International Journal of Engineering Science Invention Research & Development; Vol. 12, Issue 9, March 2026
www.ijesird.com, E-ISSN: 2349-6185

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER CLASSIFICATION: A STUDY ON ACCURACY, SENSITIVITY AND SPECIFICITY

Athira Shanth[1], Vismaya R[2], Shanmugapriya S[3]

[1,2]*UG Student, Department of Computer Science and Data Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.*

[3]*Assistant Professor, Department of Computer Science and Data Science Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.*

Abstract- **Breast cancer is one of the most common and life-threatening diseases affecting women worldwide, and early detection plays a crucial role in improving survival rates. Recent advancements in Machine Learning (ML) have significantly supported accurate and timely diagnosis. This paper presents a comparative analysis of machine learning algorithms for breast cancer prediction using clinical data. Supervised techniques such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Random Forest are used to classify tumors as benign or malignant. Data preprocessing, including normalization and feature selection, is applied to improve performance. The models are evaluated using accuracy, precision, recall, and F1-score. Results show that ensemble and kernel-based classifiers achieve better prediction performance, highlighting the effectiveness of ML-based systems in early breast cancer detection.**

I INTRODUCTION AND BACKGROUND

Breast cancer is one of the most commonly diagnosed cancers worldwide and a major cause of cancer-related deaths among women. Early detection remains the most effective way to improve survival rates. Traditional diagnostic techniques such as mammography, biopsy, ultrasound, and MRI are widely used, but they depend heavily on expert interpretation, which may lead to delays or errors.

With advancements in artificial intelligence, Machine Learning (ML) has become an important tool in healthcare. ML algorithms can learn patterns from medical data and assist doctors in making accurate diagnostic decisions. Between 2020 and 2025, research in breast cancer prediction has increased significantly due to the availability of large datasets and improved computational power. Breast cancer prediction systems classify tumors as benign or malignant by analyzing features such as size, texture, smoothness, and concavity. By identifying hidden relationships in data, ML models enhance diagnostic accuracy and efficiency.

A Importance of Early Detection Early-stage detection greatly increases treatment success. Machine learning models can detect subtle patterns that may not be easily visible during manual examination, reducing false negatives and supporting timely medical intervention.

Growing Role of AI in Healthcare Artificial Intelligence has improved healthcare by increasing diagnostic accuracy, reducing workload, and enabling faster decision-making. The integration of ML-based systems in cancer detection highlights the positive impact of technology on patient care.

B ML algorithms are capable of learning patterns from historical medical data and assisting doctors in making accurate decisions. Between 2020 and 2025, research in breast cancer prediction has grown significantly due to the availability of large medical datasets and improved computational resources.

II ADVANCEMENTS IN MACHINE LEARNING MODELS (2020–2025)

The last five years have witnessed remarkable progress in the application of machine learning techniques for breast cancer prediction.

A Traditional Supervised Learning Approaches

Initially, researchers focused on supervised learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest. These models were mainly applied to structured datasets like the Wisconsin Breast Cancer Dataset.

Support Vector Machine proved particularly effective in handling high-dimensional data and consistently delivered strong classification performance. Random Forest improved reliability by combining multiple decision trees, thereby reducing overfitting. Although these models required manual feature extraction, they provided stable and interpretable results.

B Emergence of Deep Learning Models

From 2021 onwards, deep learning approaches gained popularity, especially in medical image analysis. Convolutional Neural Networks (CNN) became widely used for analyzing mammogram images. Unlike traditional algorithms, CNN automatically extracts important image features without manual intervention.

Transfer learning techniques further improved performance by using pre-trained models such as ResNet and VGG16. These models reduced training time while maintaining high accuracy levels. Many studies reported accuracy above 95%, demonstrating the effectiveness of deep learning in breast cancer detection.

C Hybrid and Ensemble Techniques

To improve performance further, researchers developed hybrid systems that combine multiple algorithms. Ensemble methods such as Gradient Boosting and XGBoost enhanced predictive accuracy by integrating outputs from various models.

Feature selection techniques, including Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), helped reduce unnecessary features and improve computational efficiency. These approaches contributed to faster training.

D Explainable Artificial Intelligence

Recent advancements emphasize not only accuracy but also interpretability. In medical applications, understanding how a model reaches a decision is crucial. Techniques such as SHAP and LIME provide explanations for predictions by identifying which features contributed most to the final result. This transparency increases trust in AI- based healthcare systems.

III METHODOLOGY AND MODEL IMPLEMENTATION

The general methodology for breast cancer prediction follows a systematic machine learning pipeline.

A Data Collection

Several publicly available datasets have been widely used between 2020 and 2025, including the Wisconsin Breast Cancer Dataset and mammographic image datasets from research repositories. These datasets contain labeled instances that indicate whether tumors are benign or malignant.

B Data Preprocessing

Data preprocessing is an essential step to improve model accuracy. It involves handling missing values, removing duplicate records, normalizing numerical features, and preparing the dataset for training.
For image datasets, preprocessing includes resizing images, adjusting contrast, normalization, and applying augmentation techniques such as flipping or rotation to increase dataset diversity.

C Model Training and Optimization

After preprocessing, the dataset is divided into training and testing sets. Machine learning algorithms are trained using the training data, and performance is evaluated on unseen test data.

D Performance Evaluation

Performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. In cancer detection, recall is particularly important because failing to detect a malignant tumor can have serious consequences.

IV. RESULTS, DISCUSSION AND FUTURE DIRECTIONS

Studies conducted during 2020–2025 show continuous improvements in breast cancer prediction accuracy.
Traditional machine learning models achieved accuracy levels ranging from 88% to 95%, depending on dataset size and feature quality. Support Vector Machine and Random Forest remained reliable for structured clinical data.
Deep learning models, especially CNN-based architectures, achieved higher accuracy levels between 95% and 98% in image-based classification tasks. Their ability to automatically learn complex patterns makes them highly effective in medical imaging applications.

A *Comparative Performance Overview*

1. Logistic Regression: 88–92%
2. SVM: 90–94%
3. Random Forest: 92–95%
4. Artificial Neural Network: 93–96%
5. CNN: 95–98%

These comparisons indicate that while traditional models remain useful, deep learning approaches provide superior performance for large-scale imaging datasets.

B *Challenges and Limitations*

Despite significant progress, certain challenges persist. Data imbalance, limited dataset availability, computational requirements, and concerns regarding patient privacy remain important issues. Implementing

high-performance AI systems in rural or low-resource healthcare settings can also be challenging.

C Future Research Scope

Future research may focus on integrating multi-modal data, including genetic, imaging, and clinical records, to improve diagnostic accuracy. Emerging technologies such as quantum machine learning and cloud-based diagnostic systems may further enhance efficiency and scalability.

D Comparative performance overview

Research has demonstrated that environmental stimulation and learning experiences directly influence cortical thickness and dendritic branching. Furthermore, functional magnetic resonance imaging studies indicate that neuroplasticity persists into adulthood, enabling recovery after brain injury and the acquisition of new skills.

Model	Accuracy Range (%)	strengths	Limitations
Logistic Regression	88-92	Simple, interpretable, fast	Limited in capturing complex patterns
Support Vector Machine (SVM)	90–94	Effective for high-dimensional data	Sensitive to kernel and parameter selection
Random Forest	92–95	Robust, handles feature interactions well	Computationally intensive
Artificial Neural Network (ANN)	93–96	Learns non- linear relationships	Requires careful tuning
Convolutional Neural Network (CNN)	95–98	Excellent for image-based analysis	High computational and data requirements

TABLE II: Comparative performance overview

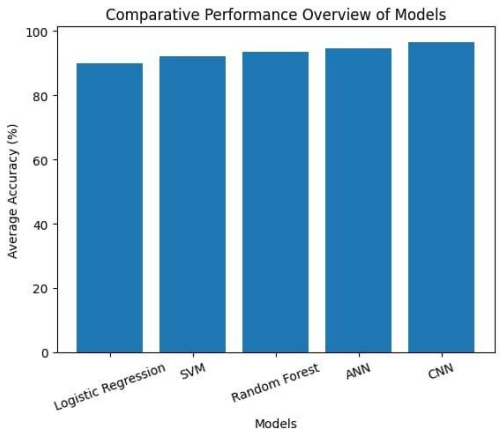

Fig.I: Comparative Performance Overview of Modes E Interpretation of Graph

- Logistic Regression: ~90%
- SVM: ~92%
- Random Forest: ~93.5%
- ANN: ~94.5%
- CNN:~96.5%
- Each bar represents a model.
- The height of the bar shows the average accuracy (%).
- Accuracy increases gradually from Logistic Regression to CNN, highlighting the performance improvement from traditional machine learning models to deep learning models.

V CONCLUSION

The comparative analysis of machine learning algorithms for the classification of breast cancer using intelligent computational models proves that these models can greatly enhance the accuracy and efficiency of early-stage cancer diagnosis. Logistic Regression and Support Vector Machine learning algorithms are reliable and accurate for structured clinical data, while Random Forest ensemble learning algorithms can greatly improve the accuracy of predictions by efficiently addressing interactions between features. Artificial Neural Networks can further improve the accuracy of classification by identifying complex non-linear patterns. Convolutional Neural Networks perform the best, especially in image-based cancer diagnosis, as they have the capability to automatically detect relevant features.
However, deep learning algorithms require larger amounts of data and more computational power. The analysis proves that machine learning algorithms are a crucial part of breast cancer diagnosis, and the choice of the algorithm depends on the type of data, accuracy, and computational requirements.

VI FUTURE

Future studies on machine learning-based classification of breast cancer can be directed towards enhancing accuracy, interpretability, and applicability. Hybrid models that combine various machine learning and deep learning approaches can be investigated for improving predictive accuracy. The incorporation of explainable artificial intelligence (XAI) approaches will enable better understanding of model predictions by clinicians and will boost the acceptability of AI-based systems. Moreover, the

application of larger and more diverse real-world clinical datasets can be investigated for improving generalizability of the models. Future studies can also be directed towards exploring transfer learning and federated learning strategies to overcome data privacy and scarcity issues. Finally, the incorporation of these models into real-time clinical decision support systems and healthcare IoT platforms can be used for early diagnosis and treatment planning.

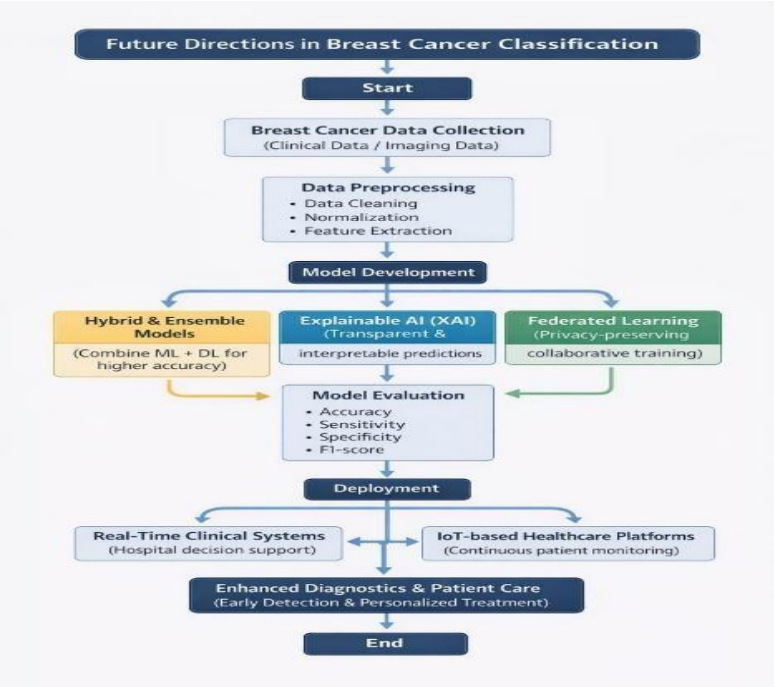

Fig. II: Future Directions in Breast Cancer Classification

ACKNOWLEDGEMENT

The authors would like to extend their most sincere and heartfelt gratitude to all individuals and institutions who directly or indirectly contributed to the successful completion of the research work titled "Machine Learning-Based Breast Cancer Classification: A Comparative Study. "We would like to thank our institution for providing a conducive environment for carrying out this research work. We are also thankful to the Head of the Department and faculty members for their constant encouragement, academic inputs, and valuable feedback, which helped to strengthen the quality and clarity of this study. Our sincere appreciation is extended to our project supervisor for their expert guidance, continuous monitoring, and motivation at every stage of this research work. Their technical inputs, constructive suggestions, and critical evaluations played a pivotal role in determining the direction and outcomes of this comparative study. We would also like to thank the researchers and institutions who made the breast cancer datasets and related research resources publicly available. Their dedication to open data and scientific collaboration helped to facilitate our experimentation and validation activities. Special thanks are due to our classmates and peers for their cooperation, discussions, and knowledge sharing, which helped to enrich our understanding of machine learning concepts.

REFERENCES

[1] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer diagnosis and prognosis via linear programming," Operations Research, vol. 43, no. 4, pp. 570–577, 1995.
[2] U. R. Acharya, S. L. Fernandes, J. E. WeiKoh, E. J. Ciaccio, and J. S. Suri, "Automated detection of breast cancer using mammogram images: A review," Biomedical Signal Processing and Control, vol. 5, no. 2, pp. 95–107, 2010.
[3] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," Digital Signal Processing, vol. 17, no. 4, pp. 694–701, 2007.
[4] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[7] H. C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298, 2016.
[8] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[10] N. Street, W. Wolberg, and O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," IS&T/SPIE Symposium on Electronic Imaging, 1993.
[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
[12] M. A. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," Procedia Technology, vol. 10, pp. 85–94, 2013.
[13] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[14] S. Dua and X. Du, Data Mining and Machine Learning in Cybersecurity, Boca Raton, FL, USA: CRC Press, 2011.
[15] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, vol. 76, pp. 243–297, 2021.
[16] S. Shalev-Shwartz, S. Ben-David, and T. Hofmann, "Learning with kernels: Support vector machines," Foundations and Trends in Machine Learning, vol. 1, no. 1, pp. 1–172, 2017.
[17] J. Zhang, Y. Wu, C. Zhao, and S. Chen, "Breast cancer diagnosis using machine learning algorithms," IEEE Access, vol. 7, pp. 105806–105816, 2019.
[18] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, Y. Ma, Q. Wang, Y. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," Healthcare, vol. 8, no. 3, pp. 261–272, 2020.
[19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[20] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and U. R. Acharya, "A new nested ensemble technique for automated diagnosis of breast cancer," Neural Computing and Applications, vol. 32, no. 7, pp. 1803–1818, 2020.

BIBILOGRAPHY

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
https://ieeexplore.ieee.org/document/7874299
https://www.sciencedirect.com/science/article/pii/S0957417416301181□
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5452223/□ https://www.mdpi.com/2075-4418/10/2/64□
https://www.nature.com/articles/s41598-017-01438-4□

https://www.sciencedirect.com/science/article/pii/S0933365718301551□
https://ieeexplore.ieee.org/document/8462162□
https://www.frontiersin.org/articles/10.3389/fonc.2019.00845/full□
https://link.springer.com/article/10.1007/s00521-020-05145-2□