Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

Machine Learning–Driven Classification Techniques for Early-Stage Breast Cancer

[1] T. Vinothini, [2] Dr. V. Arun, [3] Dr. S. Meenakshi Sundaram

[1] Department of Computer Science and Engineering, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India.

[2] Department of Electronics and Communication Engineering, Anna University Regional Campus, Madurai, Tamil Nadu, India.

[3] Department of Computer Science and Engineering, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India.

(Received: 27 September 2025 Revised: 05 october 2025 Accepted:01 November 2025)

KEYWORDS

Random Forest, Decision Tree, Machine Learning, Logistic Regression, Breast Cancer, Prediction, and Detection.

ABSTRACT:

One of the most prevalent cancers in women worldwide is breast cancer. In addition to developing standardized procedures for data collecting and analysis of breast cancer using mammograms and other medical imaging data, more research is needed to overcome the difficulties and constraints of the conventional detection process. In recent years, a number of studies have been conducted to develop and evaluate machine learning models for breast cancer diagnosis, recurrence prediction, and treatment planning. Even before a tumor appears on a mammogram, as is typically the case, machine learning algorithms may be trained to identify minute alterations in breast tissue that could be signs of cancer. Higher accuracy and specificity have been achieved in the detection of breast cancer with the use of machine learning techniques, which makes it a useful tool for supporting doctors in making clinical decisions. Overall, the results suggest that further research may be necessary to develop standardized procedures for data collecting and analysis as well as to address the issues and constraints of machine learning in the detection of breast cancer.

1. Introduction

Despite significant advancements in breast cancer screening and treatment, early detection remains critical for improving prognosis and survival rates. However, mammography has limitations, including its low sensitivity, it can also boost the accuracy and efficiency of existing screening methods. It allows for the identification of patterns and relationships within large amounts of data that may be difficult for human experts to identify. These limitations have alternative approaches to breast cancer screening, including the use of ML techniques.

Machine learning has widely used in medical imaging analysis, including breast cancer detection using mammograms and other medical images. ML models can be trained to recognize specific patterns in medical images that are indicative of breast cancer, allowing for earlier detection and more accurate diagnosis. Machine learning algorithms can also be trained to recognize subtle changes in breast tissue that may be indicative of cancer, even before a tumor is visible on a mammogram. by selecting the most relevant features from the input data. The use of ML techniques has demonstrated high accuracy and specificity in breast cancer detection, making it a promising tool for aiding physicians in clinical decision-making.

The use of ML techniques has demonstrated high accuracy and specificity in breast cancer detection, making it a promising tool for aiding physicians in clinical decision- making. ML models can analyze large amounts of data in a short amount of time and provide accurate predictions, reducing the need for unnecessary diagnostic tests and surgeries. Furthermore, ML models can help identify subgroups of patients with a high risk of breast cancer, enabling personalized screening and treatment plans. One major challenge is the lack of standardized protocols for data acquisition and analysis, which can lead to inconsistent results across studies. Additionally, the lack of diverse and representative datasets can limit the generalizability of ML models, leading to bias and errors.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

The use of machine learning in detection and diagnosis can improve the accuracy and efficiency of existing screening methods. It allows for the identification of patterns and relationships within large amounts of data that may be difficult for human experts to discern. Machine learning algorithms can also be trained to recognize subtle changes in breast tissue that may be indicative of cancer, even before a tumor is visible on a mammogram. In recent years, ML models can be trained to recognize specific patterns in medical images that are indicative of breast cancer, allowing for earlier detection and more accurate diagnosis. These studies have shown promising results, with machine learning algorithms outperforming traditional screening methods in some cases. However, there are still challenges to be addressed in the development and implementation of these algorithms, such as the need for high-quality and diverse datasets for training and validation.

In conclusion, early detection using ML holds great potential for improving the prognosis and survival rates of patients with this disease. The use of ML models in diagnosis and treatment planning can lead to earlier detection, personalized treatment plans, and reduced healthcare costs. However, further research is needed to address the challenges and limitations of ML in breast cancer detection and to develop standardized protocols for data acquisition and analysis.

2. Literature Review

In this section, we are going to review the researches on the use of machine learning in the field of breast cancer detection and diagnosis.

David A. Omondiagbe, Shanmugam Veeramani, and Amandeep S. Sidhu [1] They suggested that integrating machine learning models with clinical data can provide a more comprehensive approaches for diagnosis and treatment. The study evaluated the performance of several random forest. The algorithm can handle both categorical and continuous variables, and can also handle missing data. It is also robust to outliers and noise in the data, and can handle nonlinear relationships between the input variables and the output variable. SVM had the highest sensitivity, while Naive Bayes had the highest specificity. The study also found that feature selection improved the performance of the models.

Arpita Joshi and Dr. Ashish Mehta [2] Based on the findings, the authors recommended the use of Random Forest and SVM for diagnosis using ML. The study concluded that ML algorithms with clinical data can provide a more comprehensive approach for breast cancer diagnosis and treatment diagnosis, and further research is needed to explore the use of these techniques in clinical practice.

S. Karthik, R. Srinivasa Perumal, and P. V. S. S. R. Chandra Mouli [3] The study evaluated performance of three DNN models, including AlexNet, GoogleNet, and VGG16, The authors utilized mammographic data for detection and diagnosis. The studies were categorized based on their methodology, including machine learning, deep learning, feature extraction, and ensemble methods. SVM had the highest sensitivity, while Naive Bayes had the highest specificity. The result VGG16 model had highest accuracy, sensitivity, AUC-ROC, followed by GoogleNet and AlexNet. The study also found that increasing the number of layers and filters in the models improved their performance.

Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong [4], The study showed SVM had the highest sensitivity, while Naive Bayes had the highest specificity. The results showed that the RBM and backpropagation model had a higher accuracy, compared to backpropagation model alone. The study also found that increasing RBM improved performance. The input features are split into nodes based on their importance in predicting the outcome variable. The decision tree algorithm uses a top-down approach to split the features into nodes, with each node representing a decision rule.

Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei and Tianfu Wang [5], The authors reviewed a total of 92 studies that utilized mammographic data for detection and diagnosis. The studies were categorized based on their methodology, including machine learning, deep learning, feature extraction, and ensemble methods. the lack of diverse and representative datasets can limit the generalizability of ML models, leading to bias and errors. The authors also found that combining different techniques can further improve the accuracy and efficiency of breast cancer diagnosis.

Saleem Z. Ramadan [6], The author reviewed a total of 55 studies that utilized CAD for breast cancer detection using mammogram. The study was categorized

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

methodology, including machine learning, deep learning, feature extraction, and ensemble methods. The studies were categorized based on their methodology, including machine learning, deep learning, feature extraction, and ensemble methods. the lack of diverse and representative datasets can limit the generalizability of ML models, leading to bias and errors. The author also found that combining different techniques can further improve the accuracy and efficiency of breast cancer diagnosis.

M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath and M. A. Bamiah [7], The study evaluated performance of several ML can also be trained to recognize subtle changes in breast tissue that may be indicative of cancer, even before a tumor is visible on a mammogram. In recent years, ML models can be trained to recognize specific patterns in medical images that are indicative of breast cancer, allowing for earlier detection and more accurate diagnosis. The study also found that using a combination of features, including age, family history, and laboratory test results, improved the performance of the model.

Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza and Nikahat Kazi [8] The results showed ANN model had highest accuracy in all terms of ML requirement compared to the other models. The study also found that using a combination of features, including age, tumor size, and hormone receptor status, improved the model. Based on findings, authors recommended use of artificial neural networks for breast cancer diagnosis and recurrence prediction. They also suggested that integrating machine learning models with clinical data can provide a more comprehensive approach for breast cancer diagnosis and treatment.

Shubham Sharma, Archit Aggarwal and Tanupriya Choudhury [9] The study evaluated the performance of several random forest. The algorithm can handle both categorical and continuous variables, and can also handle missing data. It is also robust to outliers and noise in the data, and can handle nonlinear relationships between the input variables and the output variable. In terms of performance, Random Forest has been shown to outperform many other machine learning algorithms on a variety of datasets, including the Breast Cancer Wisconsin dataset.

Ram MurtiRawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal [10] The studies were categorized based on their methodology, including machine learning, deep learning, feature extraction, and ensemble methods. the lack of diverse and representative datasets can limit the generalizability of ML. The the input features are split into nodes based on their importance in predicting the outcome variable. The decision tree algorithm uses a top-down approach to split the features into nodes, with each node representing a decision rule.

M. Elgedawy [11] The study evaluated performance of three DNN models, including AlexNet, GoogLeNet, and VGG16, Random Forest has been shown to outperform many other machine learning algorithms on a variety of datasets, including the Breast Cancer Wisconsin dataset. SVM had the highest sensitivity, while Naive Bayes had the highest specificity. The results showed that the RBM and backpropagation model had a higher accuracy, compared to backpropagation model alone. The study also found that increasing RBM improved performance.

T. K. Avramov and D. Si [12] The performance of each model is evaluated using several metrics, including accuracy, sensitivity, and specificity. The results show that PCA and MIFS are the most effective feature reduction methods, while the SVM model is the most accurate ML model for diagnosis. The authors also compare their results with previous studies and demonstrate that their approach outperforms other methods in terms of accuracy. It allows for the identification of patterns and relationships within large amounts of data that may be difficult for human experts to discern.

3. Methodology

A. Dataset

We will classify tumours using the Breast Cancer Wisconsin (Diagnostic) Dataset, which consists of 570 rows and 32 columns. One of the main obstacles in tumour detection is accurately differentiating between malignant (cancerous) and benign (non-cancerous) tumours. The dataset has been prepared by extracting features from each image into csv file.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

B. Data Visualization

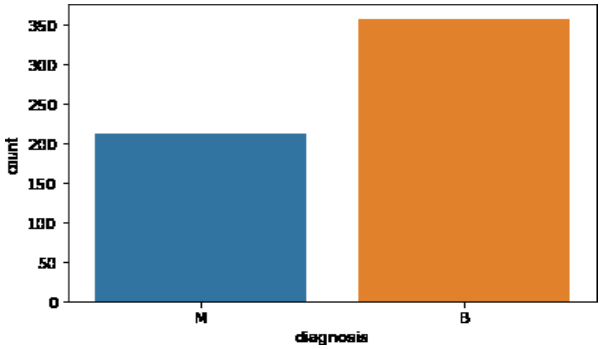

Fig 1: Number of malignant and benign data in dataset

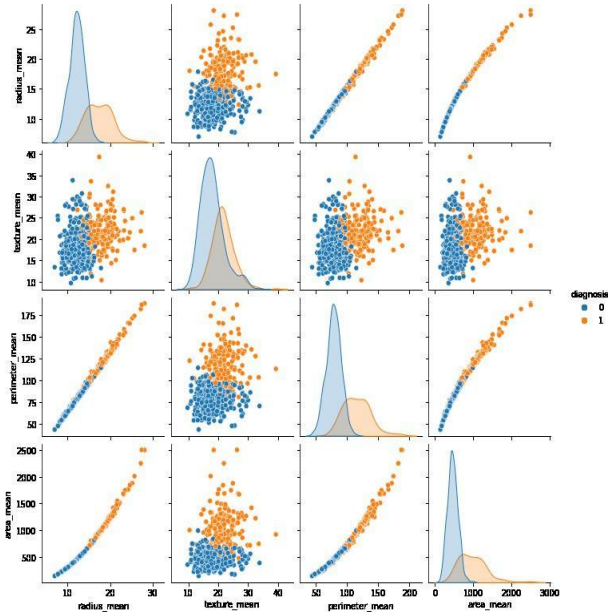

Fig 2: malignant and benign tumor data distributed in two classes

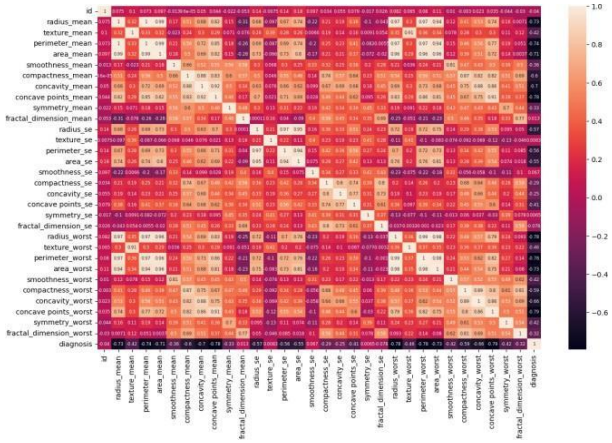

Fig 3: Correlation of each feature with the target

C. Algorithms

Logistic Regression, in machine learning, it is common to use this as a classification algorithm. It is a statistical model that is used to predict the probability of occurrence of a binary outcome. In the case of our project need, logistic regression is used to predict the probability of a tumor being malignant or benign based on feature extracted from images and data collected in csv file. The logistic regression algorithm uses this function to predict the probability of the binary outcome based on the input features. It is a type of ensemble learning method that uses multiple decision trees to make predictions. The algorithm works by creating a large number of decision trees and combining their predictions to make a final prediction. The algorithm works by analyzing a set of input variables, such as the size and shape of the tumor, and using these variables to construct a decision tree. The decision tree is built by recursively splitting the data into smaller and smaller subsets based on the input variables until a stopping criterion is met. This is done using an optimization algorithm called gradient descent. Random Forest algorithm can be attributed to its ability to handle high-dimensional datasets and reduce overfitting. It also has the advantage of combining multiple decision trees to improve the prediction accuracy. The logistic regression algorithm learns the parameters of the linear function that maps the input features to the output probability. The decision tree is built by recursively splitting the data into smaller and smaller subsets based on the input variables until a stopping criterion is met the testing set performing in terms of predicting the correct diagnosis of the tumor. In the case of breast cancer detection, the logistic regression algorithm has shown promising results.

Decision Tree algorithm mostly used for regression and classification problems. It is a tree-based model that is used to make decisions based on a set of input features. In the case of breast cancer detection, decision predict the tumor based on the features from mammogram images. In decision tree, the input features are split into nodes based on their importance in predicting the outcome variable. The decision tree algorithm uses a top-down approach to split the features into nodes, with each node representing a decision rule. The algorithm continues to split the nodes until it reaches a point where further splitting does not affect. The decision tree algorithm will be trained on training set using the features of

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

mammogram images. The objective is to create a tree that can accurately classify tumors into benign or malignant categories. The decision tree algorithm selects the most important features are then used to create a tree structure that consists of decision nodes and leaf nodes. The decision nodes are based on the feature values and determine the next node in the tree, while the leaf nodes represent the final decision or prediction. Once the decision tree is created, it is evaluated on the test set to determine its accuracy in classifying tumors and avoid false positives and false negatives. Decision trees have several advantages for breast cancer detection. They are easy to understand and interpret, making it possible for medical professionals to interpret the results and make informed decisions about patient care. These trees can also handle numerical and categorical data, making them versatile for a wide range of applications. However, they are prone to overfitting and may not perform well on datasets with noisy or irrelevant features.

Random Forest, usage of this as a machine learning algorithm is prevalent in various tasks such as classification, regression, and others. It is a type of ensemble learning method that uses multiple decision trees to make predictions. The algorithm works by creating a large number of decision trees and combining their predictions to make a final prediction. Each tree is built using a subset of data and features, which helps to reduce overfitting and increase accuracy. It is often used in combination with other algorithms to improve performance even further, such as in a stacked ensemble approach. The algorithm works by analyzing a set of input variables, such as the size and shape of the tumor, and using these variables to construct a decision tree. The decision tree algorithm selects the most important features are then used to create a tree structure that consists of decision nodes and leaf nodes, the stopping criterion is typically based on the purity of the subsets, or the degree to which all members of a subset belong to the same class. Random Forest algorithm can be attributed to its ability to handle high-dimensional datasets and reduce overfitting. It also has the advantage of combining multiple decision trees to improve the prediction accuracy. This is typically done by taking the majority vote of the predictions, although other methods can also be used. The result is a highly accurate prediction of tumor is benign or malignant. One of the advantages of Random Forest is its ability to handle large datasets with many input variables. The algorithm can handle both categorical and continuous variables, and can also handle missing data. It is also robust to outliers and noise in the data, and can handle nonlinear relationships between the input variables and the output variable. In terms of performance, Random Forest has been shown to outperform many other machine learning algorithms on a variety of datasets, including the Breast Cancer Wisconsin dataset. It is often used in combination with other algorithms to improve performance even further, such as in a stacked ensemble approach.

```
Logistic Regression
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        90
           1       0.93      0.94      0.93        53

    accuracy                           0.95       143
   macro avg       0.95      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143

Accuracy :  0.951048951048951
Decision Tree
              precision    recall  f1-score   support

           0       0.98      0.92      0.95        90
           1       0.88      0.96      0.92        53

    accuracy                           0.94       143
   macro avg       0.93      0.94      0.93       143
weighted avg       0.94      0.94      0.94       143

Accuracy :  0.9370629370629371
Random Forest
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

Accuracy :  0.965034965034965
```

Fig 4: Classification report of all algorithms

4. Conclusion

In conclusion, this study has explored the application of machine learning algorithms for breast cancer detection using the Breast Cancer Wisconsin (Diagnostic) Dataset. The study considered three algorithms, Random Forest, Decision Tree and Logistic Regression. The showed results of all three algorithms were effective in predicting the diagnosis of a tumour, with Random Forest achieving the highest accuracy.

Logistic Regression and Decision Tree both performed well with accuracies of 95.1% and 93.7% respectively. However, Random Forest was better the other two algorithms with an accuracy of 96.5%. This indicates that

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 2412-2417 | ISSN:2251-6727

Random Forest is better suited for breast cancer detection tasks and can provide more accurate results. The high accuracy of the Random Forest algorithm can be attributed to its ability to handle high-dimensional datasets and reduce overfitting. It also has the advantage of combining multiple decision trees to improve the prediction accuracy.

This makes Random Forest a powerful tool for breast cancer detection. In conclusion, this paper highlights the importance of machine learning in medical diagnosis and provides insights into the performance of different algorithms for breast cancer detection. The results suggest that Random Forest can be an effective tool for medical professionals in diagnosing breast cancer and can potentially lead to earlier detection and improved patient outcomes.

Acknowledgment:

We would like to thank Dr.S.Meenakshi Sundaram for his expertise and guidance as a Professor of Computer Science and Engineering at AAA College of Engineering and Technology Sivakasi, India have been instrumental in shaping the direction and quality of our publication.

References

1. David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu “Machine Learning Classification Techniques for Breast Cancer Diagnosis” (2019).
2. Arpita Joshi and Dr. Ashish Mehta “Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer” (2017).
3. S. Karthik, R. Srinivasa Perumal and P. V. S. S. R. Chandra Mouli Breast Cancer Classification Using Deep Neural Networks (2019).
4. Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation. (2018).
5. Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei and Tianfu Wang "Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review" (2019).
6. Saleem Z. Ramadan Methods Used in Computer-Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review (2020).
7. M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath and M. A. Bamiah "Early Detection of Breast Cancer Using Machine Learning Techniques".
8. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza and Nikahat Kazi Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques (2015).
9. Shubham Sharma, Archit Aggarwal and Tanupriya Choudhury "Breast Cancer Detection Using Machine Learning Algorithms" (2018).
10. Ram Murti Rawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning" (2020)
11. M. Elgedawy, “Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes,” International Journal of Engineering and Computer Science, 2017, vol. 6, no. 1, pp. 19884- 19889.
12. T. K. Avramov and D. Si, “Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis,” Proc. Int. Conf. Comput. Data Anal. - ICCDA ’17, pp. 69–74, 2017.
13. M. Rmili, and A. El, “A Combined Approach for Breast Cancer Detection in Mammogram,” 2016 13th International Conference on Computer Graphics, Imaging and Visualization, pp. 350–353, 2016.
14. Bhola A, Tiwari AK, "Machine learning based approaches for cancer classificaition using gene expression data", MLAIJ, vol. 2, December 2015.
15. Agarap AFM, "On breast cancer detection: an artificial intelligence learing algorithms on the Wisconsin diagnostic dataset", ICMLSC 2018, March 2018.