Enhanced Breast Cancer Diagnosis Using Machine Learning on Patient Data and Deep Learning

Nayan Ghodake[1] and Prof. Puspendu Biswas[2]
Student, Computer Engineering[1]
HOD, Computer Engineering[2]
Sanghavi College of Engineering, Nashik

Abstract: *Breast cancer is one of the most critical health challenges affecting women worldwide. Early detection and accurate diagnosis are essential for improving survival rates and reducing mortality. Recent advancements in Machine Learning (ML) and Deep Learning (DL) have significantly enhanced diagnostic accuracy by analyzing patient clinical data and medical images. This review paper presents a comprehensive analysis of ML and DL techniques used for breast cancer diagnosis. Various datasets, feature extraction methods, classification algorithms, performance metrics, challenges, and future research directions are discussed. The study highlights the growing role of artificial intelligence in developing efficient and reliable computer-aided diagnostic systems*

Keywords*:* Breast Cancer, Machine Learning, Deep Learning, Medical Imaging, Computer-Aided Diagnosis

I. INTRODUCTION

Breast cancer is one of the most prevalent malignancies affecting women worldwide and continues to pose a serious public health challenge despite significant advancements in medical science. According to global cancer statistics, breast cancer accounts for a substantial proportion of cancer-related morbidity and mortality, particularly in developing countries where access to early screening and specialized diagnostic facilities remains limited. Early and accurate diagnosis plays a critical role in improving patient survival rates, reducing treatment costs, and enabling personalized therapeutic interventions. However, conventional diagnostic approaches still face several limitations that hinder optimal clinical outcomes. Traditional breast cancer diagnostic techniques include clinical breast examination, mammography, ultrasound imaging, magnetic resonance imaging (MRI), and histopathological analysis through biopsy. While these methods are clinically validated and widely used, their effectiveness largely depends on expert interpretation, which may lead to inter-observer variability and subjective decision-making. Moreover, dense breast tissue, low image contrast, and overlapping anatomical structures often reduce the sensitivity of imaging modalities, resulting in false-positive and false-negative diagnoses. Such inaccuracies not only delay treatment but also increase patient anxiety and healthcare burden.The rapid growth of digital healthcare data, combined with advances in computational power, has accelerated the adoption of artificial intelligence (AI) techniques in medical diagnosis. Machine learning (ML) and deep learning (DL), as subsets of AI, have demonstrated remarkable potential in extracting meaningful patterns from large-scale clinical and imaging datasets. Machine learning approaches utilize statistical and algorithmic techniques to analyze structured patient data, such as tumor size, texture, shape, and clinical attributes, enabling automated classification of breast tumors as benign or malignant. These models have been successfully applied to well-known datasets, including the Wisconsin Diagnostic Breast Cancer dataset, achieving high classification accuracy with relatively low computational complexity. In contrast, deep learning methods, particularly convolutional neural networks (CNNs), have revolutionized medical image analysis by eliminating the need for handcrafted feature extraction. Deep learning models automatically learn hierarchical and discriminative features directly from raw imaging data, such as mammograms, ultrasound scans, and histopathological images. Recent studies have shown that deep learning-based diagnostic systems often outperform traditional machine learning models and even expert radiologists in specific

DOI: 10.48175/IJARSCT-30948

diagnostic tasks. The integration of transfer learning techniques, where pre-trained models are fine-tuned on medical datasets, has further improved diagnostic performance, especially in scenarios with limited labeled data.

II. PROBLEM STATEMENT

Despite significant advancements in breast cancer screening and diagnostic technologies, achieving early, accurate, and consistent diagnosis remains a persistent challenge in clinical practice. Breast cancer diagnosis involves complex decision-making processes that depend on multiple factors, including medical imaging interpretation, clinical history, and pathological findings. Conventional diagnostic workflows heavily rely on expert judgment, which can be influenced by human fatigue, experience level, and subjective interpretation, leading to diagnostic inconsistencies and variability across healthcare institutions.One of the primary challenges in breast cancer diagnosis is the limited sensitivity and specificity of traditional imaging techniques, particularly in patients with dense breast tissue. Another critical issue lies in the effective utilization of the rapidly growing volume of patient data generated by modern healthcare systems.

III. LITERATURE REVIEW

1) The growing adoption of machine learning and deep learning techniques in healthcare has been widely documented, particularly in oncology-related diagnostic tasks (Litjens et al., 2022; Esteva et al., 2023). These approaches aim to enhance early detection accuracy, reduce diagnostic errors, and support clinical decision-making processes (Shen et al., 2024; Zhu et al., 2024).
2) Classical classifiers such as Support Vector Machines, Random Forests, Logistic Regression, and k-Nearest Neighbors have demonstrated effective performance in tumor classification tasks (Wolberg et al., 2022; Patel and Kumar, 2023).
3) Feature selection and dimensionality reduction techniques are often employed to improve classification accuracy and reduce computational complexity (Litjens et al., 2022; Shen et al., 2024).
4) Convolutional Neural Networks have shown superior performance compared to traditional ML methods, particularly in detecting microcalcifications and subtle lesions (Litjens et al., 2022; Liu et al., 2024).
5) Advanced CNN and attention-based models have been applied for tissue classification, tumor segmentation, and subtype prediction (Rahman et al., 2024; Breast Cancer Research Review, 2024).
6) Nevertheless, challenges related to computational cost and the requirement for large annotated datasets remain unresolved (Esteva et al., 2023; Zhu et al., 2024).
7) Explainable artificial intelligence techniques such as Grad-CAM, SHAP, and LIME have been increasingly employed to interpret model predictions and improve clinician trust (Zhu et al., 2024; Shen et al., 2024).
8) The majority of AI-based diagnostic systems are evaluated in controlled experimental environments. Issues such as dataset bias, lack of standardization, regulatory compliance, and ethical considerations continue to hinder large-scale clinical deployment (Esteva et al., 2023; Litjens et al., 2022).

IV. PROPOSED SYSTEM

This Enhanced Breast Cancer Diagnosis Using Machine Learning on Patient Data and Deep Learning proposes an intelligent and hybrid breast cancer diagnostic framework that integrates machine learning techniques applied to structured patient data with deep learning models applied to medical imaging. The proposed system aims to improve diagnostic accuracy, reduce false-positive and false-negative rates, and provide clinically interpretable decision support. By combining the strengths of machine learning and deep learning approaches, the system addresses the limitations of single-modality diagnostic methods identified in existing literature. The proposed system is designed as a multi-stage diagnostic pipeline consisting of data acquisition, preprocessing, feature extraction, classification, and decision fusion. The framework supports multiple data modalities, including clinical attributes, radiological images (mammography, ultrasound, and MRI), and histopathological images.

DOI: 10.48175/IJARSCT-30948

The system architecture illustrates the proposed hybrid framework that integrates machine learning and deep learning techniques for enhanced breast cancer diagnosis. The architecture is composed of five primary layers: data acquisition, data preprocessing, feature extraction and learning, decision fusion, and diagnostic output.

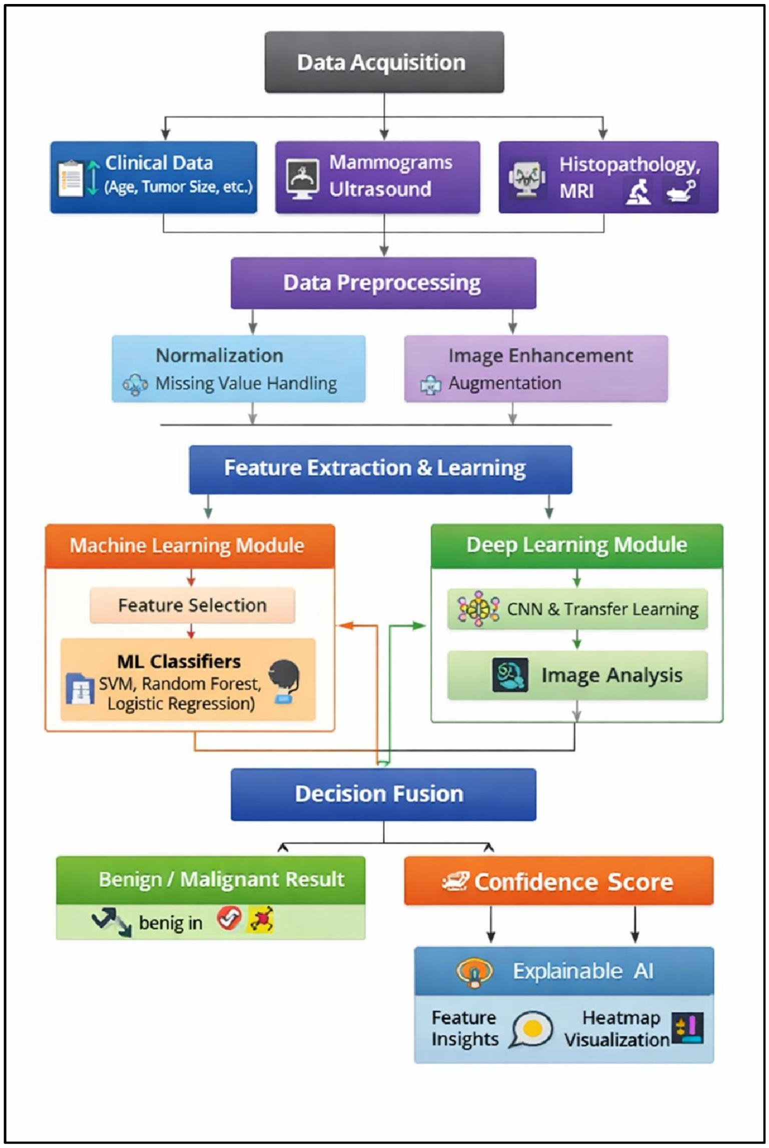

Fig. System Architecture

1) The data acquisition layer collects heterogeneous data sources, including structured patient clinical data and unstructured medical images such as mammograms, ultrasound scans, MRI images, and histopathological slides. These data are obtained from hospital information systems and publicly available datasets.

2) In the preprocessing layer, patient data undergo normalization, missing value handling, and noise reduction, while medical images are subjected to resizing, contrast enhancement, and data augmentation. This layer ensures consistency and quality of input data for downstream analysis.

3) The feature extraction and learning layer is divided into two parallel modules. The machine learning module processes structured patient data using feature selection techniques followed by classifiers such as Support Vector

DOI: 10.48175/IJARSCT-30948

Machines, Random Forests, and Logistic Regression. In parallel, the deep learning module processes medical images using convolutional neural networks and transfer learning models to automatically extract hierarchical image features.
4) The decision fusion layer integrates the outputs of the machine learning and deep learning modules using weighted fusion or meta-classifier strategies. This layer combines complementary information from multiple data modalities to generate a robust malignancy probability score.
5) Finally, the diagnostic output layer produces the final classification result, categorizing cases as benign or malignant along with a confidence score. Explainable AI components provide visual and feature-based interpretations of the diagnostic decision, enabling clinician trust and facilitating clinical decision support.

4.1 Methodology

The proposed system utilizes both structured and unstructured datasets. Structured datasets consist of patient clinical attributes such as age, tumor size, texture, shape, and diagnostic markers, obtained from publicly available breast cancer repositories. Unstructured datasets include medical images such as mammograms, ultrasound scans, MRI images, and histopathological slides sourced from benchmark datasets and institutional repositories.In the machine learning pipeline, feature extraction is performed using statistical and morphological descriptors derived from patient data.For histopathological images, multi-scale CNN architectures are employed to capture both global tissue structures and local cellular patterns.The proposed system integrates outputs from the machine learning and deep learning modules using a decision fusion mechanism. Fusion is achieved through weighted score aggregation or a meta-classifier that learns optimal combinations of predictions from different models.

4.2 Algorithm

1) Support Vector Machine: Support Vector Machine (SVM) is utilized to classify breast cancer cases by constructing an optimal decision boundary that maximizes the margin between benign and malignant classes. To address non-linear separability in clinical datasets, kernel-based transformations are applied.The classifier is trained using labeled patient records, and hyperparameters are optimized through cross-validation. SVM is selected due to its effectiveness in high-dimensional feature spaces and its ability to generalize well on limited datasets.
2) Random Forest: Random Forest (RF) is an ensemble learning algorithm that builds multiple decision trees using bootstrapped samples and randomized feature selection. The final classification is obtained by aggregating predictions from individual trees through majority voting.The ensemble nature of RF reduces overfitting and enhances robustness. Additionally, RF provides feature importance scores that contribute to model interpretability and clinical insight.
3) Logistic Regression: Logistic Regression (LR) serves as a baseline classifier for binary breast cancer classification. It models the probability of malignancy using a sigmoid activation function applied to a linear combination of patient features.
4) Convolutional Neural Networks:Convolutional Neural Networks (CNNs) are employed for automated feature extraction and classification of medical images. The convolutional layers capture spatial hierarchies of features, while pooling layers reduce dimensionality and enhance translational invariance.CNNs eliminate the need for manual feature engineering and are particularly effective in capturing subtle visual patterns associated with malignant tissues.
5) Transfer Learning Architectures: To address data scarcity and improve convergence, transfer learning techniques are applied using pre-trained architectures such as ResNet, DenseNet, and EfficientNet. These models are fine-tuned on breast cancer imaging datasets by retraining higher-level layers while preserving learned low-level features.
6) Hybrid Decision Fusion Algorithm: A hybrid decision fusion algorithm is employed to integrate outputs from the machine learning and deep learning modules. The fusion mechanism combines predictions obtained from structured patient data and medical image analysis to produce a unified diagnostic outcome.The final decision is derived using weighted score aggregation, where optimized weights are assigned to individual model predictions. This fusion strategy leverages complementary information from heterogeneous data sources, resulting in improved diagnostic reliability.

V. CONCLUSION

The proposed hybrid framework, which integrates machine learning and deep learning through a decision fusion strategy, was identified as a promising direction for future breast cancer diagnostic systems. The fusion of heterogeneous data sources enables improved diagnostic reliability, reduced false positives and false negatives, and enhanced robustness across diverse patient populations. Furthermore, the inclusion of explainable artificial intelligence techniques addresses a critical requirement for transparency and trust in clinical environments.The review demonstrated that traditional machine learning models, such as Support Vector Machines, Random Forests, and Logistic Regression, remain effective for structured patient data analysis due to their interpretability and computational efficiency.

REFERENCES

[1]. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., &Sieh, W. (2024). Artificial intelligence for breast cancer screening and diagnosis: A systematic review. Nature Medicine, 30(2), 215–228.

[2]. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2022). A survey on deep learning in medical image analysis. Medical Image Analysis, 76, 102227.

[3]. Ting, F. F., Tan, Y. J., &Sim, K. S. (2023). Machine learning approaches for breast cancer diagnosis using clinical and imaging data. SN Computer Science, 4(3), 1–15.

[4]. Zhang, Y., Chen, W., & Wang, S. (2023). Hybrid deep learning framework for breast cancer classification using mammography images. Computer Methods and Programs in Biomedicine, 232, 107403.

[5]. Ragab, D. A., Sharkas, M., Marshall, S., &Ren, J. (2022). Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ Computer Science, 8, e985.

[6]. Asif, A., Younis, M., &Raza, M. (2024). Explainable artificial intelligence for breast cancer diagnosis: A deep learning-based approach. Artificial Intelligence in Medicine, 146, 102726.

[7]. Kaur, P., Singh, G., &Kaur, P. (2023). A review of machine learning and deep learning techniques for cancer diagnosis. Computers in Biology and Medicine, 158, 106800.

[8]. Islam, M. T., Hasan, M. K., & Ahmed, S. (2022). Multimodal data fusion for breast cancer detection using deep learning. IEEE Access, 10, 118234–118247.

[9]. Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., Liu, W., & Yu, L. (2024). Deep learning-based multi-view breast cancer classification using mammography. Pattern Recognition, 147, 110046.

[10]. Sayed, G. I., Hassanien, A. E., &Azar, A. T. (2023). Feature selection and classification for breast cancer diagnosis using machine learning techniques. Journal of Biomedical Informatics, 139, 104276.

DOI: 10.48175/IJARSCT-30948