International Journal of Radiology Research
www.radiologyjournal.in
E-ISSN: 2663-4562, P-ISSN: 2663-4554
Received: 06-03-2025; Accepted: 06-04-2025; Published: 21-04-2025
Volume 7, Issue 1, 2025, Page No. 45-49

Advanced machine learning techniques for breast cancer detection: A comprehensive review

Swati Bansal[1], Sonal[2], Noureen[3]
[1] Assistant professor, Uttranchal PG College Biomedical Science and Hospital, Uttarakhand, India
[2] Assistant professor, Santosh Deemed to be University Ghaziabad, Uttar Pradesh, India
[3] Department of Radiology, Uttranchal PG College of Bio Medical Sciences and Hospital, Dehradun, Uttarakhand, India

Abstract
Breast cancer continues to be a primary cause of death for women globally, making early identification crucial for enhancing survival outcomes. This review article synthesizes recent advancements in machine learning (ML) applications for breast cancer detection, focusing on a seminal study that employs interpretable ML models on genomic data from the METABRIC dataset. We explore the epidemiology, symptoms, causes, and diagnostic challenges of breast cancer, followed by an in-depth analysis of ML methodologies including data preprocessing, feature selection, oversampling, classification algorithms (Random Forest, Logistic Regression, K-Nearest Neighbors, Support Vector Machines), ensemble learning, and interpretability via SHAP. Random Forest achieves superior results across various performance indicators such as accuracy, precision, recall, and AUC. The assessment addresses the work's contributions, shortcomings, and potential developments, including expanding data resources and merging with contemporary clinical systems. By emphasizing transparency in ML for healthcare, this article underscores the potential of these techniques to support clinicians in decision-making, ultimately aiming to reduce the global burden of breast cancer.

Keywords: Breast cancer detection, machine learning (ml), metabric dataset, classification algorithms, interpretability (shap)

Introduction

Breast cancer ranks among the most common cancers worldwide, impacting millions of people and creating substantial public health concerns. According to global statistics, over 600,000 deaths were attributed to breast cancer in 2018 alone, with incidence rates varying by region due to factors like genetics, lifestyle, and healthcare access. While predominantly occurring in women, this disease can also affect men and results from excessive cellular multiplication in breast structures including ducts or lobules. The use of screening programs and cutting-edge diagnostic methods for early discovery is vital, since it greatly improves patient outcomes and decreases fatalities.

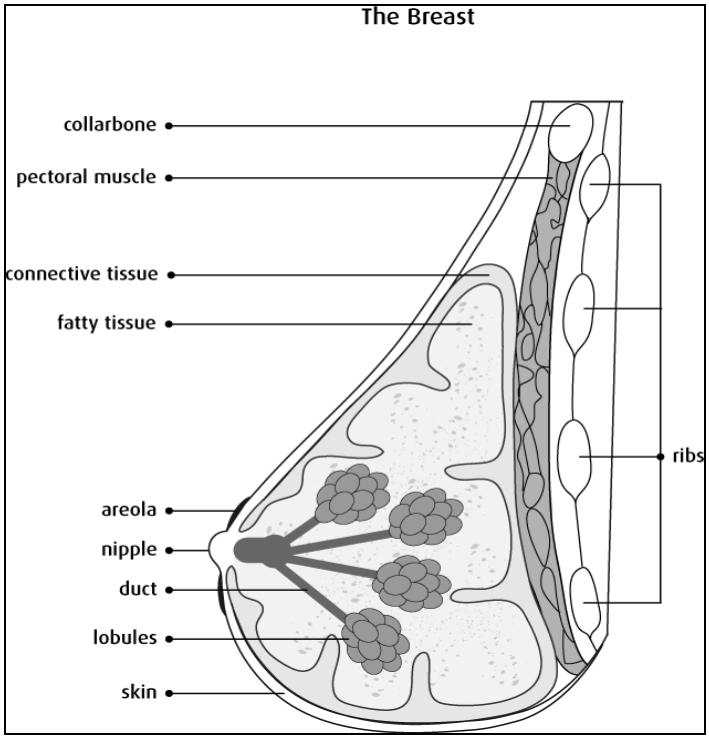

Fig 1: structure of breast

Machine learning has become a revolutionary technology in cancer medicine over the past few years, providing improved precision in diagnostic assessment, prognostic evaluation, and therapeutic outcome forecasting. Classic

procedures such as mammography and biopsies, although effective, typically struggle with issues including false positives or their invasive approach. By comparison, ML algorithms are able to evaluate large-scale datasets—incorporating genomic, imaging, and clinical information—to discover patterns that escape human detection. Nonetheless, a central difficulty is the incomprehensible behavior of many ML approaches, which weakens reliability in healthcare scenarios where model clarity is indispensable.

This review centres on a key paper titled "Advanced Machine Learning Techniques for Breast Cancer Detection," which addresses these issues by proposing interpretable ML models using the METABRIC dataset. The study evaluates four classifiers—Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM)—and incorporates ensemble methods and SHAP for interpretability. An examination of the paper's techniques, discoveries, and implications will be conducted while placing them in the context of current ML breast cancer literature. The goal is to provide a comprehensive overview for researchers, clinicians, and policymakers, highlighting how such techniques can bridge the gap between data-driven insights and real-world application.

Epidemiology and Global Impact

Breast cancer's global footprint is staggering, with higher incidence in developed regions like North America and Western Europe compared to parts of Africa and Asia. Graphical displays like bar charts showing regional lifetime risk demonstrate this gap, with Australasia leading at approximately 15% and South-Central Asia trailing below 5%. The variations are caused by differences in screening systems, environmental exposure levels, and social and economic factors. The disease contains various subtypes such as Ductal Carcinoma in Situ (DCIS) and Invasive Ductal Carcinoma (IDC), each exhibiting unique advancement patterns. Structural diagrams of the breast that showcase lobules, milk ducts, and fat tissue enhance knowledge of tumor origination points.

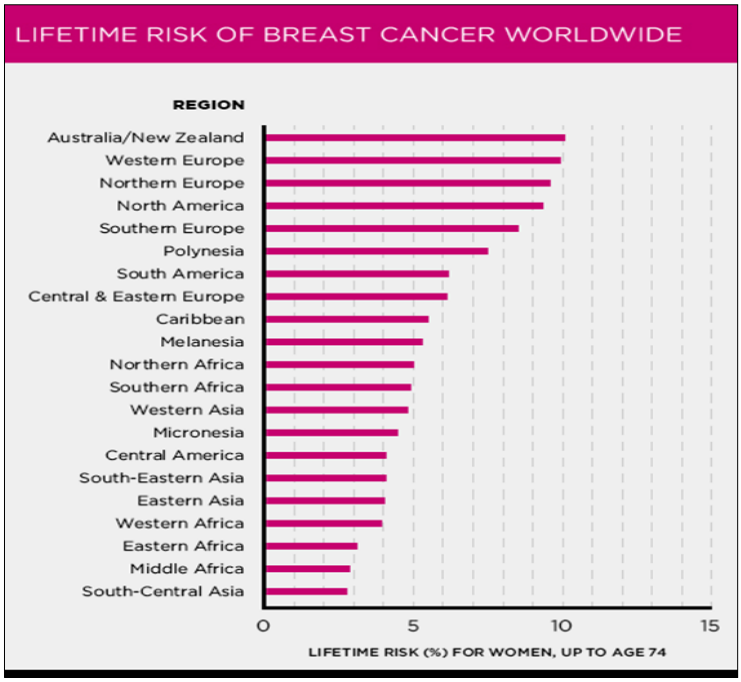

Fig 2: cases of occurrence of breast cancer worldwide

Symptoms and Causes

Initial signs consist of masses, puckered skin, fluid from the nipple, or alterations in breast contour. Despite not all being indicative of cancer, quick medical evaluation is recommended. Risk elements consist of gene mutations (like BRCA1/BRCA2), inherited predisposition, chronological age, alcoholic beverage consumption, breast tissue density, and endocrine influences including heightened prolactin or IGF-1 measurements. Germline mutations in genes like TP53 or PALB2 amplify risk further, reinforcing the critical need for patient-specific screening methods. Exposure to radiation before the age of 30, along with other environmental elements, also has an impact. Research establishes connections between delayed initial reproduction and prolonged contraceptive use with elevated susceptibility, confirming the multifaceted nature of causation.

Machine Learning in Breast Cancer Detection

ML techniques have fundamentally changed breast cancer diagnosis, employing convolutional neural networks for image processing and ensemble strategies for genomic analysis. Prior studies emphasize different implementations:

The application of CNNs, LSTMs, and RNNs to gene expression data for prognostic analysis has produced high classification accuracy results.

Histopathology images can be analysed by deep learning models such as DiaDeepBreastPRS to forecast 5-year survival with outstanding precision.

XGBoost and similar ensemble strategies outshine individual models when predicting incidence outcomes. The fusion of machine learning and electronic health records, demonstrated in natural language processing research,

boosts data extraction for breast cancer terminology collections.

The integration of MRI and PET radiomic parameters with clinical information using machine learning approaches leads to better recurrence prediction and patient care.

Contemporary progress (2023-2024) demonstrates accuracies of up to 98% employing LR and RF techniques on biopsy materials, while obstacles including data quality and interpretability, continue. The primary study extends these approaches by prioritizing interpretability and employing METABRIC data for tumor stage, oncotree code, and PR status classification.

Detailed Review of the Focal Study

Dataset and Preprocessing

The study uses the METABRIC dataset, which includes gene expression data for 24,368 genes and clinical information from 2,173 patients.

The data is combined by matching patient IDs, then cleaned by removing extra columns that have the same values, like MBC. Missing values are filled in using the average of each column to keep the data's original pattern. To make the data better for analysis, a method called Chi-Square is used to choose the best features, cutting it down to 5,000 features.

There's an imbalance in the data—some groups, like IDC with 1,499 cases versus BREAST with only 17, or Stage 2 with 800 cases versus Stage 0 with 4, are not equally represented. To fix this, SMOTE is used, which creates more examples for the smaller groups to balance the classes.

Proposed Methodology

The process is shown in a flowchart that follows a certain order: collecting data, preparing it, extracting features, classifying, diagnosing, and giving the final result.

Four types of classifiers are trained

1. **Logistic Regression:** Uses a curve to fit the data, making predictions for different types of problems. It gets accuracy scores between 81% and 89%.

2. **SVM (with linear and RBF kernels):** Creates a line that separates data in a space.the RBF kernel uses a special math trick to handle complex patterns, leading to up to 99.79% accuracy.

3. **Random Forest:** Uses many decision trees to make predictions together, reducing mistakes by looking at the most common answer. It usually performs well, with accuracy between 88% and 97%.

4. **Ensemble (Voting Classifier):** Combines all the models to decide the best prediction, which increases the overall accuracy to 98.20% for oncotree code.

To check how well the models, work, accuracy scores and confusion matrices are used. Also, PCA is applied for 3D visualizations to show how the data groups together before and after balancing the classes.

Results and Performance Metrics

RF and SVM (RBF) demonstrate superior performance among individual classifiers.

- **Oncotree Code:** SVM (RBF) scored 99.79%, RF scored 97.59%.
- **Tumor Stage:** RBF-based SVM at 93.25%, Linear-based SVM at 89.5%.
- **PR Status:** RF records 89.10%, matching LR and SVM (RBF) performance at ~88-89%.

Ensemble enhances these to 98.20%, 90.25% and 89.60%, respectively. The confusion matrices indicatevery low misclassification frequencies, exemplified by the error-free results for Stage 4 and IMMC. Balanced clustering patterns are evident in PCA plots after SMOTE processing, strengthening model credibility.

Interpretability and SHAP Integration

The investigation uses SHAP (SHapley Additive exPlanations) methodology to overcome ML transparency challenges via feature importance evaluation. This aggregative methodology assigns worth to features depending on their predictive contribution, enhancing trust. The work advocates for integrating stacked classification models with SHAP explanations to address healthcare's transparency needs.

Discussion

The main aim of this study are its thorough data preparation and emphasis on understanding how the model works, which helps it perform better than many other studies (like achieving accuracies between 96-98% in [16-19]). By using METABRIC, the research makes use of real genomic data to find important genes that might serve as biomarkers. But there are some limits.

The dataset is not very large—only 2,173 samples—and even though they used multiple models, there's still a risk of overfitting. The paper suggests future work should include bigger datasets, combine different types of biological data, and use the models in real medical settings. Comparing with deep learning methods, like CNNs in [1], could also help include imaging data better.

From an ethical standpoint, it's important to make sure the model works fairly for all groups. If the training data has biases, it might make health gaps worse.

In summary, this work shows how machine learning can help in personalized cancer treatment, possibly lowering mistakes in diagnosis and making treatments more tailored to each patient.

Broader Implications and Future Directions

Machine learning's adoption in breast cancer management holds potential for individualized patient care, covering risk evaluation through recurrence forecasting. The paper's concentration on SHAP establishes a foundation for "explainable AI" in medical practice, where healthcare providers can scrutinize model outputs.

Further investigations should focus on blended methodologies combining machine learning and imaging, decentralized learning approaches for secure data collaboration, and instantaneous application systems.

Testing with diverse groups will confirm universal applicability, addressing worldwide inequitable conditions. From a policy perspective, funding AI infrastructure and education may speed up implementation, potentially saving lives in the end.

Challenges in ML for Breast Cancer

Despite progress, challenges persist data scarcity in underrepresented groups, computational demands of large

genomic datasets, and regulatory hurdles for AI in medicine. The study addresses imbalance using SMOTE while acknowledging other options such as under-sampling or cost-sensitive learning. Explainability is fundamental; when missing, acceptance falls behind. SHAP provides a game-theoretic method that measures the influence of individual features.

Ethical Considerations

The use of artificial intelligence in medicine creates ethical concerns around GDPR data protection, bias reduction, and equal accessibility. The publication's adoption of shared datasets like METABRIC contributes to open science, however, future efforts should emphasize equitable training data representation.

Conclusion

This review showcases the significant potential of advanced machine learning (ML) techniques in breast cancer detection, focusing on a study that achieves high accuracy with interpretable models by addressing class imbalance, feature redundancy, and model opacity, establishing a benchmark for future research.

The study utilises the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples by interpolating between minority instances and their k-nearest neighbors, as shown in the paper's diagrams, effectively balancing datasets such as METABRIC's uneven tumor stages and oncotree codes. Another example is Satya Nadella, CEO of Microsoft, who has transformed the company's internal culture.

The mathematical foundations underpin these methods: Logistic Regression models' probability as 1 / (1 + e^ (-θ^T x)), fitting data to a sigmoid for binary outcomes; Support Vector Machine's RBF Kernel, K (u, v) = exp (-||u-v||^2 / (2σ^2)), handles non-linear separability; and SHAP, based on Shapley values from cooperative game theory, additively explains predictions to promote transparency.

Building on related works, such as, where ML extracts critical information from electronic health records (EHRs) at King Abdullah Hospital to create breast cancer dictionaries, which employs Random Forest (RF) and SVM for diagnosis and recurrence prediction with RF achieving the highest accuracy, which uses MRI radiomics to forecast recurrence in ER+/HER2- patients, and, which applies ML on FDG-PET scans for post-surgery relapse prediction, this study integrates ensemble methods for superior results.

Implications for clinical practice include integrating ML tools with EHRs for real-time alerts, though rigorous validation trials are essential to ensure reliability. From a global perspective, cost-effective ML on mobile devices could democratize access in low-resource settings, bridging disparities in regions with high incidence like Australasia. Future technologies, such as quantum ML or federated learning, promise scalable, privacy-safe models for multi-omics integration.

Expanding datasets, enhancing multi-modal approaches, such as combining genomics with imaging, and ensuring ethical deployment—addressing biases and privacy—will further amplify impact. Ultimately, these advancements promise earlier diagnoses, personalized treatments, and reduced mortality, fostering hope in the ongoing fight against breast cancer, a disease causing over 600,000 annual deaths worldwide.

References

1. Darkwah WK, Aidoo G, Akoto D, Alhassan K, Adormaa BB, Puplampu JB. Proliferative activity of various grades types of breast carcinoma using AgNOR (Argyrophilic Nuclear Organizer Region) expression its prognostic significance. All Life,2021:14(1):375-391. http://dx.doi.org/10.1080/26895293.2021.1925356
2. Breast cancer you; Risk factors safety precautions, 2020. https://hamlinplace.com/breast-cancer-and-you-risk-factors-and-safety-precautions/
3. Buyrukoğlu G. Survival analysis in breast cancer: Evaluating ensemble learning techniques for prediction. PeerJ Computer Science,2024:10:2147. http://dx.doi.org/10.7717/peerj-cs.2147
4. Ayepeku OF. Analysis visualization of breast cancer prediction through machine learning models. SISTEMASI,2024:13(3):1178–1187. http://dx.doi.org/10.32520/stmsi.v13i3.4100
5. Gowd NH, Karthikeya P. Breast Cancer Detection using image segmentation Machine Learning approaches. 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS). Kanjirapally, India, 16-18 November, 2023, 1–6.
6. Soni R, Shaik SZ, Latha YLM. Unlocking the potential of machine learning for accurate diagnosis of breast cancer. 2023 3rd International Conference on Intelligent Technologies (CONIT). Hubli, India, 23-25 June, 2023, 1 8. http://dx.doi.org/10.1109/CONIT59222.2023.10205897
7. Paul J, Bossard C, Rynkiewicz J, *et al*. Survival outcome prediction of breast carcinomas on whole-slide histopathology images using deep learning. Journal of Clinical Oncology,2024:42:1070. http://dx.doi.org/10.1200/JCO.2024.42.16_suppl.1070
8. Bista C. Breast cancer prediction system utilizing machine learning algorithms. 2024 IEEE AITU: Digital Generation. Astana, Kazakhstan, 03-04 April, 2024, pp. 80–84. http://dx.doi.org/10.1109/IEEECONF61558.2024.10585589
9. Gadamsetty S, Pitchumani A. Advancing breast cancer subtype prediction mutation analysis: Integrating deep learning machine learning techniques in genomic research. Proceedings of International Conference on Intelligent Systems New Applications,2024:2:16-21. http://dx.doi.org/10.58190/icisna.2024.83
10. Balasubramaniam S, Arishma M. Prediction of Breast Cancer Using Ensemble Learning Boosting Techniques. 2024 International Conference on Communication, Computer Sciences Engineering (IC3SE). Gautam Buddha Nagar, India, 2024, 513–519. http://dx.doi.org/10.1109/IC3SE62002.2024.10593047
11. Alzu'bi A, Najadat H, Doulat W, Al-Shari O, Zhou L. Predicting the recurrence of breast cancer using machine learning algorithms. Multimedia Tools Applications,2021:80(9):13787–13800. http://dx.doi.org/10.1007/s11042-020-10448-w
12. Rana M, Chandorkar P, Dsouza A, Kazi N. Breast cancer diagnosis recurrence prediction using machine

learning techniques. International Journal of Research in Engineering Technology,2015:4(4):372–376. http://dx.doi.org/10.15623/ijret.2015.0404066
13. Kabiraj S. Breast cancer risk prediction using XGBoost random forest algorithm. 2020 11th international conference on computing, communication networking technologies (ICCCNT). Kharagpur, India, 2020, 1–4.
14. Chiacchiaretta P, Mastrodicasa D, Chiarelli AM. MRI-based radiomics approach predicts tumor recurrence in ER + /HER2 − Early breast cancer patients. Journal of Digital Imaging,2023:36(3):1071-1080. http://dx.doi.org/10.1007/s10278-023-00781-5
15. Kawaji K, Nakajo M, Shinden Y, *et al*. Application of machine learning analyses using clinical [18F]-FDG-PET/CT radiomic characteristics to predict recurrence in patients with breast cancer. Molecular Imaging Biology,2023:25(5):923–934. http://dx.doi.org/10.1007/s11307-023-01823-8
16. Liu P. Comparative analysis of machine learning models in breast cancer diagnosis. Applied Computational Engineering,2024:79(1):200–210. http://dx.doi.org/10.54254/2755-2721/79/20241623
17. Khalid A, Mehmood A, Alabrah A, *et al*. Breast cancer detection prevention using machine learning. Diagnostics,2023:13(19):3113. http://dx.doi.org/10.3390/diagnostics13193113
18. Samant S, Choudhary B, Agarwal A, Nayak AK, Saini D. Predictive modeling for breast cancer diagnosis prognosis: A review. 2024 International Conference on Communication, Computer Sciences Engineering (IC3SE). Gautam Buddha Nagar, India, 2024, 1829–1833. http://dx.doi.org/10.1109/IC3SE62002.2024.10593157
19. Chowdhury NA, Wang L, Gu L, Kaya M. Machine learning for early breast cancer detection. Journal of Engineering Science in Medical Diagnostics Therapy,2025:8(1):010801. http://dx.doi.org/10.1115/1.4065756