Recent Advances in Breast Cancer Prediction Using Machine Learning: A Comprehensive Review

The landscape of breast cancer prediction has been significantly transformed by the integration of machine learning and artificial intelligence technologies, with recent research demonstrating remarkable progress in both traditional machine learning approaches and deep learning methodologies. Current investigations reveal that sophisticated algorithms can achieve accuracy rates exceeding 96% in breast cancer detection, with some models reaching perfect classification performance[1][3]. The field has evolved from conventional demographic-based risk assessment models to personalized prediction systems that leverage diverse data sources including medical imaging, genomics, and clinical information[2]. Systematic reviews indicate that Convolutional Neural Networks (CNNs) have emerged as the most accurate and extensively utilized approach for breast cancer detection, while traditional machine learning models such as Random Forest, Gradient Boosting, and AdaBoost continue to demonstrate exceptional performance in specific applications[3][4]. These developments represent a paradigm shift toward more precise, accessible, and patient-centric approaches to breast cancer diagnosis and risk assessment, though significant technological, ethical, and integration challenges remain to be addressed for successful clinical implementation.

Machine Learning Approaches and Methodologies

Deep Learning Architectures

The application of deep learning techniques has revolutionized breast cancer prediction, with Convolutional Neural Networks establishing themselves as the predominant architecture in this domain. Research demonstrates that CNN-based approaches can be categorized into two primary methodologies: transfer learning-based models and de novo trained models[4]. Transfer learning approaches utilize previously trained neural networks such as AlexNet, ResNet, and VGG, while de novo models are generated and trained from scratch specifically for breast cancer detection tasks[4]. The CNN Improvements for Breast Cancer Classification (CNNI-BCC) model exemplifies this advancement, helping physicians identify breast cancer through trained deep learning neural network systems that categorize breast cancer subtypes[1].

Beyond traditional CNNs, researchers have explored various sophisticated architectures including Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), and Autoencoders (AEs)[4]. Generative Adversarial Networks (GANs) have also gained prominence, particularly in addressing data augmentation challenges and improving

mammography diagnoses[4]. These GAN-based approaches have shown promise in generating synthetic training data and creating super-resolution images, with some implementations demonstrating superior performance compared to conventional image-augmentation methods[4].

Traditional Machine Learning Models

Despite the prominence of deep learning, traditional machine learning algorithms continue to demonstrate exceptional performance in breast cancer prediction tasks. Recent comparative studies have evaluated six different classification models: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Classifier (SVC), and Linear Support Vector Classifier (Linear SVC)[1]. Notably, ensemble methods have shown particular promise, with Random Forest, Gradient Boosting, and AdaBoost achieving perfect 100% accuracy in some implementations[3]. This exceptional performance is complemented by other models such as Bagging (99.56%), KNN (95.82%), and Multilayer Perceptron (96.92%), indicating that multiple approaches can achieve near-optimal results[3].

The effectiveness of these traditional approaches stems from their ability to handle diverse feature types and their interpretability, which remains crucial for clinical applications. The Random Forest algorithm, in particular, has consistently demonstrated superior performance across multiple studies, achieving accuracy rates of 96.49% in large-scale implementations[1]. These results suggest that while deep learning approaches capture complex patterns in raw data, traditional machine learning models remain highly competitive when applied to well-engineered features.

Data Types and Feature Selection Strategies

Imaging Data Integration

Contemporary breast cancer prediction models leverage multiple imaging modalities to enhance diagnostic accuracy and robustness. Mammographic imaging remains the cornerstone of screening programs, with recent research incorporating both craniocaudal and mediolateral views to create comprehensive datasets[1]. Advanced implementations utilize datasets containing thousands of merged images, such as the 3,002 combined pictures from 1,501 individuals who underwent digital mammography between February 2007 and May 2015[1]. The integration of dynamic contrast-enhanced (DCE) MRI provides additional morphological and functional lesion information, offering excellent sensitivity for breast cancer detection[1].

Digital breast tomosynthesis represents another significant advancement in imaging-based prediction, with researchers applying GANs to detect anomalies and complete images without requiring training photos with abnormalities[4]. This approach has produced encouraging results by locating suspicious areas through sophisticated pattern recognition algorithms[4]. The combination of multiple imaging modalities enables more comprehensive feature extraction and reduces the likelihood of false negatives or positives in clinical settings.

Feature Engineering and Selection

Effective feature selection has emerged as a critical component in developing robust breast cancer prediction models. Recent research employs three distinct modules for optimal feature selection: removal of low-variance features, univariate feature selection, and recursive feature elimination[1]. These sophisticated preprocessing techniques ensure that only the most informative features contribute to model training, thereby improving both accuracy and computational efficiency.

Genomic data integration represents another frontier in feature engineering, with studies utilizing validated gene expression data to detect clinical outcomes in breast cancer[4]. Histopathological imaging provides additional feature sources, enabling the identification of mitosis processes for invasive breast cancer diagnosis and the classification of tumor-related stroma[4]. The Wisconsin Breast Cancer Diagnostic dataset, containing 569 observations and 32 features, serves as a standard benchmark for evaluating feature selection strategies and model performance[3].

Performance Evaluation and Accuracy Achievements

Comparative Model Performance

Recent evaluations reveal significant variations in performance across different machine learning approaches, with several models achieving exceptional accuracy rates. The most striking results come from ensemble methods, where Random Forest, Gradient Boosting, and AdaBoost have demonstrated perfect 100% accuracy in controlled studies[3]. These results represent a substantial advancement over earlier approaches and suggest that sophisticated ensemble techniques can effectively capture the complex patterns associated with breast cancer development.

Traditional machine learning models have also shown impressive performance metrics, with Random Forest achieving 96.49% accuracy in large-scale implementations[1]. Support Vector Machines and other classical approaches continue to demonstrate competitive results, particularly when combined with effective feature engineering techniques[1]. The consistency of these high-performance results across multiple studies indicates that current machine learning approaches have reached a level of maturity suitable for clinical consideration.

Evaluation Metrics and Validation

Contemporary breast cancer prediction research employs comprehensive evaluation frameworks that extend beyond simple accuracy measurements. Performance assessment typically includes classification accuracy, specificity, sensitivity, F1-score, and precision

metrics[3]. These multi-dimensional evaluations provide clinicians with detailed insights into model reliability and help identify potential biases or limitations in specific applications.

The validation methodologies employed in recent studies demonstrate increasing sophistication, with researchers implementing rigorous training, testing, and validation protocols[3]. Cross-validation techniques and holdout datasets ensure that reported performance metrics reflect genuine predictive capability rather than overfitting to training data. This methodological rigor is essential for building confidence in machine learning applications for clinical decision-making.

Systematic Review Findings and Research Trends

Comprehensive Literature Analysis

Systematic reviews provide valuable insights into the broader landscape of breast cancer prediction research, with recent analyses examining 98 high-quality articles following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[4]. These comprehensive evaluations reveal that CNN-based approaches dominate the literature, with accuracy metrics serving as the most popular performance evaluation method[4]. The systematic nature of these reviews ensures that findings represent genuine trends rather than isolated successes.

The evolution of research focus has shifted toward personalized risk models that leverage individual patient information from medical imaging and associated reports[2]. This transition from demographic-based screening policies to AI-driven personalized approaches represents a fundamental change in how breast cancer risk is assessed and managed[2]. The integration of diverse data sources, including imaging, genomics, and clinical information, enables more sophisticated risk stratification than previously possible.

Methodological Advancements

Recent systematic reviews highlight significant methodological improvements in breast cancer prediction research. The application of artificial intelligence techniques, particularly deep learning, has enabled the development of models capable of diagnosing breast cancer up to 12 months earlier than conventional clinical procedures[4]. This early detection capability represents a crucial advancement that could significantly improve patient outcomes through timely intervention.

The integration of radiomics approaches with traditional clinical features has emerged as a particularly promising research direction[2]. These hybrid methodologies combine quantitative imaging features with clinical and genomic data to create more comprehensive prediction models[2]. The systematic evaluation of these approaches indicates that multi-modal integration consistently outperforms single-modality approaches across various performance metrics.

Current Challenges and Future Research Directions

Technical and Implementation Challenges

Despite remarkable progress in accuracy and sophistication, significant challenges remain in translating machine learning advances into clinical practice. Computational requirements represent a primary concern, as many deep learning approaches require substantial processing power for imaging methods and preprocessing[1]. This computational burden can limit the accessibility of advanced prediction systems, particularly in resource-constrained healthcare environments.

Data quality and standardization present additional challenges, as machine learning models require large, high-quality datasets for effective training and validation. The heterogeneity of imaging protocols, patient populations, and clinical practices across institutions can introduce biases that affect model generalizability[4]. Addressing these technical challenges requires collaborative efforts to establish standardized data collection and sharing protocols.

Ethical and Regulatory Considerations

The integration of machine learning into breast cancer prediction raises important ethical and legal considerations that must be addressed before widespread clinical adoption. Privacy concerns related to genomic and imaging data require sophisticated security measures and clear consent protocols[2]. The potential for algorithmic bias, particularly regarding underrepresented populations, necessitates careful attention to training data diversity and model fairness.

Regulatory pathways for AI-based medical devices continue to evolve, with healthcare authorities developing frameworks for evaluating machine learning applications in clinical settings. The interpretability of complex deep learning models remains a concern for clinicians who need to understand and validate algorithmic recommendations[4]. Future research must balance model performance with explainability to ensure that advanced prediction systems can be effectively integrated into clinical workflows.

Conclusion

The current state of breast cancer prediction using machine learning represents a remarkable convergence of technological advancement and clinical need, with recent research demonstrating that sophisticated algorithms can achieve exceptional accuracy rates while processing diverse data types including medical imaging, genomics, and clinical information. The dominance of CNN-based approaches in systematic reviews, combined with the continued excellence of traditional ensemble methods like Random Forest and Gradient Boosting, indicates that multiple technological pathways can lead to clinically relevant performance levels.

However, the translation of these research achievements into routine clinical practice requires addressing significant challenges related to computational requirements, data standardization, ethical considerations, and regulatory approval processes.

Future research directions must focus on developing more efficient algorithms that maintain high accuracy while reducing computational demands, establishing robust data sharing and standardization protocols, and creating interpretable models that can gain clinician trust and regulatory approval. The ultimate goal of creating more precise, accessible, and patient-centric approaches to breast cancer diagnosis and risk assessment remains achievable, but will require sustained collaboration between machine learning researchers, clinical practitioners, and healthcare policy makers. As the field continues to mature, the integration of personalized prediction models into standard care protocols has the potential to significantly improve early detection rates and patient outcomes through timely, accurate risk assessment and intervention strategies.

Citations:

1. https://pmc.ncbi.nlm.nih.gov/articles/PMC10572157/
2. https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2024.1343627/full
3. https://thesai.org/Publications/ViewPaper?Volume=14&Issue=2&Code=IJACSA&SerialNo=72
4. https://pmc.ncbi.nlm.nih.gov/articles/PMC9818155/
5. https://www.mdpi.com/2077-0383/13/21/6486
6. https://pubs.rsna.org/doi/abs/10.1148/ryai.220159
7. https://www.jatit.org/volumes/Vol103No1/10Vol103No1.pdf
8. https://www.nature.com/articles/s41586-021-04278-5
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC10698602/
10. https://radiologybusiness.com/topics/artificial-intelligence/fda-authorizes-1st-ai-tool-predict-5-year-breast-cancer-risk-routine-mammograms
11. https://pubmed.ncbi.nlm.nih.gov/39907762/
12. https://pubmed.ncbi.nlm.nih.gov/39839787/
13. https://www.scirp.org/journal/paperinformation?paperid=127241
14. https://www.nature.com/articles/s41598-024-81734-y
15. https://www.rsna.org/news/2024/march/deep-learning-for-predicting-breast-cancer
16. https://www.nature.com/articles/s41698-024-00666-y
17. https://pmc.ncbi.nlm.nih.gov/articles/PMC10625863/
18. https://www.science.org/doi/10.1126/scitranslmed.abo4802
19. https://www.nature.com/articles/s41598-025-01920-4
20. https://pubmed.ncbi.nlm.nih.gov/38663911/
21. https://www.igi-global.com/article/survey-of-breast-cancer-detection-using-machine-learning-techniques-in-big-data/227680
22. https://www.sciencedirect.com/science/article/pii/S1877050924016582
23. https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1446270/full

24. https://www.sciencedirect.com/science/article/pii/S2666521224000607
25. https://www.sciencedirect.com/science/article/pii/S2352914823001636
26. https://www.mdpi.com/2073-431X/13/11/294
27. https://www.nature.com/articles/s41598-024-57740-5
28. https://www.kaggle.com/code/junkal/breast-cancer-prediction-using-machine-learning
29. https://www.sciencedirect.com/science/article/pii/S2666307424000366
30. https://www.cancerimagingarchive.net/analysis-result/tcga-breast-radiogenomics/
31. https://portal.gdc.cancer.gov/projects/TCGA-BRCA
32. https://pmc.ncbi.nlm.nih.gov/articles/PMC7909418/
33. https://www.cancerimagingarchive.net/collection/tcga-brca/
34. https://www.oncology-central.com/asco-2025-can-ai-assist-in-identifying-her2-low-and-her2-ultralow-breast-cancers/
35. https://www.breastcancer.org/news/ai-tool-mammograms-predict-risk
36. https://internationalpubls.com/index.php/cana/article/view/3964
37. https://pubmed.ncbi.nlm.nih.gov/34875674/
38. https://pmc.ncbi.nlm.nih.gov/articles/PMC9028992/
39. https://pmc.ncbi.nlm.nih.gov/articles/PMC10393322/
40. https://academic.oup.com/bib/article/26/1/bbae628/7916277
41. https://journals.physiology.org/doi/abs/10.1152/physiolgenomics.00033.2023
42. https://www.nature.com/articles/s41598-020-66907-9
43. https://pubmed.ncbi.nlm.nih.gov/37370847/
44. https://www.nature.com/articles/s41598-024-76331-y
45. https://pmc.ncbi.nlm.nih.gov/articles/PMC9628143/
46. https://www.statnews.com/2025/05/29/ai-breast-cancer-detection-radiologists-slow-to-trust-new-mammogram-screening-tools/
47. https://pubs.rsna.org/doi/10.1148/radiol.233067
48. https://ecancer.org/en/news/25975-ai-supported-breast-cancer-screening-new-results-suggest-even-higher-accuracy
49. https://www.icadmed.com/breast-cancer-blog/icad-leadership-in-breast-health-ai-showcased-at-ecr-2025/
50. https://www.bangkokhospital.com/en/bangkok-cancer/content/ai-mammography-increases-the-efficiency-of-breast-cancer-detection
51. https://www.openpublichealthjournal.com/VOLUME/18/ELOCATOR/e18749445372257/PDF/
52. https://thesai.org/Publications/ViewPaper?Volume=16&Issue=1&Code=ijacsa&SerialNo=129
53. https://pubmed.ncbi.nlm.nih.gov/38571502/
54. https://researchers.mq.edu.au/en/publications/breast-cancer-risk-prediction-using-machine-learning-a-systematic