International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Review Article

A Review of Machine Learning Techniques for Breast Cancer Prediction

Sajal Kumar Kar[1*], Achyut Pandey[2] and Chandra Shekhar Gautam[3]

[1]APS University, Rewa (M.P.), India
[2]Govt. TRS College Rewa (M.P.), India
[3]AKS University, Satna (M.P.), India

Received 01 May 2025, Accepted 12 May 2025, Available online 13 May 2025, **Vol.15, No.3 (May/June 2025)**

Abstract

Cancer is currently one of the most widespread diseases among humans, both in occurrence and mortality. Cancer care is a growing area of focus for developing interventions to improve the overall quality of life and longevity. Regular physical exercise is widely recognized as a key element of rehabilitation for various chronic conditions, contributing to an improved quality of life and a reduced risk of all-cause mortality. Recent observational studies indicate that moderate physical activity may lower the risk of cancer-related mortality, suggesting that exercise could be an effective approach to enhancing both overall and long-term survival. This research work extensively explores the classification of cancer modalities using machine learning models.

Keywords: *Machine Learning, Breast Cancer Prediction, Deep Learning, Data Mining, Ensemble Techniques.*

1. Introduction

Breast cancer is one of the deadliest and most diverse diseases of the modern era, leading to a significant number of deaths among women worldwide. It ranks as the second leading cause of female mortality [1]. Various machine learning [2] and data mining techniques are utilized for predicting breast cancer. Identifying the most effective and suitable algorithm for this prediction remains a crucial task. Breast cancer originates from malignant tumors when uncontrolled cell growth occurs [3]. The excessive development of fatty and fibrous tissues in the breast contributes to the disease. Cancer cells spread throughout tumors, leading to different stages of breast cancer. There are multiple types of breast cancer [4], each classified based on how the affected cells and tissues disseminate in the body.

Ductal Carcinoma in Situ (DCIS) is a type of breast cancer where abnormal cells spread beyond the breast; it is also known as non-invasive cancer [5]. Another type, Invasive Ductal Carcinoma (IDC) [6], also called infiltrative ductal carcinoma [7], occurs when abnormal breast cells extend into surrounding breast tissues. IDC is more commonly observed in men [8]. The third type, Mixed Tumors Breast Cancer (MTBC), also referred to as invasive mammary breast cancer [9], is caused by abnormal duct and lobular cells [10]. Lobular Breast Cancer (LBC) [11], the fourth type, develops within the lobules and increases the risk of other invasive cancers. Mucinous Breast Cancer (MBC) [12], also known as colloid breast cancer, is the fifth type and results from invasive ductal cells. It occurs when abnormal tissues expand around the ducts [13]. The final type, Inflammatory Breast Cancer (IBC), leads to breast swelling and redness. It is an aggressive and fast-growing cancer that emerges when lymph vessels become blocked by cancerous cells [14].

Data mining is the process of extracting valuable information from large datasets. Various data mining techniques and functions play a crucial role in detecting diseases. Techniques such as machine learning, statistics, databases, fuzzy sets, data warehousing, and neural networks aid in the diagnosis and prognosis of different types of cancer [15], including prostate cancer, lung cancer [16], and leukemia [17].

The traditional approach to cancer detection relies on the "gold standard" method, which involves three key tests: clinical examination, radiological imaging, and pathology testing [18]. While this conventional method identifies the presence of cancer using a regression-based approach, modern machine learning techniques employ model-based designs. These models are developed to predict unseen data, yielding highly accurate results during training and testing phases [19]. The machine learning process follows three primary steps: pre-processing, feature selection or extraction, and classification [20]. Among these,

*Corresponding author's ORCID ID: 0000-000-0000-0000
DOI: https://doi.org/10.14741/ijcet/v.15.3.3

feature extraction is the most critical, as it significantly aids in cancer diagnosis and prognosis. This process helps distinguish between benign and malignant tumours, enhancing the accuracy of cancer detection [21].

Data mining and machine learning algorithms play a crucial role in diagnosing and predicting various types of breast cancer, as illustrated in Figure 1. Data mining techniques [22], such as classification, regression, and clustering, help extract meaningful insights about breast cancer patients. These algorithms [23] utilize training datasets, which enable the prediction of different types of breast cancer [24]. This paper is structured into multiple sections. Section 2 discusses key machine learning algorithms used for breast cancer prediction. Section 3 focuses on major ensemble techniques applied in breast cancer prediction. Section 4 covers deep learning approaches for breast cancer diagnosis. Section 5 presents a survey on breast cancer, while Section 6 reviews various machine learning and deep learning algorithms. Section 7 details the study selection and materials used in this research. Section 8 provides a discussion, and Section 9 concludes this review paper.

2. Machine learning algorithms for breast cancer prediction

Machine learning is a self-learning method [25] where algorithms are designed to learn from past datasets. By inputting a large volume of data, a machine learning model analyzes the information and, based on the trained model, makes predictions about future outcomes [24], [26], [27]. For breast cancer prediction, the key machine learning algorithms include:

2.1. Decision tree (DT)

The Decision Tree [37] is a classification and regression model. It divides a dataset into smaller subsets, enabling predictions with high accuracy. The Decision Tree method includes models such as CART [38], C4.5 [39], C5.0 [40], and the conditional tree [32], [41].

2.2. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is commonly used in pattern recognition and is an effective approach for predicting breast cancer. In this method, each class is given equal importance when recognizing patterns. KNN [36] extracts similar-featured data from large datasets and classifies them based on feature similarity [32].

2.3. Logistic regression (LR)

Logistic Regression (LR) is a supervised learning algorithm that involves multiple dependent variables. The output of this algorithm is binary. Logistic regression [35] is capable of producing continuous outcomes for specific data. It is based on a statistical model that uses binary variables [32].

2.4. Artificial neural network (ANN)

An Artificial Neural Network (ANN) [28] is a widely used algorithm in the data mining process. It comprises an input layer, hidden layer, and output layer. This method is particularly effective for identifying complex patterns [29]. The algorithm relies on parallel processing [30], distributed memory [31], collective solutions, and network architecture [32][34].

2.5. Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm employed for both classification and regression problems [44]. It utilizes both theoretical and numerical functions to address regression issues. SVM provides high accuracy when predicting outcomes for large datasets and is a powerful machine learning technique based on 2D and 3D modelling [32], [45].

2.6. Naive Bayes algorithm (NB)

The Naive Bayes algorithm is used to make assumptions based on large training datasets. It calculates probabilities using the Bayesian method [42]. This algorithm achieves high accuracy when determining the probabilities of noisy data used as input [43]. It is an analogy classifier that compares training datasets with training tuples [32].

2.7. K-mean algorithm

The K-Means algorithm is a clustering method that partitions data into smaller clusters. It is used to identify similarities between data points. Each data point belongs to at least one cluster, which is suitable for evaluating large datasets [48].

2.8. Random forest (RF)

The Random Forest algorithm [46] is a supervised learning method [47] used to solve classification and regression problems. It serves as a building block of machine learning, helping to predict new data based on previous datasets [32].

2.9. Gaussian mixture algorithm

The Gaussian Mixture algorithm is a popular technique in unsupervised learning, often referred to as a soft clustering method. It computes the probabilities of various types of clustered data. The implementation of this algorithm is based on expectation maximization [51].

2.10. Hierarchical algorithm

The Hierarchical algorithm evaluates raw data in the form of a matrix. Each cluster is separated

hierarchically, and every cluster contains similar data points. A probabilistic model measures the distance between clusters [50].

2.11. C-mean algorithm

Clusters are identified based on similarity, with each cluster consisting of data points from a single "family." In the C-Means algorithm, each data point belongs to one cluster. This method is primarily used in medical image segmentation and disease prediction [49].

3. Ensemble techniques for breast cancer prediction

Ensemble techniques in machine learning combine the predictions of multiple models to improve accuracy and reduce the likelihood of overfitting. These methods are particularly useful in complex tasks like breast cancer prediction, where the combination of various classifiers can enhance prediction performance. Below are some common ensemble techniques used for breast cancer prediction:

3.1 Bagging

Bagging is a technique where multiple models (typically decision trees) are trained on different subsets of the training data, each selected randomly with replacement. The predictions of these models are averaged (for regression) or voted upon (for classification) to provide a final output. Random Forest is a popular ensemble method based on bagging, often used in breast cancer prediction.

3.2 Boosting

Boosting methods iteratively train models, with each new model focusing on the mistakes of the previous ones. This approach helps to reduce bias and improve the model's performance. Algorithms like AdaBoost, Gradient Boosting, and XGBoost are widely used for breast cancer prediction, as they focus on boosting the accuracy of weak learners by giving more weight to misclassified instances.

3.3.Stacking

Stacking involves training several different models (or base learners) and then combining their predictions using a meta-learner. The base learners can be diverse algorithms (such as decision trees, support vector machines, or logistic regression), and the meta-learner learns to weight the predictions of these models optimally. Stacking is highly effective for complex problems like breast cancer prediction, as it can leverage the strengths of different algorithms.

3.4.Voting

The Voting Ensemble method combines predictions from multiple models (e.g., decision trees, KNN, logistic regression) and produces a final prediction by majority voting (for classification problems). This technique is simple but can be very effective in improving prediction accuracy by aggregating the strengths of diverse models.

3.5 Random Forests

A Random Forest is an ensemble method based on bagging, specifically using decision trees. It builds multiple decision trees and combines their outputs to generate the final prediction. Random Forest is robust to overfitting and has been widely used in breast cancer prediction, offering high accuracy by leveraging diverse decision trees on random subsets of data.

3.6 XGBoost

Extreme Gradient Boosting (XGBoost) is an efficient and powerful ensemble technique that builds decision trees in a sequential manner. It is known for its speed and performance in predictive modeling, especially in tasks like breast cancer prediction. XGBoost minimizes errors by focusing on hard-to-classify instances

4. Deep learning techniques for breast cancer prediction

Deep learning techniques have shown remarkable success in medical diagnostics, including breast cancer prediction. These techniques leverage complex neural network architectures to learn intricate patterns in data, which can significantly improve the accuracy of predictions. Below are some key deep learning techniques used for breast cancer prediction:

4.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are particularly effective in image-based tasks, such as mammogram and ultrasound image classification for breast cancer detection. CNNs use convolutional layers to automatically detect features (like edges, textures, and patterns) from images. This is particularly useful in breast cancer prediction, where image-based data (such as histopathological images or radiological scans) are prevalent. CNNs can classify images into categories (e.g., benign vs. malignant) with high accuracy.

4.2. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are specialized for sequential data analysis, which makes them suitable for temporal datasets. In breast cancer prediction, RNNs can be used when the data involves time-series information, such as changes in tumour markers over time or patient medical histories. These networks capture the temporal dependencies between inputs, which helps in predicting future outcomes.

4.3. Long Short-Term Memory Networks (LSTMs)

LSTMs, a type of RNN, are designed to handle long-range dependencies in sequential data. This technique is useful in breast cancer prediction when considering patient history or sequences of medical events (e.g., previous treatments, hormone therapy, or chemotherapy cycles). LSTMs excel at remembering long-term patterns in the data, which makes them ideal for analyzing patient records or clinical time-series data.

4.4. Deep Neural Networks (DNNs)

Deep Neural Networks (DNNs) are multi-layered neural networks that can model complex relationships between features. These networks consist of multiple hidden layers, each learning higher-level abstractions of input data. DNNs are widely used for structured data like clinical features (age, genetic information, biopsy results) and for classifying patients as high-risk or low-risk for breast cancer.

4.5. Auto encoders

Auto encoders are unsupervised deep learning models used for feature extraction and dimensionality reduction. In breast cancer prediction, auto encoders can learn to compress high-dimensional data, such as genetic information or medical imaging, into a lower-dimensional latent space. This compressed representation can then be used for classification tasks. Auto encoders are useful when there is a need to extract meaningful features from large datasets before performing prediction.

4.6. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are used to generate synthetic data. In breast cancer prediction, GANs can be used to generate synthetic images of tumours, allowing the model to learn from a larger variety of cases, even if real labelled data is scarce. This technique can improve the generalization ability of models and is particularly useful when there is limited data for rare types of cancer.

4.7. Transfer Learning

Transfer Learning involves using a pre-trained model on a large dataset (e.g., ImageNet for image classification) and adapting it to the breast cancer prediction task. This technique is useful when you have limited labeled data, as it allows the model to leverage knowledge learned from a different but related task. For example, a CNN pre-trained on general image data can be fine-tuned to classify breast cancer images.

4.8. Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a fully connected neural network consisting of multiple layers of nodes, each representing a non-linear transformation of the input data. MLPs are widely used for structured data, where they can model complex relationships between various features, such as patient age, family history, and clinical tests, to predict breast cancer risk.

4.9. Deep Reinforcement Learning (DRL)

Deep Reinforcement Learning (DRL) combines reinforcement learning with deep neural networks, where an agent learns optimal strategies through trial and error. In the context of breast cancer, DRL can be used to predict treatment outcomes, where the system "learns" the best treatment plans based on patient data and ongoing results, adjusting its approach for optimal outcomes.

4.10. Attention Mechanisms and Transformers

Attention mechanisms and Transformer models, which are particularly popular in natural language processing, can also be applied to breast cancer prediction. These models can focus on important parts of the input data, such as specific tumor features in medical images or critical time periods in clinical data. Transformers have shown promise in processing large datasets with complex relationships, making them suitable for combining structured and unstructured data for prediction tasks.

5. Survey on breast cancer

According to the most recent GLOBOCAN data, the ratio of breast cancer in males to females is extremely skewed towards females, with only a very small percentage (around 0.5-1%) of breast cancers occurring in men, essentially making breast cancer almost exclusively a female disease; meaning the ratio is overwhelmingly in favor of females, with significantly more cases diagnosed in women compared to men. China is the most populous country in the world. according to the report by (2018), the breast cancer incidence rate in males is 8.6%, while in females, it is 19.2%. Each year, 1.2 million people die from this disease. The American Cancer Society reported 48,100 cases of ductal carcinoma in situ (DCIS) cancer found in women. The 2019 US report predicts that 500 men and 41,760 women will die from breast cancer. Additionally, 3.8 million women in the US are living with breast cancer. In 2019, there were 59,838 cases of DCIS in women in the US. Globally, breast cancer deaths total 458,000. In 2012, the breast cancer death rate in China was 48%, compared to the global death rate of 52%. A 2015 study analyzing data from 1,517 women found a recurrence rate of 100 and a death rate of 132 for breast cancer.

6. Review of machine learning algorithms for breast cancer prediction

The primary goal of this research is to review various machine learning and data mining algorithms that have been used for breast cancer prediction. Our focus is to identify the most accurate and suitable algorithm for predicting breast cancer. To achieve this, we have reviewed and analyzed past studies on breast cancer prediction algorithms, including research papers that explore linear methods (such as Linear Regression, Logistic Regression, and Linear Discriminant Analysis), nonlinear methods (including Classification and Regression Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines), and ensemble algorithms (like Decision Tree, Random Forest, Boosting, and AdaBoost). Many researchers have used combinations of linear and nonlinear algorithms, or nonlinear and ensemble algorithms. To organize this review, we have categorized it into different sections that provide a comparative analysis of each algorithm based on their accuracy rates. Following this comparison, we will highlight the most suitable machine learning algorithm for breast cancer prediction.

6.1. Nonlinear algorithms

For breast cancer prediction, the authors employed and compared various nonlinear algorithms, including Random Forest, Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor. They utilized a Bioinformatics and Medical Science classification approach, which involved selecting the best classifier by comparing data mining algorithms to identify the most effective one for prediction. After evaluating the four classification techniques, the authors concluded that the Support Vector Machine (SVM) outperformed the others, achieving an accuracy of 97.9% [2].

For the prediction and detection of breast cancer, the authors utilized several data mining classification algorithms, including the Bagging Algorithm, IBk (Instance-Based Learning with specific parameters), Random Committee Algorithm, Random Forest Algorithm, and the Simple Classification and Regression Tree (Simple CART Algorithm). The Antenna dataset was employed to evaluate the accuracy of each algorithm. The results were analyzed across various Weka categories such as Bayes, Function, Meta, Lazy, and Trees. After thorough analysis, the authors determined that the Random Forest Algorithm achieved the highest accuracy, making it the most suitable algorithm for breast cancer prediction. The Random Forest Algorithm yielded an accuracy rate of 92.2%, while the Bagging, IBk, and Random Committee Algorithms achieved accuracy rates of 90.9%, 90%, and 90.9%, respectively.

For breast cancer prediction, the authors utilized Gene Expression (GE) and DNA methylation data. They employed three algorithms—Support Vector Machine (SVM), Decision Tree, and Random Forest—to classify nine models for cancer prediction. To evaluate the accuracy and error rates of these algorithms, the authors conducted a comparative analysis using two data mining tools: Weka and Spark. The GE and DNA methylation datasets were filtered to identify common genes, with the primary goal of detecting the presence of tumors. After comparing the performance of the algorithms on both tools, the authors found that SVM achieved the highest accuracy, with 99.68% on Spark and 98.03% on Weka, outperforming the other algorithms.

The authors employed Naive Bayes, K-Nearest Neighbors (KNN), and the J48 algorithm to predict nine different types of cancer, including breast cancer. The dataset, comprising 61 attributes and 1,059 records, was collected with the assistance of various doctors and experts. Using a training set of data, the authors initially compared symptoms to test results to determine whether they were true or false. If the symptoms matched, the result was considered true. Through this approach, the authors predicted various types of breast cancer and evaluated each algorithm based on its accuracy rate. During the breast cancer detection process, Naive Bayes (NB) and KNN demonstrated higher accuracy rates compared to the J48 decision tree classifier, with accuracies of 98.2%, 98.8%, and 98.5%, respectively.

For breast cancer detection, the authors utilized the Support Vector Machine (SVM) combined with a recursive feature elimination technique and a predictive machine learning model. The goal was to identify the most relevant features from a dataset containing data from both benign and malignant cases. The dataset was sourced from the Wisconsin Diagnostics Breast Cancer (WDBC) database. The recursive feature elimination technique was applied to evaluate the SVM algorithm, and a performance matrix was designed to assess the accuracy of the SVM model across different kernel types. The results showed that SVM achieved 99% accuracy with a linear kernel, 98% with an RBF kernel, 97% with a polynomial kernel, and 84% with a sigmoid kernel.

In another study, a classification model was applied to predict breast cancer using a dataset organized into clusters, where each cluster contained data with similar characteristics. To enhance the accuracy of the classification model, the authors employed Hyperparameter Optimization (HPO). The dataset was obtained from the National Cancer Institute (NCI) of Egypt, with the primary objective of predicting breast cancer among Egyptian individuals. The HPO technique was used to improve prediction accuracy. The authors first collected the dataset from the NCI, applied a clustering approach to group similar data patterns, and then used a feature selection method to identify relevant features for prediction. A Decision Tree model was employed to categorize the data, and the HPO technique was applied to detect the presence of breast cancer.

In a separate experiment, the authors used Support Vector Machine (SVM), Decision Tree, Naive Bayes, and K-Nearest Neighbor (KNN) for breast cancer risk prediction and diagnosis. The dataset was again sourced from the Wisconsin Diagnostics Breast Cancer (WDBC). The experiment was conducted using the Weka tool, and the authors employed K-Fold cross-validation by splitting the data into training and testing sets. SVM achieved the highest accuracy at 97.13% with an execution time of 0.07 seconds, outperforming the other algorithms. However, SVM's execution time was longer compared to that of the KNN algorithm.

6.2. Ensemble algorithm

The authors employed several algorithms, including Support Vector Machine (SVM), Nearest Neighbor Algorithm, Logistic Regression, and Naive Bayes, to analyze a breast cancer dataset categorized as malignant or benign. The SVM technique was implemented using two kernels: linear and Gaussian. The Nearest Neighbor algorithm was applied using both Manhattan and Euclidean distances, while Naive Bayes was implemented using normal distribution and kernel distribution methods. The dataset was sourced from the UCI repository, specifically WDBC (Wisconsin Diagnostic Breast Cancer) and WPBC (Wisconsin Prognostic Breast Cancer). The analysis was conducted using the MATLAB tool to classify the data accurately based on their performance. The WPBC dataset contained 34 attributes for breast cancer prediction, while the WDBC dataset included 32 attributes for diagnosis. Among the algorithms, the K-Nearest Neighbors (KNN) algorithm was identified as the most suitable for breast cancer diagnosis and prognosis.

In another study, the authors utilized dimensionality reduction techniques for feature selection and extraction to analyze a breast cancer dataset. They applied three machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LR). The performance of these algorithms was evaluated based on accuracy, precision, and sensitivity. Logistic Regression, which relies on a logistic function, was used to predict outcomes based on independent variables. The KNN algorithm was analyzed using Euclidean distance, with the value of K varying depending on the dataset. The dataset was obtained from the UCI repository, and the Spyder tool was used to measure the accuracy of each algorithm. SVM achieved the highest accuracy at 92.78% [77].

Data from breast cancer patients was collected from the Iranian Centre for Breast Cancer (ICBC) and analyzed using three machine learning techniques: Decision Tree (C4.5), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The authors evaluated the performance of these algorithms based on accuracy, sensitivity, and specificity. For the ANN algorithm, they focused on the multi-layer perceptron (MLP) model to determine its accuracy. The results revealed that SVM outperformed the other algorithms, achieving the highest accuracy of 95.7% for breast cancer prediction [78].

In a separate study, the authors compared two algorithms—Support Vector Machine (SVM) and Artificial Neural Network (ANN)—for breast cancer diagnosis. SVM was used for pattern recognition on the Wisconsin Breast Cancer dataset, considering factors such as the patient's age and tumor size. It classified tumors as either benign or malignant, while ANN was employed to model nonlinear functions. Both algorithms were evaluated using the K-Fold validation technique, with accuracy serving as the primary metric. The results showed that SVM achieved a higher accuracy rate of 96.9%, compared to ANN's accuracy of 95.4% .

6.2.1 Linear and nonlinear algorithm

Feature selection and extraction techniques were applied to Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Naïve Bayes (NB) for breast cancer prediction. The dataset was sourced from the Wisconsin Diagnostic Breast Cancer database. Feature selection involves identifying a subset of relevant features from a large dataset to enhance computational efficiency. The authors compared different feature selection techniques, including Correlation-Based Feature Selection (CFS), Linear Discriminant Analysis (LDA), and Recursive Feature Elimination (RFE). Their analysis revealed that ANN achieved the highest accuracy among the algorithms, with an accuracy rate of 97.0%, followed by SVM at 96.4% and NB at 91%.

Additionally, data mining tools were utilized to classify algorithms such as Naïve Bayes, Bayesian Logistic Regression, Simple CART, and J48 based on various parameters. The dataset was collected from the Wisconsin Breast Cancer database (WBCO). The Decision Tree algorithm was used to partition the data into subsets, while J48 employed decision nodes to predict outcomes from the dataset. The study aimed to determine the best classifier based on Kappa Statistics, Error Rate, and accuracy. Weka tools were used for evaluation, and the results indicated that Simple CART was the most effective algorithm, achieving an accuracy of 98.13%.

6.2.2 Nonlinear and ensemble algorithm

Decision Tree, Naïve Bayes, and K-Nearest Neighbor (KNN) were applied and compared for breast cancer prediction using the Wisconsin original dataset from the UCI Machine Learning Repository. The dataset contained 10 attributes, with 458 benign and 241

malignant cases. Three key matrices were designed based on two classes: actual healthy and actual not healthy, to evaluate data sensitivity. Weka was used to analyze the performance of each algorithm, and the results showed that Naïve Bayes achieved the highest accuracy of 95.99%, outperforming Decision Tree and KNN.

Additionally, the authors analyzed patient data using Naïve Bayes classifiers, Support Vector Machine (SVM), K-Star, Decision Tree, and Artificial Neural Networks (ANN) with the Weka data mining tool. SMO was applied with the RBF kernel to normalize attributes, while KNN was implemented using Multi-Layer Perceptron (MLP), which consisted of an input layer, a hidden layer, and an output layer. The K-Star algorithm was used to assess data similarity. After evaluating these algorithms on a dataset from the University of Medical Centre, Institute of Oncology, the authors found that the J48 Decision Tree had the highest accuracy of 75.52%, surpassing all other models.

6.2.3 Deep learning algorithm

To predict breast cancer in tumor cells, the authors applied deep learning techniques with various activation functions, including Tanh, Rectifier, Maxout, and Exprectifier, and compared them with machine learning algorithms such as Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and Random Forest. The Wisconsin dataset, consisting of 457 benign and 241 malignant tumor cases, was used for analysis. The study found that the algorithm using the Exponential Rectifier Linear Unit (ELU) activation function achieved the highest accuracy of 96.99%.

Additionally, a model was proposed to predict the recurrence of breast cancer, integrating two key algorithms: Extreme Learning Machine and the Bat algorithm. The Bat algorithm was employed to generate biases and random weights, while the dataset, sourced from the Wisconsin Breast Cancer Prognostic database, was analyzed using MATLAB. Relevant attributes were selected using the coefficient correlation method, followed by the application of the Bat algorithm and Extreme Learning parameters to assess recurrence. Deep learning activation functions such as sigmoid, sine, and Tanh were tested at different training stages, with Tanh achieving the highest accuracy of 93.75%.

For error-free breast cancer detection using mammograms, deep learning techniques such as Stack Sparse Autoencoder (SSAE), Sparse Autoencoder (SAE), and Convolutional Neural Networks (CNN) were employed. The preprocessing stage involved noise removal, background elimination, and artifact suppression. ROI segmentation was then applied to detect tumors by removing the pectoral muscle. Finally, a deep neural network was constructed for training and testing, with input generation leading to the final detection phase. The dataset consisted of 322 digitized mammogram images from the MIAS database. A confusion matrix was used to evaluate accuracy, sensitivity, and precision, revealing that SSAE achieved 98.9% accuracy, SAE 98.5%, and CNN 97%. Among these, SSAE demonstrated the highest accuracy for early-stage breast cancer detection.

7. Overview of study selection

The study organizes research papers on breast cancer prediction using machine learning based on publication year. Table 1 categorizes the reviewed papers into journals and conference proceedings, with the final column showing the total number of papers selected each year. Most of the reviewed studies were published between 2016 and 2023, with the highest number in 2013, highlighting recent advancements in machine learning and deep learning techniques for breast cancer prediction.

Figure 4 presents a bar plot illustrating the yearly distribution of selected research articles. The graph emphasizes the focus on recent studies, beginning with a combined count of papers before 2016, followed by a year-wise breakdown from 2016 onward. Journal and conference papers are distinctly marked using different colours. The highest number of reviewed papers is from 2023, demonstrating the study's objective of identifying the most recent and relevant methods for breast cancer.

Table 1. Year wise Number of Journal and Conference Papers

Year	Journal Paper	Conference Paper	Total
2016	6	1	7
2017	1	3	4
2018	4	7	11
2019	19	6	25
2020	5	1	6
2021	97	134	231
2022	135	154	289
2023	189	267	456
2016-2023	426	573	999

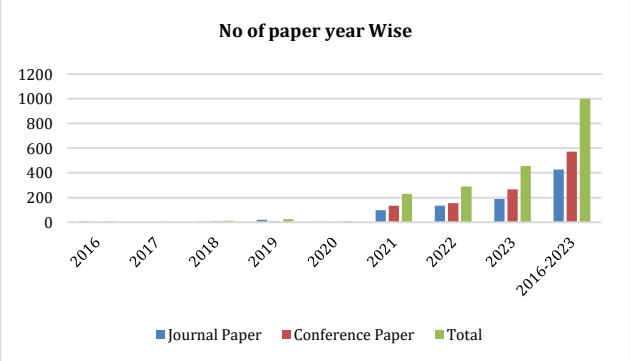

Figure 1. Number of papers per year

Table 2. Comparative Review of Machine Learning techniques for Breast Cancer Prediction

Algorithm	Tool	Dataset	Number of Attribute	Data types	Pre-Processing	Data Processing Method	Evaluation Method	Validation Technique	Accuracy	R#
Logistic Regression	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	99.26%	
Support Vector Machine (SVM)	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	
K – Nearest Neighbour(KNN)	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	
Random Forest	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	
Decision Tree (DT)	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	
Naïve Baise (NB)	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	
Artificial Neural Network (ANN)	Python (Jupyter Notebook)	Wisconsin Diagnostic Breast Cancer Dataset	Not specified	Numerical	Feature selection to improve data set	Model training and optimization	Accuracy, Precision Recall, F1 Score, ROC Curve	Not Specified	Not Specified	

8. Discussion

This research provides a comprehensive summary of various machine learning, deep learning, and data mining algorithms used for breast cancer prediction. Table 2 presents a comparative overview of different machine learning techniques based on tools, data sources, data types, preprocessing methods, evaluation methods, validation techniques, and accuracy levels across different scenarios.

Table 3 highlights the accuracy levels of key machine learning techniques, while Table 4 outlines the advantages and disadvantages of significant studies reviewed. Breast cancer prediction is categorized into three main approaches: Machine Learning Techniques, Ensemble Techniques, and Deep Learning Techniques. Table 5 summarizes the number of reviewed papers corresponding to each of these approaches.

Additionally, the study examines five types of algorithms: Non-Linear Algorithms, Ensemble Algorithms, Deep Learning Algorithms, a combination of Linear and Non-Linear Algorithms, and a combination of Non-Linear and Ensemble Algorithms. Table 6 provides a summary of the number of research papers analyzed for breast cancer prediction based on these algorithm categories.

Each machine learning and deep learning technique performs differently depending on the dataset and conditions. After a comparative analysis, it was observed that the Support Vector Machine (SVM) algorithm is the most effective for breast cancer prediction. Several researchers [2], have analyzed prediction algorithms using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, consistently showing that SVM achieves higher accuracy than other machine learning models.

Sara Al Ghunaim *et al*. compared machine learning algorithms using two tools, Weka and Spark, and found that SVM outperformed other methods, achieving an accuracy of 98.03% on Weka and 99.68% on Spark. Additionally, for breast cancer prediction using deep learning techniques, a study [60] utilized the MIAS database and applied CNN, SAE, and SSAE, with SSAE achieving the highest accuracy of 98.9%.

In total, different researchers [2], [52] have reviewed 24 different algorithms for breast cancer prediction. Despite advancements in machine learning and deep learning, the accuracy of these algorithms varies depending on the dataset used. Therefore, further research is needed to develop more advanced models that can improve prediction accuracy and generalize across different datasets.

Conclusion

This paper presents a comprehensive review of various machine learning, deep learning, and data mining algorithms used for breast cancer prediction. The primary objective is to identify the most effective algorithm for accurately predicting the occurrence of breast cancer. The review aims to provide insights into

previous studies on machine learning algorithms applied to breast cancer prediction, serving as a foundational resource for beginners looking to understand and analyze these techniques as a stepping stone to deep learning.

The review begins with an overview of breast cancer types, symptoms, and causes, based on an analysis of fourteen research papers. It then explores major machine learning, ensemble, and deep learning techniques, offering a detailed examination of the algorithms commonly used for breast cancer prediction.

Despite advancements in predictive techniques, several challenges remain. Future research should address the issue of limited datasets by employing data augmentation techniques. Additionally, the imbalance between positive and negative cases must be considered, as it can introduce bias in predictions. Another critical challenge is the disproportionate number of breast cancer images compared to affected tissue patches, which impacts accurate diagnosis and prediction. Addressing these issues will enhance the reliability and effectiveness of breast cancer prediction models.

Acknowledgments

I sincerely thank Dr. Achyut Pandey, Head of the Physics Department, Govt. TRS College Rewa (M.P.) for his invaluable supervision and guidance throughout this work. I also acknowledge Dr. Chandra Shekhar Gautam for his insightful contributions as a co-author. Their support was instrumental in the completion of this review paper.

References

[1] Y.-S. Sun *et al*., "Risk factors and preventions of breast cancer," International journal of biological sciences, vol. 13, no. 11, p. 1387, 2017.

[2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–5, IEEE.

[3] Y. Lu, J. Y. Li, Y. T. Su, and A. A. Liu, "A review of breast cancer detection in medical images," in 2018 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4, IEEE.

[4] F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier," in 2013 13th International Conference on Intellient Systems Design and Applications, pp. 121–125, IEEE.

[5] R. Hou *et al*., "Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation," IEEE Transactions on Biomedical Engineering, 2019.

[6] R. Chaudhury, R. Iyer, K. K. Iychettira, and A. Sreedevi, "Diagnosis of invasive ductal carcinoma using image processing techniques," in 2011 International Conference on Image Information Processing, pp. 1– 6, IEEE.

[7] S. Pervez and H. Khan, "Infiltrating ductal carcinoma breast with central necrosis closely mimicking ductal carcinoma in situ (comedo type): a case series," Journal of medical case reports, vol. 1, no. 1, p. 83, 2007.

[8] D. L. Page, W. D. Dupont, L. W. Rogers, and M. Landenberger, "Intraductal carcinoma of the breast: follow up after biopsy only," MDPI Cancers, vol. 49, no. 4, pp. 751–758, 1982.

[9] B. Tuck, F. P. O'Malley, H. Singhal, and K. S. Tonkin, "Osteopontin and p53 expression are associated with tumor progression in a case of synchronous, bilateral, invasive mammary carcinomas," Archives of pathology and laboratory medicine, vol. 121, no. 6, p. 578, 1997.

[10] Lee *et al*., "Efficacy of the multidisciplinary tumor board conference in gynecologic oncology: a prospective study," Medicine, vol. 96, no. 48, 2017.

[11] S. Masciari *et al*., "Germline e-cadherin mutations in familial lobular breast cancer," Journal of medical genetics, vol. 44, no. 11, pp. 726–731, 2007.

[12] Memis *et al*., "Mucinous (colloid) breast cancer: mammographic and us features with histologic correlation," European journal of radiology, vol. 35, no. 1, pp. 39–43, 2000.

[13] Gradilone *et al*., "Circulating tumor cells (ctcs) in metastatic breast cancer (mbc): prognosis, drug resistance and phenotypic characterization," Annals of Oncology, vol. 22, no. 1, pp. 86–92, 2010.

[14] F. M. Robertson *et al*., "Inflammatory breast cancer: the disease, the biology, the treatment," CA: a cancer journal for clinicians, vol. 60, no. 6, pp. 351–375, 2010.

[15] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," International Journal of Information Technology, pp. 1–15, 2020.

[16] D. Delen, "Analysis of cancer data: a data mining approach," Expert Systems, vol. 26, no. 1, pp. 100–112, 2009.

[17] M. Shahbaz, S. Faruq, M. Shaheen, and S. A. Masood, "Cancer diagnosis using data mining technology," Life Science Journal, vol. 9, no. 1, pp. 308–313, 2012.

[18] Reddy, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," ICT Express, 2020.

[19] Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol," Journal of Public Health Research, vol. 8, no. 3, 2019.

[20] S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review," IOSR Journal of Dental and Medical Sciences (IOSR-JDMS).

[21] H. Witten and E. Frank, Data mining: practical machine learning tools and techniques with Java implementations, vol. 31 of Acm Sigmod Record. Elsevier, 2005.

[22] D. L. Olson and D. Delen, Advanced data mining techniques. Springer Science and Business Media, 2008.

[23] L. Li *et al*., "Research on machine learning algorithms and feature extraction for time series," in 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–5, IEEE.

[24] M. Bishop, Pattern recognition and machine learning. Springer, 2006.

[25] L. Tuggener *et al*., "Automated machine learning in practice: state of the art and recent results," in 2019 6th Swiss Conference on Data Science (SDS), pp. 31–36, IEEE.

[26] Dey, "Machine learning algorithms: a review," International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1174– 1179, 2016.
[27] M. D. Ganggayah *et al*., "Predicting factors for survival of breast cancer patients using machine learning techniques," BMC medical informatics and making, decision, vol. 19, no. 1, p. 48, 2019.
[28] Y. Uzun and G. Tezel, "Rule learning with machine learning algorithms and artificial neural networks," Journal of Seljuk University Natural and Applied Science, vol. 1, no. 2, 2012.
[29] P. Singhal and S. Pareek, "Artificial neural network for prediction of breast cancer," in 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 464–468, IEEE, 2018.
[30] Q. Dai, S.-H. Xu, and X. Li, "Parallel process neural networks and its application in the predication of sunspot number series," in 2009 Fifth International Conference on Natural Computation, vol. 1, pp. 237–241, IEEE.
[31] W. K. Tsai, A. Parlos, and B. Fernandez, "Asdm-a novel neural network model based on sparse distributed memory," in 1990 IJCNN International Joint Conference on Neural Networks, pp. 771–776, IEEE.
[32] H. Tran, "A survey of machine learning and data mining techniques used in multimedia system," September 2019.
[33] S. D. Borde and K. R. Joshi, "Enhanced signal detection slgorithm using trained neural network for cognitive radio receiver," International Journal of Electrical and Computer Engineering, vol. 9, no. 1, p. 323, 2019.
[34] P. Utomo, A. Kardiana, and R. Yuliwulandari, "Breast cancer diagnosis using artificial neural networks with extreme learning techniques," International Journal of Advanced Research in Artificial Intelligence, vol. 3, no. 7, pp. 10–14, 2014.
[35] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," The journal of educational research, vol. 96, no. 1, pp. 3–14, 2002.
[36] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbour (knn) approach for predicting economic events: Theoretical background," International Journal of Engineering Research and Applications, vol. 3, no. 5, pp. 605–610, 2013.
[37] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 2094–2097, 2016.
[38] M. Mahmood *et al*., "An improved cart decision tree for datasets with irrelevant feature," in International Conference on Swarm, Evolutionary, and Memetic Computing, pp. 539–549, Springer.
[39] Budiman, A. H. Kridalaksana, M. Wati, *et al*., "Performance of decision tree c4. 5 algorithm in student academic evaluation," pp. 380– 389, 2017.
[40] R. Pandya and J. Pandya, "C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning," International Journal of Computer Applications, vol. 117, no. 16, pp. 18–21, 2015.
[41] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," Shanghai Archives of Psychiatry, vol. 27, no. 2, p. 130, 2015.
[42] W.Wu, S. Nagarajan, and Z. Chen, "Bayesian machine learning: Eegmeg signal processing measurements," IEEE Signal Processing Magazine, vol. 33, no. 1, pp. 14–36, 2015.
[43] A. Ibrahim, A. I. Hashad, and N. E. M. Shawky, A Comparison of Open Source Data Mining Tools for Breast Cancer Classification, pp. 636–651. IGI Global, 2017.
[44] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in Advanced Course on Artificial Intelligence, pp. 249– 257, Springer, 2005.
[45] Y. Yang, J. Li, and Y. Yang, "The research of the fast svm classifier method," in 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 121–124, IEEE.
[46] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5– 32, 2001.
[47] T. O. Ayodele, "Types of machine learning algorithms," New Advances in Machine Learning, pp. 19–48, 2010.
[48] Y. Li and H. Wu, "A clustering method based on k-means algorithm," Physics Procedia, vol. 25, pp. 1104–1109, 2012.
[49] C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," Computers and Geosciences, vol. 10, no. 2-3, pp. 191–203, 1984
[50] A review of big data environment, tools and challenges Chandra Shekhar Gautam1 and Dr.Prabhat Pandey2 Journal of Emerging Technologies and Innovative Research, Volume 6, Year 2019, Pages 569-575
[51] An improving query optimization process in Hadoop MapReduce using ACO-Genetic algorithm and HDFS map reduce Technique Chandra Shekhar Gautam1 and Dr.Prabhat Pandey2 International Journal of Current Engineering and Technology, Volume 13, Year 2023,
[52] A review on genetic algorithm models for hadoop mapreduce in big data Chandra Shekhar Gautam1 and Dr.Prabhat Pandey2 International Journal of Recent Scientific Research, Volume 13, Year 2022, Pages 771-775
[53] Predicting heart disease using machine learning classification technique, Miss Sanjana Chaudhary,Chandra Shekhar Gautam1 and Dr. Akhilesh A Waoo2,Journal of the Maharaja Sayajirao University of Baroda,Volume 10,Year 2024.