

A Machine Learning Approach for Early Detection of Breast Cancer: Performance Evaluation and Analysis

Tanusree Saha, Sarmistha Santra

Abstract: Breast cancer is one of the most prevalent cancers affecting women globally and continues to be a leading cause of cancer-related deaths. Early and accurate diagnosis significantly improves survival rates, but conventional diagnostic techniques are often time-consuming, costly, and prone to subjective interpretation. To address these challenges, this study focuses on developing an efficient breast cancer detection system using machine learning (ML) and deep learning (DL) algorithms. By leveraging Convolutional Neural Networks (CNNs) and several traditional ML models, the system aims to classify breast cancer efficiently based on imaging data. Utilising the Wisconsin Breast Cancer Dataset (WBCD), the project evaluates the performance of various machine learning models, including SVM, KNN, Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, AdaBoost, and XGBoost, based on accuracy, precision, and robustness. The outcomes indicate that machine learning can significantly enhance early detection efforts, providing clinicians with reliable decision-support tools.

Keywords: Breast Cancer, Machine Learning, Deep Learning, CNN, SVM, KNN, Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, AdaBoost, Boost

Abbreviations:

DL: Deep Learning
ML: Machine Learning
WBCD: Wisconsin Breast Cancer Dataset
CNNs: Convolutional Neural Networks
MRI: Magnetic Resonance Imaging
CAD: Computer-Aided Diagnostic
AI: Artificial Intelligence
SVM: Support Vector Machines
KNN: K-Nearest Neighbours
CNNs: Convolutional Neural Networks
RF: Random Forest
DT: Decision Trees
NB: Naïve Bayes
LR: Logistic Regression
FNA: Fine-Needle Aspirates
WHO: World Health Organisation

Manuscript received on 28 April 2025 | First Revised Manuscript received on 21 May 2025 | Second Revised Manuscript received on 18 October 2025 | Manuscript Accepted on 15 November 2025 | Manuscript published on 30 November 2025.

*Correspondence Author(s)

Tanusree Saha*, Department of Information Technology, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: tanusreesah@gmail.com, ORCID ID: [0000-0001-6011-1237](https://orcid.org/0000-0001-6011-1237)

Sarmistha Santra, Student, Department of Computer Applications, JIS College of Engineering, Kalyani (West Bengal), India. Email ID: sarmisthasan6@gmail.com

© The Authors. Published by Lattice Science Publication (LSP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

I. INTRODUCTION

Breast cancer remains the most common malignancy and a leading cause of cancer-related deaths among women worldwide. According to the Centres for Disease Control and Prevention (CDC) and the World Health Organisation (WHO), breast cancer accounts for approximately 2.1 million new cases globally each year. Despite advances in medical research and treatment options, survival rates for breast cancer vary greatly depending on several factors, including the cancer type, the stage at which the disease is diagnosed, and the availability of healthcare services. Breast cancer originates when abnormal cells in the breast begin to grow uncontrollably. Most breast cancers develop in the ducts (ductal carcinomas) or lobules (lobular carcinomas) that carry milk to the nipple. However, malignancies can also arise from the fibrous connective tissue or the fatty tissue within the breast. If left untreated, cancerous cells may invade neighbouring healthy tissues and spread to lymph nodes, most commonly under the arms, and potentially metastasise to other parts of the body. Early detection is pivotal in improving the prognosis and survival rates for breast cancer patients. When diagnosed at an early stage, breast cancer is highly treatable, often requiring less aggressive therapies and offering better survival outcomes. Unfortunately, many individuals are diagnosed at later stages due to the asymptomatic nature of early-stage breast cancer or due to a lack of access to regular screening programs. This delay in diagnosis often leads to a more complicated disease course and reduces the effectiveness of available treatment options. Conventional methods for detecting breast cancer include clinical breast examinations, mammography, ultrasound imaging, magnetic resonance imaging (MRI), and tissue biopsy. While these methods have proven effective, they have certain limitations, such as being time-consuming, expensive, and occasionally prone to human error or subjectivity in interpretation. Moreover, the accuracy of these diagnostic techniques can be influenced by factors such as the density of breast tissue and the radiologist's or pathologist's skill level. In response to these challenges, there has been an increasing focus on developing computer-aided diagnostic (CAD) systems that leverage advances in machine learning (ML) and artificial intelligence (AI). Machine learning algorithms, particularly in the realm of deep learning, offer the potential to automate and enhance breast cancer detection processes by learning complex patterns from large datasets with minimal human

intervention. These models can analyze mammographic images, histopathological slides, or even clinical data to assist healthcare professionals in making more accurate and timely diagnoses. The lack of robust and exact prognosis models further complicates treatment planning for breast cancer patients. Without reliable predictive tools, clinicians may struggle to tailor treatments effectively, which could potentially impact patient survival and quality of life. Hence, there is a pressing need to develop advanced diagnostic systems that minimize error rates and maximize predictive accuracy. Machine learning techniques, such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), Naïve Bayes (NB), AdaBoost, and XGBoost, have demonstrated promising results in classification tasks, including medical diagnoses. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have further enhanced performance by enabling the automatic extraction of features and handling large amounts of complex medical data. The objective of this research is to design and evaluate a machine learning-based system for detecting breast cancer, utilising the Wisconsin Breast Cancer Dataset (WBCD). The system aims to compare multiple machine learning algorithms, assess their predictive accuracies, and highlight the potential of deep learning models in providing a reliable and efficient diagnostic aid for breast cancer detection. Early and accurate detection enabled by these models could significantly reduce mortality rates and improve the quality of care for patients worldwide.

II. LITERATURE REVIEW

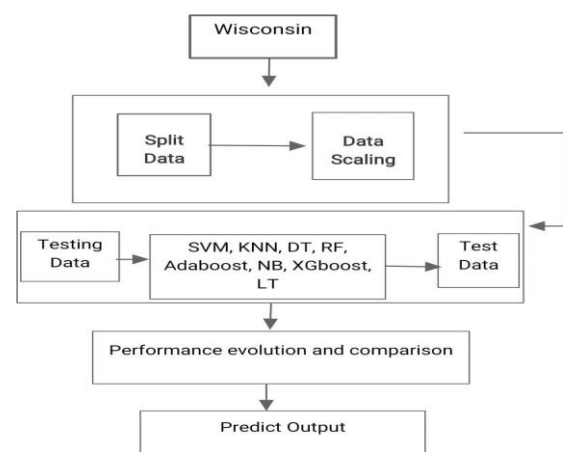
Early research into automated breast cancer detection primarily focused on traditional machine learning models, while recent developments have incorporated advanced deep learning techniques. Several key studies underpin the foundation of this research:

To detect breast cancer in computerised mammograms with varying densities, this research paper [1] developed an effective deep learning model. Three separate feature selection modules were used in this study: recursive feature elimination, univariate feature selection, and the removal of low-variance features. Convolutional Neural Networks (CNNs) are thoroughly covered in Krichen's assessment [2], which covers their fundamentals, architectural development, applications, and cutting-edge methods. Zhou X., Li Y., Gururajan R., Bargshady G [3] propose a 19-layer CNN architecture for differentiating between benign and malignant breast tumours, trained on the BreA-KHis imaging dataset and evaluated using k-fold cross-validation. Similarly, Hamidinekoo et al. (2018) [4] reviewed ensemble learning methods, such as random forests and gradient boosting, emphasising the benefits of combining multiple classifiers to enhance predictive performance. Feature selection plays a crucial role in improving model accuracy, as demonstrated by Hossain and Muhammad (2019) [5], who employed PCA and genetic algorithms in conjunction with SVM classifiers. Their approach highlighted how reducing data dimensionality could improve model training efficiency without compromising accuracy. The importance

of data augmentation techniques in medical imaging was emphasised by Hussain et al. (2020) [6], particularly in the context of ultrasound images. This work demonstrated that preprocessing strategies can significantly impact model generalisation and robustness. Transfer learning, wherein pre-trained CNN models are fine-tuned on domain-specific datasets, was explored by Rahimzadeh and Attar (2019) [7]. Their findings support the notion that leveraging previously learned features can dramatically reduce the amount of training data needed while maintaining high accuracy. Multimodal imaging techniques, which integrate information from mammography, ultrasound, and MRI, were discussed by Qiu et al. (2020) [8], suggesting that combining data sources can lead to improved detection outcomes. Evaluation metrics such as sensitivity, specificity, and AUC were thoroughly addressed by Omer et al. (2019), [9] providing guidelines for robust and meaningful model evaluation. Finally, Jain et al. (2017) [10], provided a comprehensive overview of machine learning techniques, including dimensionality reduction and classification, further highlighting the rapid evolution of computer-aided diagnosis systems.

III. METHODOLOGY

For a thorough assessment of breast cancer detection performance, eight classification models were selected: SVM, KNN, Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, AdaBoost, and XGBoost. These models were chosen due to their proven effectiveness in prior classification tasks and their complementary strengths. Ensemble learning strategies were also explored, focusing on Random Forest (which uses bagging) and stacking classifiers (which combine multiple model outputs). Each model was trained and tested on four different train-test splits: 60%-40%, 70%-30%, 80%-20%, and 90%-10%. Feature scaling was applied selectively to assess its impact on model performance. Additionally, hyperparameter tuning was conducted for each model to further optimise its performance.



[Fig.1: The Proposed Model]

Dataset: The classifiers tested in this work have been trained and evaluated using an enhanced dataset. The WBCD, a popular



dataset available on the UCI Machine Learning Repository website, served as the foundation for developing the upgraded dataset. Thirty traits are present in 569 occurrences of the WBCD. There are now only 17 features in this dataset due to the improved technique that was used. Five feature selection strategies were used to reduce the number of features. The top 17 features with the most significant weight were then chosen, and the remaining features were ignored.

All of the classifiers have been trained and tested using four distinct sets from our updated dataset. These sets are as follows: set 1 is 60% training and 40% testing, set 2 is 70% training and 30% testing, set 3 is 80% training and 20% testing, and set 4 is 90% training and 10% testing. It is well known that using various training sizes yields comprehensive studies with reliable outcomes.

Table-I: Enhanced Dataset

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity
563	20.92	25.09	143.00	1347.0	0.10990	0.22360	0.31740	0.14740	0.2149	0.06879	29.41	179.10	1819.0	0.14070	0.410	0.10990
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	26.40	166.10	2027.0	0.14100	0.21	0.11100
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	38.25	155.00	1731.0	0.11660	0.15	0.09780
566	16.60	28.08	108.30	858.1	0.09455	0.10230	0.09251	0.05302	0.1590	0.05648	34.12	126.70	1124.0	0.11390	0.30	0.09455
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	39.42	184.60	1821.0	0.16500	0.80	0.11780
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	30.37	59.16	268.6	0.08996	0.06	0.05263

IV. SYSTEM DESIGN

The proposed system for breast cancer detection using machine learning employs a structured and modular design, ensuring the effective utilisation of data, accurate model training, and efficient evaluation. The design comprises several critical phases, each integral to achieving a high-performing diagnostic system.

A. Data Acquisition

The system begins with the acquisition of data from the Wisconsin Breast Cancer Dataset (WBCD), available on the UCI Machine Learning Repository. This dataset contains features computed from digitized images of fine-needle aspirates (FNA) of breast masses. It includes 569 instances and 30 features describing characteristics such as radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry of the cell nuclei. Each record is labelled as either benign or malignant, making the dataset suitable for binary classification tasks.

B. Data Preprocessing

Effective data preprocessing is essential to ensure that the machine learning models receive clean, meaningful inputs. This phase includes:

- Handling Missing Values:** The dataset was examined for missing or null values. Any anomalies were addressed through imputation techniques or the removal of corrupted entries.
- Feature Selection:** Given the high dimensionality (30 features), feature selection techniques were applied to reduce redundancy and improve model performance. Highly correlated features (correlation > 95%) were identified and removed to prevent multicollinearity. An improved dataset was created, retaining the 17 most informative features.
- Data Normalization and Standardization:** Feature scaling was employed to normalize the dataset.

Standard Scalar was applied to normalise all feature values into a similar range, which is essential for distance-based models like SVM and KNN.

- Dataset Splitting:** The dataset was split into training and testing sets using four different ratios:

60% training / 40% testing
70% training / 30% testing
80% training / 20% testing
90% training / 10% testing

This multi-split strategy ensured robust performance evaluation across different data availability conditions.

C. Model Development

Multiple supervised learning algorithms were implemented to enable a comprehensive performance comparison:

A. Traditional Machine Learning Models:

- Support Vector Machine (SVM)
- K-Nearest Neighbours (KNN)
- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Naïve Bayes (NB)
- AdaBoost
- XGBoost

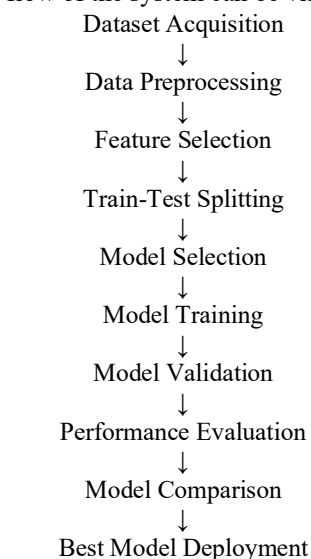
D. Model Training and Validation

- Training Phase:** The models were trained on the selected training sets, using cross-validation where necessary to prevent overfitting.
- Validation Phase:** After training, each model was evaluated against the corresponding test set to assess generalization performance.

Performance was measured using the following metrics: Accuracy, Precision, Recall, and F1-Score. This multi-metric evaluation provided a comprehensive view of model effectiveness, extending beyond simple accuracy.

E. System Architecture Diagram

The overall flow of the system can be visualized as:



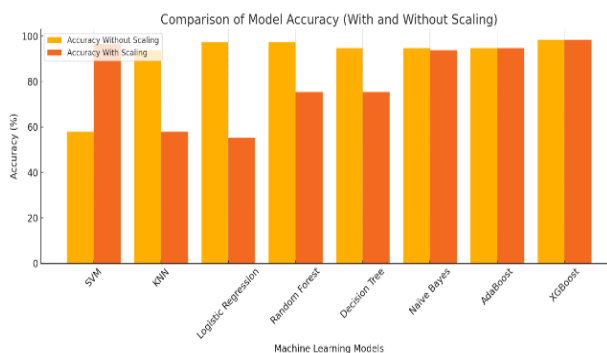
[Fig.2: Process Flow Diagram]

V. RESULTS & DISCUSSION

Experiments were conducted using an Intel Core i5-1.6 GHz processor with 8 GB RAM, in a Jupyter notebook environment. Models were implemented using Scikit-learn and Keras libraries. The prediction was made using machine learning algorithms, including SVM, K-NN, NB, LR, RF, DT, and XGB Classifier, as well as an ANN Deep Learning method. The train-test splitting technique was used to divide the data into 80-20 models for training and testing. The ANN model was implemented using the Keras framework. Sci-Kit Learn was used to implement the machine learning models. A training dataset (80%) was used to maximise model performance and document the results of the model assessment, while a test dataset (20%) was utilised to evaluate the model. First, the correlation between the features was examined, and any parameters with a high correlation (above 95%) with other features were eliminated. The chosen features were then subjected to correlation using both the standard ML models and the ANN models. Many ANN parameters were adjusted for each test, using a batch size of 32 and a specified number of epochs. In-depth documentation will be kept of each experiment's testing procedure and CV outcomes.

Table-II: Comparison Between Techniques

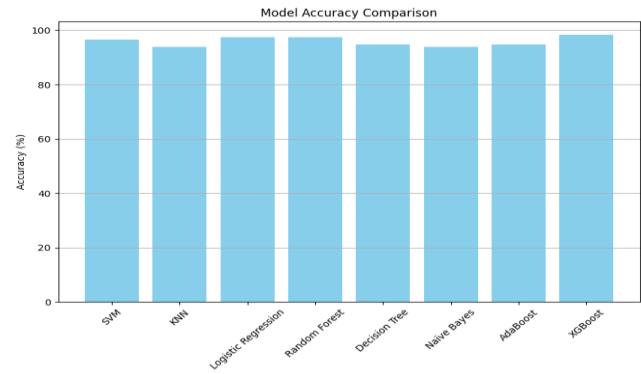
Techniques	Accuracy Without Standard Scale	Accuracy With Standard Scale
SVM	57.89%	96.49%
KNN	93.85%	57.89%
Logistic Regression	97.36%	55.26%
Random Forest	97.36%	75.43%
Decision Tree	94.73%	75.43%
Naïve Bayes	94.73%	93.85%
AdaBoost	94.73%	94.73%
XGBoost	98.24%	98.24%



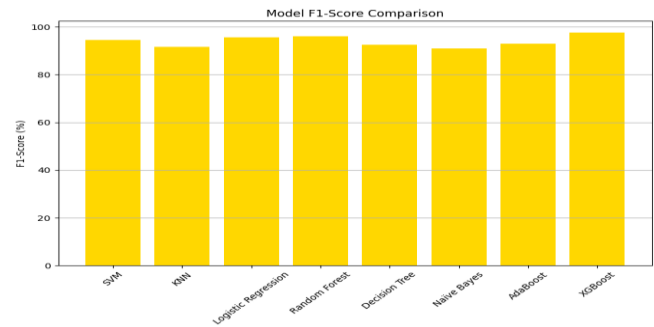
[Fig.3: Comparison of Model Accuracy (with and Without Scaling)]

Table-III: Comparison Between Evaluation Parameters

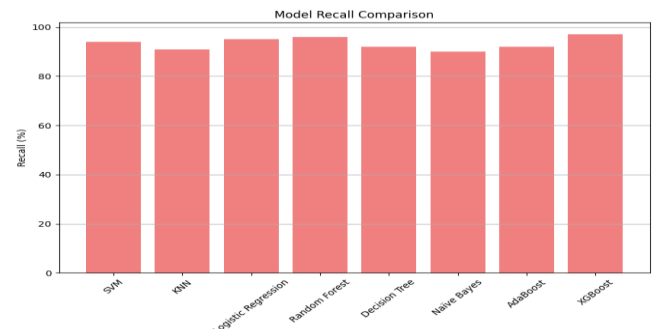
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	96.49	95	94	94.5
KNN	93.85	92	91	91.5
Logistic Regression	97.36	96	95	95.5
Random Forest	97.36	96	96	96
Decision Tree	94.73	93	92	92.5
Naïve Bayes	93.85	92	90	91
AdaBoost	94.73	94	92	93
XGBoost	98.24	98	97	97.5



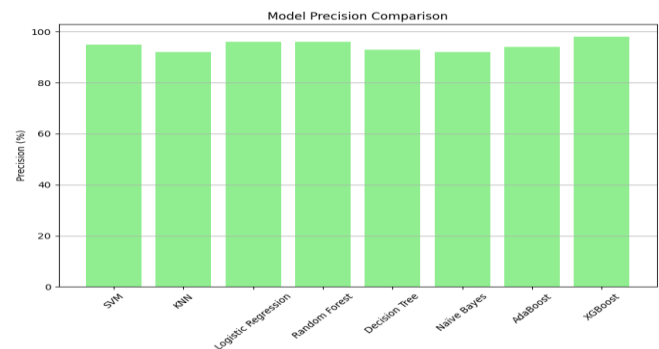
[Fig.4: Comparison of Model Accuracy]



[Fig.5: Comparison of F1-Score]



[Fig.6: Model Recall Comparison]



[Fig.7: Model Precision Comparison]

Key Findings: SVM: 96.49%, KNN: 93.85%, Logistic Regression: 97.36%, Random Forest: 97.36%, Decision Tree: 94.73%, Naïve Bayes: 93.85%, AdaBoost: 94.73%, XGBoost: 98.24%. XGBoost achieved the highest accuracy, demonstrating robustness regardless of whether feature scaling was used or not. Artificial Neural Networks (ANNs) were also trained with adjustments to batch size and epoch to optimise performance.

VI. CONCLUSION

This study explored the application of traditional machine learning and deep learning methods for breast cancer detection using the Wisconsin Breast Cancer Dataset. Models such as Random Forest, Logistic Regression, and especially XG-Boost showed high accuracy, with deep learning models like CNN further enhancing predictive performance. Machine learning techniques, when applied effectively, can significantly enhance early detection, leading to improved patient outcomes. However, challenges such as data bias, model interpretability, and ethical considerations must be addressed to ensure the responsible deployment of these models. Future work should explore personalised risk assessments and clinical decision-support systems that integrate real-world patient data and advanced imaging techniques for even greater accuracy and reliability.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Diagnostics, "Deep Learning Techniques for Breast Cancer Detection," *MDPI Diagnostics*, vol. 13, no. 19, 2023. [Online]. Available: <https://www.mdpi.com/2075-4418/13/19/3113>
2. Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers* (Vol. 12, Issue 7, Article 151). DOI: <https://doi.org/10.3390/computers12070151>
3. Zhou, X., Li, Y., Gururajan, R., Bargshady, G., Tao, X., Venkataraman, R., Barua, P. D., & Kondalsamy-Chennakesavan, S. (2020). *A New Deep Convolutional Neural Network Model for Automated Breast Cancer Detection*. In Proceedings of the 7th IEEE International Conference on Behavioural and Social Computing (BESC 2020) (pp.1–4). IEEE, Institute of Electrical and Electronics Engineers. DOI: <https://doi.org/10.1109/BESC51023.2020.9348322>
4. Hamidinekoo, A., Denton, K., Rampun, A., Honnor, S., & Zwiggelaar, A. (2018). Deep Learning in Mammography and Breast Histology: An Overview and Future Trends. In *Medical Image Analysis*, (Vol. 47, pp. 45–67). DOI: <https://doi.org/10.1016/j.media.2018.04.005>
5. Hossain, M., & Muhammad, G. (2019). Breast Cancer Detection System Based on Feature Selection and Kernel Support Vector Machine. In *Cognitive Computation* (Vol. 11, Issue 6, pp. 1361–1370). DOI: <https://doi.org/10.1007/s12559-019-09651-7>
6. Hussain, Z., Khan, S. S., Reid, J. G., Shah, M. R., & Zafar, N. A. (2020). Breast Cancer Detection Using Deep Learning Techniques in Ultrasound Images: A Comprehensive Review. In *Computers in Biology and Medicine* (Vol. 103, pp. 31–46). DOI: <https://doi.org/10.1016/j.compbiomed.2018.10.032>
7. Rahimzadeh, M., & Attar, A. (2019). Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines. In *ICT Express* (Vol. 6, Issue 2, pp. 123–129). DOI: <https://doi.org/10.1016/j.ict.2019.03.005>

8. Qiu, Y., Zhang, J., Mei, H., & Wu, S. (2020). A Review of Deep Learning Techniques for Breast Cancer Detection in Medical Imaging Modalities. In *IEEE Access* (Vol. 8, pp. 128337–128352). DOI: <https://doi.org/10.1109/ACCESS.2020.3008530>
9. Omer, S., Alkassab, A., & Saleh, M. Y. (2019). Machine Learning Techniques for Breast Cancer Computer-Aided Diagnosis Using Different Image Modalities: A Systematic Review. In *Applied Sciences* (Vol. 9, Issue 15, Article 2953). DOI: <https://doi.org/10.3390/app9152953>
10. Jain, A. K., Tripathi, R. C., & Agrawal, N. (2017). Recent Advances in Machine Learning Techniques for Breast Cancer Detection from Medical Imaging Data. In *International Journal of Computer Applications* (Vol. 160, Issue 6, pp. 31–36). DOI: <https://doi.org/10.5120/ijca2017913122>

AUTHOR'S PROFILE



Tanusree Saha, M. Tech, pursuing Ph.D. from Magadh University, Bodh Gaya, currently working as Assistant Professor, JIS, College of Engineering, Kalyani. Pursued M. Tech Degree from West Bengal University of Technology. Having 13 years of teaching experience and 3 years of research experience in the field of Digital Image Processing and allied technology. I have guided several undergraduate-level projects to date throughout my teaching journey, which began in 2021 as a certified AWS Cloud Foundation Educator. In addition to my work in the domain of Digital Image Processing research, I am also involved in Machine Learning and Artificial Intelligence-based projects. To date, we have published more than 20 IPR patents and received the AWS Educator Excellence Award in consecutive three years [2022, 2023, 2024] by AWS Academy—life members of FOSET, ISTE and CSI.



Sarmistha Santra, a Student of Master's in Computer Applications from JIS, College of Engineering, Kalyani, in the year 2024. She has completed her graduation from the West Bengal University of Technology with a Bachelor's degree in Computer Applications. Currently working as a software developer in a reputed MNC.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/ or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.