Breast Cancer Prediction Using Machine Learning Algorithms

Sarah Sejoro
Jiann-Ping Hsu College of Public Health
Georgia Southern University
Statesboro, GA, United States
ss81506@georgiasouthern.edu

Anandhi Kandaswamy
Jack N. Averitt College of Graduate Studies
Georgia Southern University
Statesboro, GA, United States
ak20427@georgiasouthern.edu

Bright Agbenu
Jiann-Ping Hsu College of Public Health
Georgia Southern University
Statesboro, GA, United States
ba09534@georgiasouthern.edu

***Index Terms*—Breast cancer, machine learning, disease classification, diagnostic prediction, feature engineering**

I. Introduction

Cancer is one of the many systemic diseases with millions of diagnoses being made each year. Breast cancer (BC), which results from cellular mutations in breast tissue cells, is the most commonly diagnosed cancer among women in the United States; in 2025, it is estimated that there will be 316,950 new cases of female breast cancer, representing 15.5% of all new cancer cases in the U.S. [2]. Female breast cancer also has a reported 5-year relative survival of 91.7% [2]. Current data indicate that the average lifetime risk of a woman developing breast cancer in the United States is about 13% [4]. Because breast cancer remains a leading cause of mortality in women worldwide, early and precise diagnostic interventions are essential. Although pathology is considered the gold standard for detection, traditional screening tools such as mammography, breast MRI, ultrasound, and PET scans can have variable diagnostic performance and may contribute to unnecessary biopsies and psychological distress [7], [8]. Advances in bioinformatics have introduced digital pathology, machine learning (ML), and deep learning (DL) as transformative approaches capable of improving diagnostic accuracy and classification performance [9], [10], [14]. This review examines studies that focus specifically on ML and DL applications in breast cancer diagnosis and motivates an ML-based classification project using reproducible methods and transparent evaluation.

II. Methodology

This literature review was developed through a systematic selection of peer-reviewed journal articles and conference papers related to the application of ML and DL in breast cancer diagnosis and disease classification. The search process employed keywords including "breast cancer diagnosis," "machine learning," "deep learning," "disease classification," "digital pathology," and "medical imaging ML models." Databases consulted included PubMed/PubMed Central, IEEE Xplore, ScienceDirect, and Google Scholar. Articles were included if they discussed ML or DL models applied to breast cancer detection, feature selection, or diagnostic classification with sufficient methodological detail to understand the learning setup and evaluation strategy. Papers were excluded if they lacked methodological detail, did not focus on cancer diagnosis, or did not apply ML or DL algorithms to a diagnostic prediction problem.

III. Literature Review

Disease classification is an ML process in which a model predicts a single label from a defined set of disease categories, using patient data such as medical images, clinical records, genetic information, laboratory results, or structured clinical variables. Its clinical importance lies in differentiating conditions that often present with overlapping manifestations, making early and accurate prediction vital for timely treatment. Prior research demonstrates rapid expansion in ML-based diagnostic methods using both structured data (e.g., demographic and clinical variables) and unstructured data (e.g., imaging) [10]. Reviews emphasize that the choice of algorithm often depends on the data type and the representational complexity of the task. For structured, tabular clinical data, supervised learning methods such as logistic regression, Naive Bayes, support vector machines, decision trees, and random forests are widely used, while deep learning models, particularly convolutional neural networks, dominate imaging applications [10], [14].

Several studies have evaluated ML models across diverse datasets and have emphasized that robust evaluation requires more than a single accuracy value. Moreno-Ibarra et al. compared multiple supervised learning models and reported performance using metrics such as sensitivity, specificity, and balanced accuracy, while also highlighting practical considerations around class distribution and bias in medical classification [11]. Scherr et al. applied diagnostic modeling to host pro-

tein measurements in infectious disease contexts, evaluating multiple model families (including decision trees, neural networks, linear models, gradient boosting, and random forests) and emphasizing reproducibility, data splitting, and generalizability to new populations as central concerns for diagnostic modeling [12]. These broader disease-classification insights inform breast cancer applications, where model choice, feature representation, and validation design strongly influence reported performance.

In breast cancer classification specifically, ML algorithms have been used to distinguish tumor categories by analyzing real-valued features of cell nuclei from digitized fine-needle aspiration (FNA) images, including features such as radius, perimeter, and concavity [15]. La Moglia and Almustafa analyzed a breast cancer dataset with tabular features and showed that logistic regression and LightGBM can achieve strong test accuracy, while also demonstrating that feature selection can shift which models perform best [8]. Yue et al. provide an overview of ML techniques used in breast cancer diagnosis and prognosis and discuss how algorithms such as ANN, SVM, decision trees, and k-NN are commonly evaluated on benchmark datasets, including the Wisconsin Breast Cancer dataset [13]. Brunyé et al. extend the diagnostic modeling space by showing that behavioral interaction features (such as viewing behavior during digital biopsy interpretation) can be used to predict diagnostic accuracy; their results report a random forest classifier with test accuracy of 0.81 and AUC of 0.86 [16]. Finally, review work in breast cancer ML highlights ongoing methodological gaps and therapeutic implications, reinforcing that model transparency, validation design, and careful feature handling are necessary for credible clinical translation [9].

Across the literature, a consistent conclusion is that disease classification performance depends heavily on data type, dataset quality, feature engineering, and evaluation strategy [10], [11]. Feature selection is repeatedly identified as a critical driver of performance and reliability, particularly when prioritizing clinically meaningful predictors to reduce complexity without sacrificing accuracy [8], [15]. Systematic reviews further suggest a transition from traditional classifiers (e.g., SVMs) toward sophisticated DL architectures that can emulate aspects of pathologist-level tissue analysis and support clinical decision-making at scale [9], [14]. However, limitations persist, including heterogeneous datasets, inconsistent preprocessing, varying metrics, and reduced external validity when models are evaluated on narrow cohorts or single-site datasets [10], [11].

IV. Synthesis and Conclusion

Overall, the reviewed literature suggests that successful disease classification in breast cancer requires careful data preprocessing, thoughtful feature engineering, and evaluation metrics that go beyond simple accuracy [10]–[12]. Comparative findings reinforce that algorithm selection must be evidence-based and tailored to dataset characteristics, with clear reporting of data splits and validation procedures [11]. While classical ML methods (e.g., SVMs and random forests) remain strong baseline approaches for structured/tabular data, ongoing advances in deep learning aim to emulate pathologist-level decision-making and support personalized medicine [9], [14]. Future research must prioritize external validation, larger and more diverse datasets, and pathways for clinical integration to ensure practical applicability [10]. Guided by these gaps, our project will implement and compare multiple ML classifiers on a well-known breast cancer benchmark dataset and will evaluate an ensemble strategy with transparent preprocessing and class-sensitive performance reporting.

References

[1] J. Makki, "Diversity of breast carcinoma: Histological subtypes and clinical relevance," *Clinical Medicine Insights: Pathology*, vol. 8, pp. 23–31, 2015, doi: 10.4137/CPath.S31563.

[2] National Cancer Institute, "Cancer Stat Facts: Female Breast Cancer," SEER, 2025. [Online]. Available: https://seer.cancer.gov/statfacts/html/breast.html. Accessed: Feb. 20, 2026.

[3] American Cancer Society, *Cancer Facts & Figures 2025*. Atlanta, GA, USA: American Cancer Society, 2025. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2025/2025-cancer-facts-and-figures-acs.pdf. Accessed: Feb. 20, 2026.

[4] American Cancer Society, "Breast Cancer Statistics: How Common Is Breast Cancer?," 2026. [Online]. Available: https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html. Accessed: Feb. 20, 2026.

[5] Centers for Disease Control and Prevention, "U.S. Cancer Statistics: Female Breast Cancer Stat Bite," 2025. [Online]. Available: https://www.cdc.gov/united-states-cancer-statistics/publications/breast-cancer-stat-bite.html. Accessed: Feb. 20, 2026.

[6] C. Shang and D. Xu, "Epidemiology of breast cancer," *Oncologie*, vol. 24, no. 4, pp. 649–663, 2022, doi: 10.32604/oncologie.2022.027640.

[7] P. Jaglan, R. Dass, and M. Duhan, "Breast cancer detection techniques: Issues and challenges," *J. Institution of Engineers (India): Series B*, vol. 100, no. 4, pp. 379–386, 2019, doi: 10.1007/s40031-019-00391-2.

[8] A. La Moglia and K. M. Mohamad Almustafa, "Breast cancer prediction using machine learning classification algorithms," *Intelligence-Based Medicine*, vol. 11, art. 100193, 2025, doi: 10.1016/j.ibmed.2024.100193.

[9] A. Jafari, "Machine-learning methods in detecting breast cancer and related therapeutic issues: A review," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 12, no. 1, art. 2299093, 2024, doi: 10.1080/21681163.2023.2299093.

[10] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare*, vol. 10, no. 3, art. 541, 2022, doi: 10.3390/healthcare10030541.

[11] M. A. Moreno-Ibarra, Y. Villuendas-Rey, D. López-Sánchez, and J. H. Arroyo-Núñez, "Classification of diseases using machine learning algorithms: A comparative study," *Mathematics*, vol. 9, no. 15, art. 1817, 2021, doi: 10.3390/math9151817.

[12] T. F. Scherr, C. E. Douglas, K. E. Schaecher, R. J. Schoepp, K. M. Ricks, and C. J. Shoemaker, "Application of a machine learning-based classification approach for developing host protein diagnostic models for infectious disease," *Diagnostics*, vol. 14, no. 12, art. 1290, 2024, doi: 10.3390/diagnostics14121290.

[13] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, art. 13, 2018, doi: 10.3390/designs2020013.

[14] M. F. Manzoor, "Machine learning for early disease diagnosis: A review of techniques in healthcare applications," *Premier Journal of Science*, 2024. [Online]. Available: https://premierscience.com/pjs-24-579/. Accessed: Feb. 20, 2026.

[15] C. G. Yedjou, S. S. Tchounwou, R. A. Aló, R. Elhag, B. Mochona, and L. Latinwo, "Application of machine learning algorithms in breast cancer diagnosis and classification," *International Journal of Science and Academic Research*, vol. 2, no. 1, pp. 3081–3086, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34825131/. Accessed: Feb. 20, 2026.

[16] T. T. Brunyé, K. Booth, D. Hendel, K. F. Kerr, H. Shucard, D. L. Weaver, and J. G. Elmore, "Machine learning classification of diagnostic accuracy in pathologists interpreting breast biopsies," *Journal of the American Medical Informatics Association*, vol. 31, no. 3, pp. 552–562, 2024, doi: 10.1093/jamia/ocad232.

[17] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)" [Dataset], UCI Machine Learning Repository, 1993, doi: 10.24432/C5DW2B.