A Data-Driven Approach to Breast Cancer Prediction: Integrating Machine Learning for Clinical Decision Support

First Author1[0000-1111-2222-3333] and Second Author2[1111-2222-3333-4444]

1 Princeton University, Princeton NJ 08544, USA

2 Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany  
lncs@springer.com

**Abstract.** The abstract should summarize the contents of the paper in short terms, i.e. 150-250 words.

**Keywords:** Breast Cancer, Machine Learning, Clinical Disease Support,Predictive Modeling, Artificial Intelligence, Healthcare Analytics.

1. Introduction

Breast cancer is one of the most prevalent cancers worldwide. It is one of the most common malignancies affecting women worldwide and represents a significant global health challenge. It occurs when cells in the breast begin to grow in an abnormal and uncontrolled way. The early symptoms in a patient often serve as warning signs of the development of cancerous cells in the body. According to GLOBOCAN 2020 report, cancer is still a major cause of global health problems with 19.3 million new cases and 10 million deaths reported in 2020[13]. Among these, female breast cancer accounted for 2.3 million new cases, highlighting a concerning rise in incidence and mortality[14]. Statistically, roughly 1 in 8 women are diagnosed with breast cancer in their lifetime. In South Asian countries like Bangladesh, breast cancer is the second most common cancer among women and is often detected at a late stage, which lowers the chances of effective treatment[15]. Some of the major methods of treating breast cancer are surgery, chemotherapy, and radiation therapy and all these play a vital role in the treatment of the disease. It is difficult to diagnose breast cancer due to diversity in tumor appearance, breast density, and false positives. Human error, late detection, and access to medical facilities also affect early diagnosis and accuracy, especially in poor settings. With this ever-growing heavy load, early diagnosis and appropriate identification of breast cancer are highly important in availing timely clinical action and lightening the load on healthcare systems.

In the midst of the development of data-driven healthcare, machine learning (ML) has emerged as a worthy technique of breast cancer prediction by recognizing complex patterns in clinical and diagnostic data. ML algorithms provide early detection by accurate interpretation of medical images to identify tumors at the initial stages. Further, ML models also evaluate a person's susceptibility towards developing breast cancer through aggregation of various patient variables such as genetics and lifestyle. It helps doctors diagnose the disease more accurately based on the classification of the intensity and type of tumor based on biopsy.

1. Literature Review

Shen et al.(2019) developed a CNN model to detect Breast Cancer from biopsies and microscopic images. Here is the efficacy of Shen et al.'s approach was rigorously evaluated on two widely recognized public datasets: the Digital Database for Screening Mammography (DDSM) / CBIS-DDSM and INbreast.For the INbreast dataset, the best single model achieved an impressive per-image AUC of 0.95, with four-model averaging further improving it to 0.98 (sensitivity: 86.7%, specificity: 96.1%) and the DDSM dataset, the best single model achieved a per-image Area Under the Curve (AUC) score of 0.88, which improved to 0.91 with three-model averaging[1].

Siham et al.(2020) focused on how to preprocess data to deal with imbalanced data that have missing values using resampling techniques to enhance the classification accuracy of detecting breast cancer. The three classifiers (NB, SMO, J48) were tested over original data. The accuracy was respectively 71.67%, 69.58%, 75.52%. After applying a discretization filter and removing the records with missing values, results improved. After that, a resample filter was applied for 7 times, then the accuracy was 98.20%(J48), 76.61%(NB), 95.32%(SMO)[2].

The study presented to detect Breast cancer that machine learning technique is good enough on linear data. But, when the data form is imaged the machine learning technique fails. For the classification of the breast cancer images data Kumar (2021) claimed that a deep learning based technique CNN gives better results as compared to machine learning techniques[3].

Kumar al(2022) trained machine learning models with some python libraries such as numpy, pandas, matplotlib. They trained K-nearest neighbor(KNN), Support vector machine(SVM), Decision tree classifier with that library. They used the Breast malignant growth dataset which was recovered utilizing the UCI archive. The accuracy of those models were 95%(KNN), 96%(SVM) and 93%(DTC)[4].

The study introduced the internal functionality of machine learning algorithms. Abien (2018) used the Wisconsin Diagnostic Dataset to train models. GRU-SVM, Linear Regression, MLP, Nearest Neighbor Softmax Regression, SVM algorithms used to train data. The activation function was used for GRU-SVM as like "Sigmoid" or "ReLU". For Nearest Neighbor used Norm function L1 and L2 both. MLP used ReLU as an activation function. Then, trained the models with 128 batch size and 3000 epochs (Nearest Neighbor's is just 1 epoch). The accuracy of those models was above 90%[5].

Breast al (2020) proposed a new method that Deep Neural Network with Support Value(DNNS) which introduced better quality images and fixed other performance parameters. They discussed the internal process of DNNS. Then, they analyzed the performance according to some other methodology like Naive Bayes, SVM, RCNN classifier, Bidirectional Recurrent Neural Networks.The accuracy of the DNNS method was 97.21 and it was the best result of all of those methods[6].

Liu et al. (2024) developed a clinical decision support tool using SHAP values to predict breast cancer recurrence. The study showed a high predictive accuracy of 0.97 with Extra Trees and 0.96 with Random Forest but also pointed out key limitations, especially concerning data quality and the limited size and diversity of the dataset[7].

Gurcan (2025) introduced an advanced framework for breast cancer diagnosis using deep learning combined with stacking ensemble techniques. The framework incorporates models such as LightGBM, CatBoost, and a CNN-based meta-predictor. The proposed model delivers high accuracy, enhanced F1 scores, and faster training times to support efficient healthcare decision-making. However, the use of complex deep learning and ensemble models increases computational demands, potentially limiting real-time use in resource-constrained settings[8].

Ayepeku (2024) conducted a comprehensive analysis of breast cancer prediction comparing various machine learning models such as Logistic Regression, Random Forest, Support Vector Classifier, and ensemble methods like Gradient Boosting and AdaBoost. The study uses comprehensive metrics and visualization to assess models but faces typical issues with data quality and availability, while lacking discussion on real-time deployment, which limits practical clinical integration despite strong analytical insights[9].

Patil et al. (2023) investigated early breast cancer prediction by comparing multiple machine learning and deep learning techniques, including SVM, KNN, Naïve Bayes, Logistic Regression, Random Forest, Decision Tree, XGB Classifier, and Artificial Neural Networks (ANN). The study highlights feature selection’s role in boosting accuracy, with SVM reaching 98.24%, but also notes ongoing challenges with data quality and availability. Moreover, the paper highlights model performance comparisons but overlooks real-time deployment and clinical integration, limiting practical use for early breast cancer detection[10].

Öznacar & Ergene (2024) examined the potential of machine learning techniques, including AdaBoost, SVM, Random Forest, and Logistic Regression, for early detection and malignancy prediction in breast cancer. The study highlights that the AdaBoost model showed the highest performance reaching 93.60% AUC and 95.65% precision. However, the study emphasizes the limitations of traditional breast cancer diagnostics, advocating for AI-based improvements, yet it overlooks challenges like data bias and the need for broad clinical validation[11].

Rb et al. (2024) proposed a novel approach for early breast cancer prediction using an ensemble of machine-learning algorithms, including KNN, Naive Bayes, SVM, and Decision Tree Classifier, enhancing predictive capabilities and improving early detection and patient outcomes. The authors demonstrated that using ensemble machine learning algorithms enhances early breast cancer detection accuracy, offering practical benefits for regions with limited medical resources. Additionally, the study emphasizes that Logistic Regression offers efficient and interpretable results, while SVM excels in accuracy with high-dimensional data[12].

### Sample Heading (Third Level). Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

#### Sample Heading (Forth Level). The contribution should contain no more than four levels of headings. The following Table 1 gives a summary of all heading levels.

**Table 1.** Table captions should be placed above the tables.

|  |  |  |
| --- | --- | --- |
| Heading level | Example | Font size and style |
| Title (centered) | **Lecture Notes** | 14 point, bold |
| 1st-level heading | **1 Introduction** | 12 point, bold |
| 2nd-level heading | **2.1 Printing Area** | 10 point, bold |
| 3rd-level heading | **Run-in Heading in Bold.** Text follows | 10 point, bold |
| 4th-level heading | *Lowest Level Heading.* Text follows | 10 point, italic |

Displayed equations are centered and set on a separate line.

*x* + *y* = *z* ()

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

**Fig. 1.** A figure caption is always placed below the illustration. Short captions are centered, while long ones are justified. The macro button chooses the correct format automatically.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], as well as a URL [5].

References

1. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, *9*(1), 12495.
2. Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5* (pp. 108-117). Springer Singapore.
3. Priyanka, K. S. (2021). A review paper on breast cancer detection using deep learning. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012071). IOP Publishing.
4. Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 114-118). IEEE.
5. Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd international conference on machine learning and soft computing* (pp. 5-9).
6. Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *Ict Express*, *6*(4), 320-324.
7. Liu, Y., Fu, Y., Peng, Y., & Ming, J. (2024). Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. *Heliyon*. <https://doi.org/10.1016/j.heliyon.2024.e24876>
8. Gurcan, F. (2025). *Enhancing Breast Cancer Prediction Through Stacking Ensemble and Deep Learning Integration*. <https://doi.org/10.7717/peerj-cs.2461>
9. Ayepeku, O. F. (2024). Analysis and Visualization of Breast Cancer Prediction through Machine Learning Models. *Jurnal Sistem Informasi*, *13*(3), 1178. <https://doi.org/10.32520/stmsi.v13i3.4100>
10. Patil, S. B., Patil, R. V., & Mahalle, P. N. (2023). Early Breast Cancer Prediction using Machine Learning and Deep Learning Techniques. *International Journal on Recent and Innovation Trends in Computing and Communication*. <https://doi.org/10.17762/ijritcc.v11i10s.7603>
11. Öznacar, T., & Ergene, N. (2024). A Machine Learning Approach to Early Detection and Malignancy Prediction in Breast Cancer. *International Journal of Computational and Experimental Science and Engineering*. <https://doi.org/10.22399/ijcesen.516>
12. Rb, S., V, S., U, M., & Ts, Y. (2024). *Early Breast Cancer Prediction Using Machine-Learning Algorithms*. 1–7. <https://doi.org/10.1109/adics58448.2024.10533612>
13. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249. <https://doi.org/10.3322/caac.21660>
14. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*(6), 394–424. <https://doi.org/10.3322/caac.21492>
15. International Agency for Research on Cancer. (2022). *Global Cancer Observatory*. <https://gco.iarc.fr>