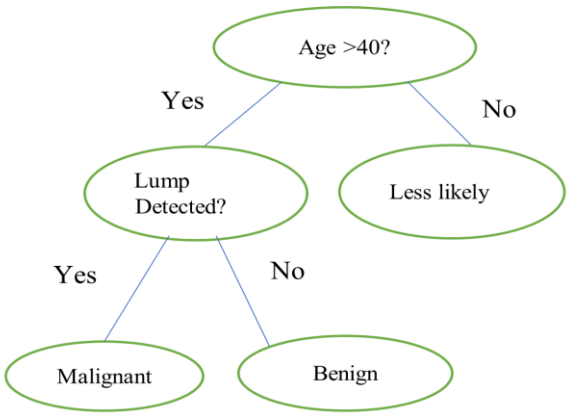

Fig. 2 Decision Tree

2.4 Support vector machine (SVM)

SVM is a highly effective supervised learning system designed for categorization. The goal is to identify the best hyperplane in the breast cancer dataset that maximally divides the two groups (malignant and benign). High-dimensional data can be handled by “SVM”, and it works by finding the support vectors, which are the data points bordering the decision boundary, to classify new instances [11]. SVM is more computationally expensive than the other algorithms especially when dealing with a large amount of data; requires proper calibration of the parameters.

2.5 Artificial neural network (ANN)

ANNs are the tool for breast cancer diagnosis and are able to realize complex patterns within medical data. These networks learn from a vast array of patient information, such as mammogram images and

clinical records, to separate refined nuances indicative of malignancy. These advancements hold promise for more efficient and reliable breast cancer detection. It has very high computational cost, especially deep learning models requires extensive training time and higher performance hardware.

2.6 K-nearest neighbor (KNN)

As a non-parametric algorithm KNN doesn't assume anything about the distribution of the underlying data. Research by [32] demonstrated an impressive 94% accuracy in classifying breast lesions using a CNN architecture. “Ensemble learning is a technique that combines the predictions of multiple base models known as weak learners to create a more robust and accurate final prediction”. It leverages the concept that aggregating the wisdom of multiple models often yields better results than relying on a single model. KNN is simple to implement but not scalable for large datasets and has high computational cost during prediction, as it involves calculating distances for all instances in the training dataset.

2.7 Ensemble machine learning

Ensemble machine learning is focused on the provision of the final prediction by aggregating several models to acquired higher performance. The basic concept of ensemble methods is that combining different models will help to minimize the biases, variances and errors inherent in each individual model. These include bagging for instance the random forests which directly work to reduce variance by training different models on bootstrapped sample data, and boosting for instance the AdaBoost which in its theory, concentrates on correcting for mistakes done in the preceding models. Stacking is the process in which multiple models’ prediction is aggregated by another model for the final result. Ensemble methods are used in all complex tasks such as classification and regression, to obtain accurate and more reliable outcomes.

Each of the aforementioned algorithms has advantages and disadvantages that should be taken into account when creating models for breast cancer diagnosis. The particular needs, dataset properties, and intended performance indicators all play a role in choosing the best method. The performance of ML algorithms can be generalized by considering key factors such as accuracy, sensitivity, and F1-score. Fig. 3 shows the performance of various ML algorithms and branches out to include decision tree and SVM applied to specific datasets, such as WDBC and ultrasound images, with their respective performance metrics presented. Another branch explores deep neural networks applied to the BreakHis dataset, along with the expected accuracy, sensitivity, and specificity values. In the end, the process leads to a conclusion and insights, likely summarizing which algorithms performed best on different datasets and providing valuable information for breast cancer detection.

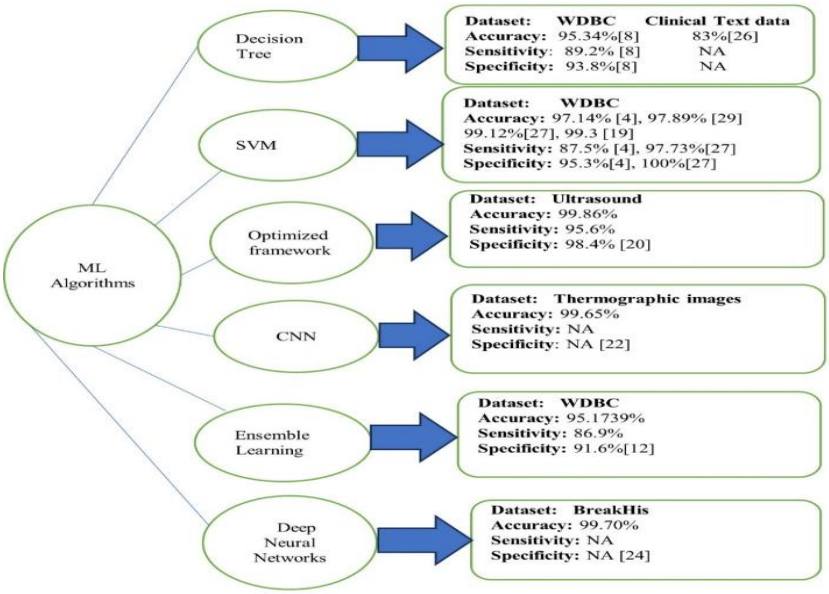

Fig. 3 Generalized view of performance of various ML algorithms in breast cancer diagnosis

3. Datasets used in breast cancer diagnosis

The studies reviewed in this paper draw upon a variety of datasets, each tailored to specific research objectives and methodologies. Table 1 provides various datasets used in breast cancer diagnosis, as analysed in this study. Most of the papers in this review are based on the "Wisconsin Diagnostic Breast Cancer (WDBC)" dataset.

Table 1 Various Datasets used in Breast Cancer Diagnosis

Dataset Name	Description	Features	Labels	Source	Application
Wisconsin Diagnostic Breast Cancer (WDBC)	Features computed from digitized FNA images of breast mass	Radius, texture, smoothness, compactness, concavity, symmetry and fractal dimension, etc.	Benign (B), Malignant (M)	UCI Machine Learning Repository	Developing predictive models for classifying breast masses as benign or malignant based on extracted features.
Breast Cancer Histopathological Database (BreakHis)	Histopathological images of breast tissue samples	High-resolution images of cellular structures	Types of breast cancer: invasive ductal carcinoma, lobular carcinoma, etc.	BreakHis database	Developing image-based classification and segmentation models for identifying and classifying different types of breast cancer from histopathological images.

Mammography Databases	Mammographic images captured using X-ray imaging	Grayscale images of breast tissue density and anomalies	Normal, Benign, Malignant, BI-RADS categories	Various medical imaging repositories and institutions	Developing machine-learning models for mammogram interpretation, anomaly detection, and breast cancer risk assessment.
Ultrasound Image Databases	Ultrasound images created using sound waves	Images depicting tissue density, texture, and lesions	Lesion characteristics, Clinical outcomes	Medical institutions, research repositories	Developing machine-learning algorithms for ultrasound image analysis and breast lesion detection.
Thermographic Image Databases	Infrared images capturing heat patterns emitted by the body	Images showing temperature variations in breast tissue	Abnormal temperature patterns	Medical research databases	Thermal analysis and machine learning-based detection of temperature anomalies in breast tissue.

These are some of the key types of datasets used for BC diagnosis using ML techniques. Each dataset has its unique characteristics and focuses on different aspects of breast cancer detection and classification, contributing to the creation of precise and efficient diagnostic models. Table 2 summarizes the strengths and limitations of each breast cancer dataset in a clear and organized format. These datasets serve as essential resources for advancing breast cancer diagnosis through ML techniques. They encompass various data modalities, including clinical features, histopathological images, mammograms, ultrasound images, and thermographic data, facilitating the development of good diagnostic models for improved patient care. Most of the papers in this study used the “WDBC” dataset for breast cancer diagnosis, which is described in this section. The Breast Cancer Histopathological Database (BreakHis) has rich histopathological images, offers detailed cellular characterization with different sample types, yet comes with the drawback, it is more complex and needs a high and time-consuming computational power. Mammography databases give information about the mammographic image reflecting the breast tissue density and abnormality based on multi-category labels but is invasive, involves radiation exposure, and needs the attention of a specialist. Ultrasound image databases present a non-invasive real-time information qualified for lesion detection algorithms but have the problem of limited depth penetration and being dependent on the operator. Finally, for diagnostics and potential biomarkers based on thermography, non-invasive thermographic image databases contain temperature patterns and are poorly influenced by changes in external temperature; however, the resolution of thermographic images is significantly lower compared to other types of images.

Table 2 Strengths and Limitations of each dataset

Dataset Name	Strengths	*Limitations*
Wisconsin Diagnostic Breast Cancer (WDBC)	Well-established dataset widely used in breast cancer research. Comprehensive features from digitized images. Labels for supervised learning (benign and malignant).	Limited to features derived from FNA images, may not capture all aspects of breast cancer.
Breast Cancer Histopathological Database (BreakHis)	High-resolution histopathological images for detailed cellular information. Enables image-based classification and segmentation models. Diverse types of breast cancer samples.	Requires advanced image processing techniques, potentially computationally intensive. Limited to histopathological information.
Mammography Databases	Mammographic images provide insights into breast tissue density and anomalies. Grayscale images for mammogram interpretation. Multi-category labels (Normal, Benign, Malignant).	Radiation exposure concerns with mammography. Interpretation subjective and dependent on radiologists' expertise.
Ultrasound Image Databases	Non-invasive ultrasound provides real-time information. Suitable for developing algorithms for breast lesion detection. Captures tissue density, texture, and lesion characteristics.	Operator dependence in capturing images. Limited depth penetration compared to other imaging modalities.
Thermographic Image Databases	Non-invasive imaging captures temperature patterns in breast tissue. Potential for early detection based on abnormal temperature variations. Complementary information to other modalities.	Sensitivity to external factors affecting body temperature. Limited resolution compared to other imaging techniques.

3.1 Wisconsin diagnostic breast cancer dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was obtained from the "ML UCI Repository," an online source repository. Dr. Wolberg and the hospitals of Wisconsin and the University collected the dataset irregularly in the past. It consists of 569 cases, with no specific categorization as either dangerous or generous. The dataset under consideration is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is widely used in the domain of medical research for developing predictive models. This dataset consists of 32 attributes that represent various clinical and morphological features extracted from breast cancer biopsies. These attributes are essential for distinguishing between two classes of cancer diagnoses: benign and malignant. The dataset includes 569 cases, of which 357 cases are classified as benign and 212 as malignant. This balanced distribution of cases allows for an effective evaluation of machine learning algorithms aimed at breast cancer classification. The detailed attribute measurements provide a comprehensive basis for accurately training predictive models, thereby enabling early and precise diagnosis of breast cancer. Table 3 provides a visual representation of the WDBC dataset.

Table 3 Visualization of WDBC Dataset

Dataset	Attributes	Cases	Classes	
WDBC	32	569	2	
			Benign (357)	Malignant (212)

-Description: A breast mass fine needle aspiration (FNA) sample is utilized to calculate features that describe the features of the cell nuclei seen in the images.

-Labels: Each sample is labelled as either benign (B) or malignant (M).

-Source: "UCI Machine Learning Repository".

-Application: Used to develop predictive models for classifying breast masses as benign or malignant based on the extracted features.

The dataset is widely used for classification tasks, particularly for binary classification algorithms like KNN, SVM, decision trees, and more. The dataset can be accessed from various machine learning libraries or repositories online [33]. It is a popular choice for learning and practicing ML techniques, especially in the medical domain.

The description of various datasets used in breast cancer diagnosis is shown in Fig 4. These datasets include medical imaging, WDBC, clinical records, and histopathological data. Each dataset serves a unique purpose in breast cancer detection and enabling researchers to develop more accurate and personalized diagnostic approaches. These data sources are contributing to the ongoing research progress toward improving breast cancer diagnosis and reducing patient mortality.

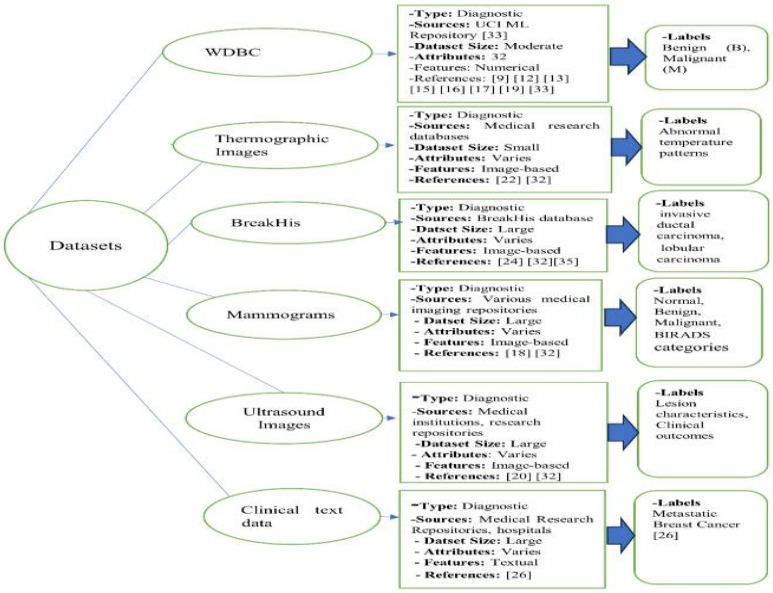

Fig. 4 Diverse datasets with description for breast cancer diagnosis

The datasets for breast cancer research can be collected from various sources. The WDBC has a moderate-size with 30 numerical attributes, which is ideal for diagnostic purposes. The Mammograms dataset presents a large-scale image-based dataset and the Breast Cancer Histopathological Database (BreakHis) is a breast cancer histopathological database, which is made of 7,909 microscopic images of tissue samples of breast cancer, which are split into benign and malignant classes [34]. The Thermographic Images Dataset, though small in size, contributes additional imaging data. Finally, the Clinical Text Data Dataset contains textual features and offers a different dimension to breast cancer analysis. Researchers can select from these datasets based on their specific research objectives, size requirements, and the type of data, like numerical, image-based, or textual, that best suits their research [37].

4. Performance evaluation and comparison

This section reviews the research to compare the performance of ML techniques in breast cancer diagnosis on a variety of datasets. It can be calculated by using the parameters accuracy, precision, recall, and F1-score. True positives (TP) and true negatives (TN) are taken into account while calculating accuracy, which is a measure of the total correctness of forecasts [38]. Accuracy =TP + TN / (TP + TN + FP + FN), Precision = *T*P/ (TP+FP), Recall = TP/ (TP+FN), F1-score = 2*(Precision * Recall) / (Precision + Recall). The terms related to these metrics are described and clarified in Table 4.

Table 4 Description of Short Terms used in Performance Metrices

Short Terms	Description	Clarification
TP	True Positives	Properly identified positive cases
TN	True Negatives	Accurately identified negative cases
FP	False Positives (Type 1 Error)	Negative cases incorrectly predicted as positive

FN	False Negatives (Type 2 Error)	Positive cases incorrectly predicted as negative

Table 5 provides an overview of diverse studies in breast cancer diagnosis. It covers the datasets utilized, the algorithms applied, and the corresponding levels of accuracy, precision, recall, and F1-score achieved in each research and presents a comparison of research and highlighting the dataset used, the algorithms employed, and the performance metrics.

Table 5 Comparison of Various Author's Works in Breast Cancer Diagnosis

Author &Year	Dataset	Algorithms	Accuracy	Precision	Recall	F1score	Limitations
Agarap (2018) [5]	WDBC	Support Vector Machine	.9714	0 .98	0.97	0.97	More dataset may be used and Deep Learning can be applied to improve
Rajaguru and S (2019) [9]	WDBC	KNN	.9561	.9452	-	-	Expensive for large datasets.
Al-Shargabi et al. (2019) [10]	WDBC	Multi-layer Perceptron	.9770	-	-	-	prone to overfitting, particular on small datasets.
Chaurasia and Pal (2020) [13]	WDBC	Ensemble Learning	.951739	-	-	-	Investigate deep learning for better classification and validate & construct an ideal classifier
MurtiRawat et al. (2020) [14]	WDBC	Ensemble Learning	.9930	1.0	.9777	.9887	More data can be added into database which will help in training of ml model and would work more accurately
Hazra et al. (2020) [16]	WDBC	Artificial Neural Networks	.9855	1.00	0.97	0.99	Incorporate the selected method into a practical strategy.
Haq et al. (2021) [17]	WDBC	PCA-SVM	.9745	-	-	0.88	Other feature selection algorithm and Deep learning can be applied.

Jabbar (2021) [18]	WDBC	Bayesian network and (RBF)	0.97	.9672	-	-	Accuracy may be increase by using other methods
Rasool et al. (2022) [20]	WDBC	Polynomial SVM	.9912	.9862	1.00	.993	Try different global datasets to check how the performance is affected by region.
Mangukiya et al. (2022) [22]	WDBC	XGboost	.9824	-	-	-	Examine scalability and efficiency on large datasets.
Hossin et al. (2023) [28]	WDBC	Logistic Regression and SVM	.9912	-	.9773	-	Massive datasets and new algorithms can be used to improve accuracy
Chen et al. (2023) [29]	WDBC	XGBoost	.974	0.960	1.00	0.980	Limited to numerical data
Kadhim and Kamil (2023) [30]	WDBC	GB	.9736	1.00	.9787	-	More datasets may be used.
Sharma et al. (2024) [31]	WDBC	Stacked based ensemble classifier	.9766	-	-	-	Deep learning technique may be used in to handle large amount of data and ensure patient's data privacy.
Allugunti (2022) [23]	Thermographic images	CNN	.9965	-	-	-	Lack of practical relevance
Aljuaid et al. (2022) [25]	BreakHis	Deep Neural Networks	.9970	.9959			Model may be improved by using different of datasets.
Dewangan et al. (2022) [26]	MRI Images	Hybrid optimization (BPBRW with HKH-ABO)	.996	1.0	.999	-	Security approach may be added to enhance performance.
Botlagunta et al. (2023) [27]	Clinical Text Data	Decision Tree	0.83	0.83	0.86	0.85	Various statistical methods and machine learning models may be deployed.

Michael et al. (2022) [21]	Ultrasound Images	Optimized framework	.9986	1.00	.996	.998	Different dataset may be used

The studies encompass a range of datasets, including the “Wisconsin Breast Cancer dataset," clinical and pathological features, mammograms, thermographic image data, and the BreakHis dataset. The limitations of each study are also addressed in this review paper. Overall, the comparison highlights the effectiveness of various ML techniques in BC across different datasets. The choice of algorithm and dataset greatly affects the accuracy achieved and other metrics.

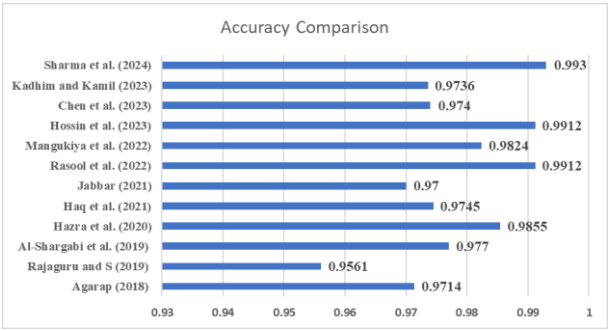

Fig. 5 Performance metrices chart of different ML algorithms on WDBC Dataset

The values of accuracy, precision, recall, and F1-score gained by using several ML algorithms on the WDBC dataset are shown in Fig. 5. It shows that, when compared to the other methods, ensemble have done the best and attained the maximum value [31]. These findings provide valuable insights for researchers in developing improved and accurate models for breast cancer.

5. Challenges

While the presented studies highlight significant achievements in BC diagnosis using diverse ML algorithms and datasets, challenges remain. Ensuring robustness across various datasets, especially when dealing with distinct modalities like mammograms, ultrasound images, and clinical data, is essential. Addressing imbalanced datasets to prevent biased predictions is crucial for clinical reliability. Future directions involve integrating multi-modal data to enhance accuracy further. Embracing explainable AI methods will bridge the gap between complex models and clinical understanding. Additionally, advancing privacy-preserving techniques and ensuring model fairness across diverse patient groups will be pivotal. As we move forward, these challenges and directions will shape the evolution of ML-driven breast cancer diagnosis.

A major issue in using machine learning algorithms for diagnosis of diseases is the availability of appropriate materials, appropriately large and diverse. [9] and [5] reported maximum accuracies of 95. 75% and 99. 5% respectively, on the WDBC dataset but they called for more diverse data and high computational costs. In the case of BreakHis, [25] employed Deep Neural Networks with high accuracy and suggested other diverse datasets.

6. Conclusion & future scope

This paper assesses and contrasts the worth of ML algorithms, including SVM, KNN, RF, DT, PCA, LR, ELM, ANN, MLP, XGBoost, and DL, as employed by different researchers in their studies. According to the analysis of various author's works among those who diagnose breast cancer, ML algorithms have shown tremendous promise in terms of producing accurate and dependable outcomes. To build predictive models for breast cancer diagnosis, many methods such as SVM, DT, RF, MLP, ANN, deep learning, and an optimized framework have been used. The reported accuracies range from .9561 to .993 on the WDBC dataset, indicating the effectiveness of these models in distinguishing between malignant and benign cases. In conclusion, the analysis of existing works highlights the potential of ML techniques in breast cancer diagnosis. It is observed that ensemble method obtains the best performance with the highest accuracy on the WDBC dataset and an optimized framework on thermographic images. Comparative analysis shows that the ensemble learning outperformed, and their accuracy, precision, recall, and F1-score are .993, 1.00, .9777, and .9887, respectively. The future scope lies in addressing the aforementioned areas of improvement to boost the accuracy, sensitivity, and specificity of these models on different datasets, ultimately contributing to more effective breast cancer management and improved patient outcomes.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Conflicts of Interest : The authors declare that there is no conflict of interest.

References

[1] WHO launches new roadmap on breast cancer. (2023, February 3). Retrieved from https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer.

[2] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[3] Rathi, M., & Pareek, V. (2016). Hybrid approach to predict breast cancer using machine learning techniques. International Journal of Computer Science and Engineering, 5(3), 125-136.

[4] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... Geessink, G. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318(22), 2199-2210.

[5] Agarap, A. F. M. (2018). On breast cancer detection: an application of machine learning algorithms on the Wisconsin Diagnostic Dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 5-9).

[6] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4).

[7] Tahmooresi, M., Afshar, A., Rad, B. B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(3-2), 21-27.

[8] Yue, Y., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[9] Rajaguru, H., & SC, S. R. (2019). Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. Asian Pac J Cancer Prev, 20(12), 3777-3781.

[10] Al-Shargabi, B., Alshami, F., & Alkhawaldeh, R. (2019). Enhancing multi-layer perception for breast cancer prediction. International Journal of Advanced Science and Technology, 0(0).

[11] Tumuluru, P., Lakshmi, C. P., Sahaja, T., & Prazna, R. (2019). A review of Machine Learning techniques for breast cancer diagnosis in medical applications. In 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 618-623).

[12] Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. Procedia Computer Science, 171, 593-601.

[13] Chaurasia, & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. SN Computer Science, 1(5), 270.

[14] Rawat, R. M., Panchal, S., Singh, V. K., & Panchal, Y. (2020). Breast Cancer detection using K-nearest neighbors, logistic regression and ensemble learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 534-540).

[15] Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. ICT Express, 6(4), 320-324.

[16] Hazra, R., Banerjee, M., & Badia, L. (2020). Machine learning for breast cancer classification with ANN and decision tree. In 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 522-527).

[17] Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., ... & Zhou, W. (2021). Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques. IEEE Access, 9, 22090-22105.

[18] Jabbar, M. A. (2021). Breast cancer data classification using ensemble machine learning. Engineering and Applied Science Research, 48(1), 65-72.

[19] Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth, 4, 1-11.

[20] Rasool, C., Bunterngchit, L., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. International Journal of Environmental Research and Public Health, 19(6), 3211.

[21] Michael, E., Ma, H., Li, H., & Qi, S. (2022). An optimized framework for breast cancer classification using machine learning. BioMed Research International.

[22] Mangukiya, M., Vaghani, A., & Savani, M. (2022). Breast cancer detection with machine learning. International Journal for Research in Applied Science and Engineering Technology, 10(2), 141-145.

[23] Allugunti, V. R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49-56.

[24] Chaudhury, S., Krishna, A. N., Gupta, S., Sankaran, K. S., Khan, S., Sau, K., ... & Sammy, F. (2022). Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer. Computational and Mathematical Methods in Medicine.

[25] Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Computer Methods and Programs in Biomedicine, 223, 106951.

[26] Dewangan, K. K., Dewangan, D. K., Sahu, S. P., & Janghel, R. (2022). Breast cancer diagnosis in an early stage using novel deep learning with hybrid optimization technique. Multimedia Tools and Applications, 81(10), 13935-13960.

[27] Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports, 13(1), 485.

[28] Hossin, M. M., Shamrat, F. J. M., Bhuiyan, M. R., Hira, R. A., Khan, T., & Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. Bulletin of Electrical Engineering and Informatics, 12(4), 2446-2456.

[29] Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience.

[30] Kadhim, R. R., & Kamil, M. Y. (2023). Comparison of machine learning models for breast cancer diagnosis. IAES International Journal of Artificial Intelligence, 12(1), 415.

[31] Sharma, D., Goyal, R., Mohana, R., (2024). An ensemble learning-based framework for breast cancer prediction. Decis. Anal. J. 10, 100372.

[32] Li, H., Dong, J., Tian, J., Wang, Y., & Yu, Z. (2020). Breast Lesion Classification in Digital Mammography Images Based on Convolutional Neural Networks. Frontiers in Oncology, 10, 300.

[33] Breast Cancer Wisconsin (Diagnostic). Retrieved from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29.

[34] Breast Cancer Histopathological Database (BreakHis). Retrieved from https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.

[35] Singh, A., & Kaswan, K. S. (2024, February). Empirical Analysis on Breast Cancer Datasets with Machine Learning. In 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) (Vol. 5, pp. 223-227). IEEE.

[36] Singh, A., & Kaswan, K. S. (2024). Improved Methodology for Breast Cancer Prediction through Integration of Hard Voting Ensemble Classifier on WDBC Data Set. Communications on Applied Nonlinear Analysis, 31(6s), 179-192.

[37] Singh, A., & Kaswan, K. S. (2024). Breast Cancer Diagnosis using Soft Voting Classifier Approach. In 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP) (pp. 292-297). IEEE Computer Society.

[38] Singh, A., Kaswan, K. S., & Rajani (2025). VELM: a voting based ensemble learning model for breast cancer prediction. *Physica Scripta*, *100*(2), 026002.