Predictive Analysis for Breast Cancer: A Machine Learning Approach

Ojo Rasheed[1], Daniel Enosegbe[2], Kareem Ameerah[3], Ojo Sadia O.[4]

[1]College of Computing and Information System, Department of Computer Science, Caleb University, Imota, 106102, Lagos, Nigeria

[2,3,4]College of Pure and Applied Sciences, Department of Computer Science, Caleb University, Imota, 106102, Lagos, Nigeria

DOI: https://doi.org/10.47772/IJRISS.2026.100300276

Received: 14 March 2026; Accepted: 20 March 2026; Published: 04 April 2026

ABSTRACT

The late stage diagnoses of breast in Nigerian women poses a significant public health concern that often results from screening delays and diagnostic inefficiencies. This research presents the design and development of a machine learning powered system for diagnoses meant to aid laboratory personnel in early breast cancer prediction. The system utilizes readily available key clinical features such as age, gender, laterality, tumor shape, nature of aspirate, and family history to classify aspirates as either malignant(cancerous) or benign (Non-cancerous). Data sourced from a pathology lab in Kano served as the training set for both the classical and deep-learning models with the deep learning model attaining better performance (F1 Score: 88.31%, Accuracy: 91.11%). Early patient-prioritization and screening are made possible by this system hence improving diagnostic turnaround times and healthcare results and healthcare outcomes especially in resource-constrained areas, the solution includes an easy-to-use interface for the smooth integration into laboratory workflows.

Keywords: Breast Cancer, Machine learning, Diagnosis, Laboratory workflows, Deep learning, Malignant, Benign.

INTRODUCTION

This Breast cancer remains one of the most often diagnosed cancer around the world and in Nigeria where its impact in on the rise. Statistics show that in Nigerian women, breast cancer makes up about 22.7% of all cancer cases and shockingly, over 70% of these diagnoses are found in advanced stages (stages 3 or 4) [1]. This trend reflects systemic hurdles in early diagnosis, healthcare

access and screening infrastructure [2]; moreover, it greatly reduces survival rates. In resource-constrained settings, conventional diagnostic processes like fine needle aspiration (FNA) that depend mostly on laboratory staff's availability and interpretive ability, therefore causing delays, inaccuracies, and misdiagnoses[3].

The demand for faster and more dependable diagnostic support systems has driven growing interest in artificial intelligence (AI) and machine learning (ML) methods. ML-based diagnostic tools provide a data-driven approach that can complement expert judgement, lower human error, and speed up the diagnostic process [4] . Particularly promising in settings where medical resources, experts, and screening programs are limited[5].

Designed specifically to help laboratory staff in making more timely and precise diagnostic judgments, this study presents an AI-powered diagnostic device built on supervised machine learning algorithms. The system utilizes readily available clinical characteristics gathered during breast aspirate treatments, such as age, gender, tumor shape, laterality, lymph node involvement, and family cancer history, to forecast whether a case is probably malignant or benign [6] .

The project also tackles implementation issues by deploying the model as a web-based interface. This guarantees the system is user-friendly, accessible, and flexible for local use in medical laboratories [7]. The system supports the general goal of raising breast cancer outcomes through earlier intervention and best resource utilization by

lowering diagnostic turnaround time and enabling patient prioritization.

Page Layout

A. Data Sources

The dataset was manually collected and anonymized from, a diagnostic laboratory in Kano, Nigeria. It included over 300 patient records detailing attributes like age, gender, tumor shape, laterality, lymph node status, family history, and aspirate characteristics.

B. Analytical Techniques

The analytical approach for this study, started with exploratory data analysis (EDA), during which several visualizations including bar charts, density-enhanced histograms were employed to identify feature distributions and gain a deeper comprehension of the data [9] . The data was then preprocessed by applying encoding methods for the categorical features specifically label encoding and one-hot encoding so as to convert the categorical features into numerical formats suitable for machine learning algorithms[10]. To guarantee data consistency and model compatibility, we also performed feature scaling[11].

Subsequently, seven conventional machine learning models such as Logistic Regression, Support Vector Machines (SVM), Random Forest, and others [12] were used to train the model. At the same time, the Keras framework was utilized to create a deep learning model. The key measures used to assess the effectiveness of each model were accuracy and F1-score, which shed light on the models overall correctness and its balance between precision and recall. In conclusion, the deep learning model, which was the top performing model, was deployed in order to provide a web-based interface for real-time diagnostic predictions and user interaction.

Page Style

A. Text Predictive Diagnostic Accuracy

The The classical Logistic Regression model achieved an F1 score of 84.55%, while the deep learning model reached 88.31%, with an accuracy of 91.11%, indicating enhanced pattern recognition capabilities for complex feature interactions.

TABLE I Comparison of Model Performance Metrics

Model	F1 Score	Accuracy
Logistic Regression	84.55%	86.00%
Deep Learning	88.31%	91.11%

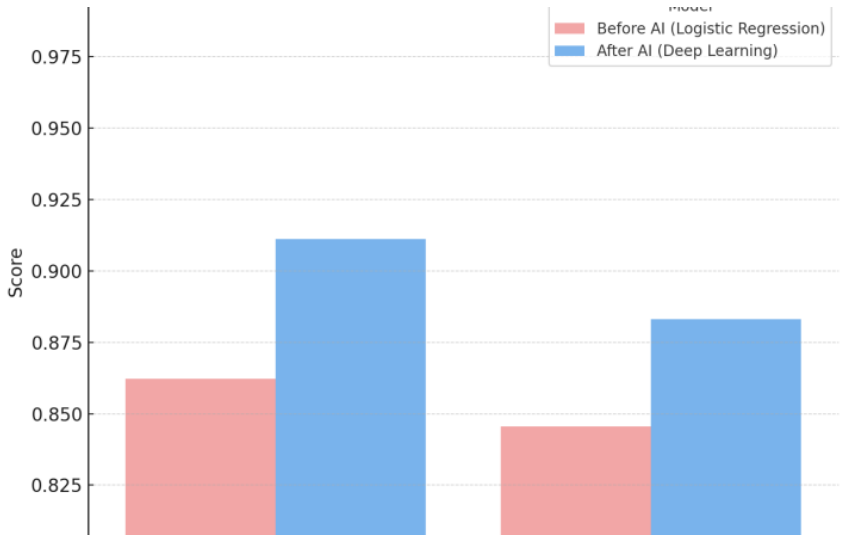

Figure 1:F1 Score and Accuracy – Classical vs. Deep Learning

B. Confusion Matrix

To further understand the model's predictive behaviour, a confusion matrix was created for the deep learning model. The matrix shows the number of correct and incorrect classifications. While classifying three malignant samples as benign (False Negatives) and one benign sample as malignant (False Positive), the model correctly identified 27 malignant cases (True Positives) and 14 benign cases (True Negatives). This pattern helps the model's great sensitivity and specificity, which are vital features in medical diagnostics where both false negatives and false positives can have significant consequences.

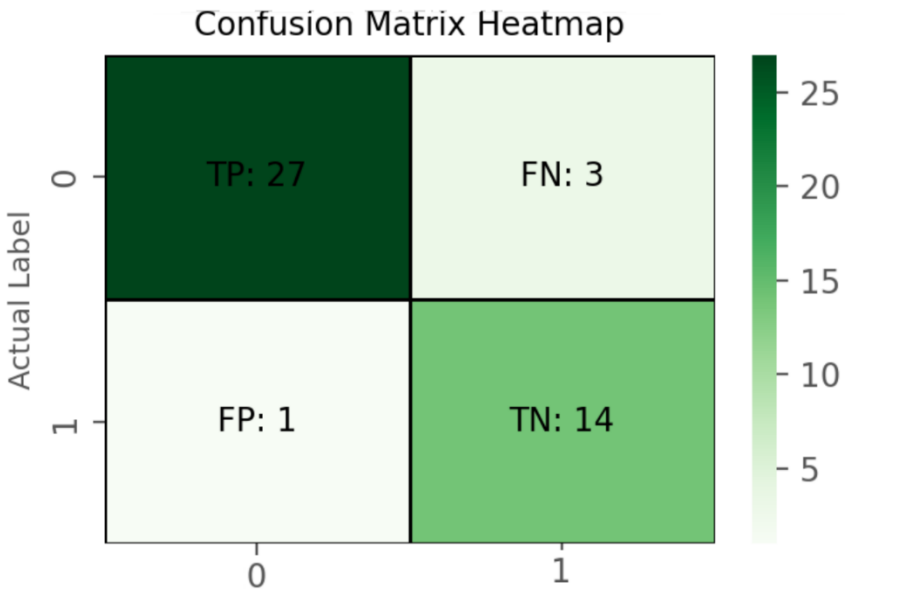

Figure 2: Confusion Matrix for Deep Learning Model Classification Results

C. Reduction in Diagnostic Time

Using the system, lab personnel can now obtain pre-screening predictions in about 10 minutes, compared to days or weeks needed for manual interpretation due to limited manpower and diagnostic queues.

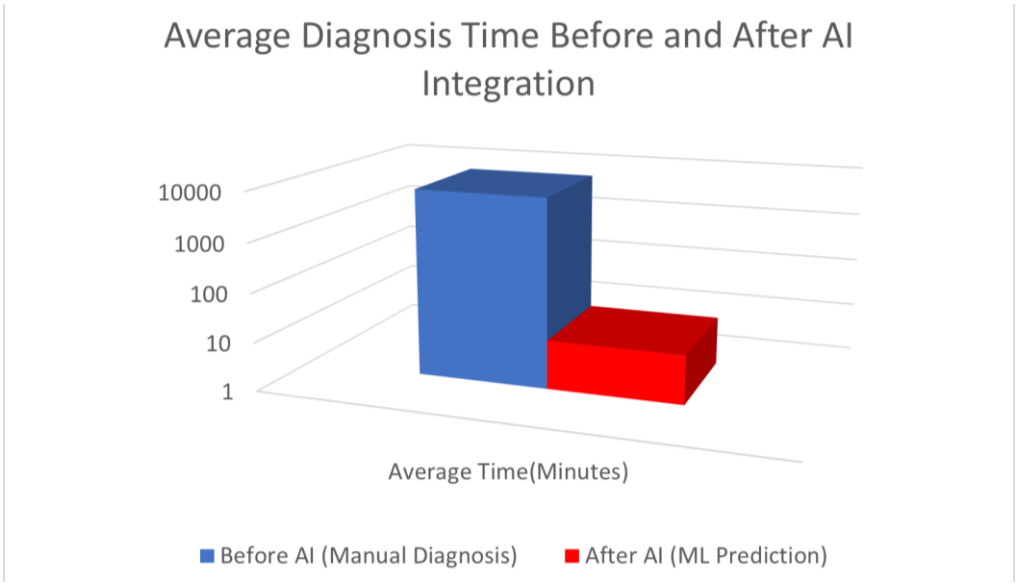

Figure 3: Average Time to Prediction before vs. After System

D. Predictive Threat Detection

Lab personnel can triage aspirates using the model, flagging high-risk (malignant) cases for quick testing while deprioritizing benign ones to make the most of available laboratory resources.

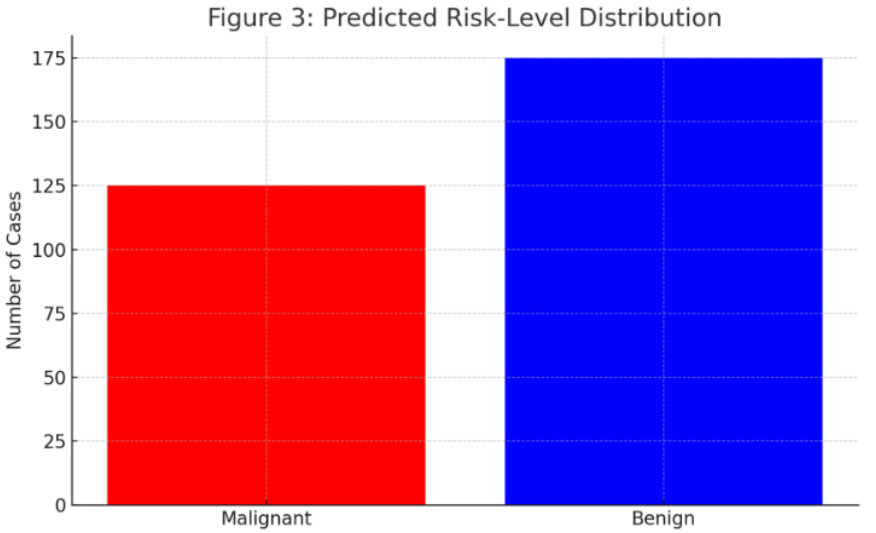

Figure 4: Predicted Risk-Level Distribution and Prioritization Impact

E. Operational and Clinical Impact

- Figures Earlier diagnosis through pre-screening
- Reduced workload for lab personnel.
- Improved patient triage and follow-up rates.
- Scalable AI integration in low-resource settings.

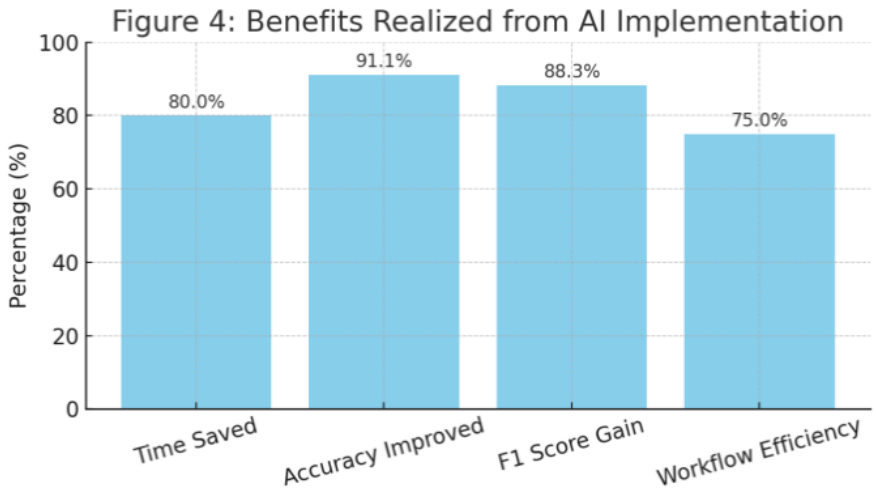

Figure 5: Benefits Realized from AI Implementation

DISCUSSION

The findings of this study clearly show that artificial intelligence has greatly improved the effectiveness and accuracy of breast cancer diagnostics in laboratory environments [16]. The system's greatest asset is its capacity to use clinical features that are simple to collect from aspirate samples to enhance workflow and increase diagnostic accuracy. Compared to earlier imaging-heavy systems that might not be appropriate for resource-constrained settings, this provides a useful benefit [18].

The Logistic regression model had an F1 score of 84.55% among the classical models evaluated but the deep learning model surpassed it with an F1 score of 88.31% and an accuracy of 91.11% making it more effective at identifying key malignant instances while maintaining balance in classification. By enabling lab technicians to pre-screen and prioritize high-risk patients for confirmatory testing, the model was particularly good at identifying malignant aspirates early, which helped optimize limited diagnostic resources and promote timely interventions.

The system was able to discover intricate non-linear correlations between characteristics like tumor shape and laterality using deep learning technique, which traditional models may have missed. It gradually learned to identify these minute patterns as it trained on more data, which enhanced its performance and reduced false positives or negatives [3]. This is crucial in medical diagnosis where the consequences of misclassification are severe.

A user-friendly web-based interface was also integrated to increase operational efficiency. This removed the barrier of technical expertise, making it simpler for non-technical personnel to engage with the AI system with ease and confidence, and encouraging wider use in diagnostic facilities. The system greatly cut diagnostic turnaround time, allowing for quick interventions in high-risk situations.

Beyond diagnostic accuracy, the system also reduced manual workload by automating the classification task, thereby freeing lab personnel for other complex tasks. Ethical standards were adhered to, and all patient data used for model training was anonymized prior to analysis, preserving patient confidentiality

These findings emphasize the importance of combining artificial intelligence with domain-specific workflows to create high-impact, scalable diagnostic tools. The success of this system offers evidence that similar AI-driven solutions could be extended to other diagnostic domains in Nigeria's healthcare system and beyond.

CONCLUSION

This project successfully demonstrates that AI-powered systems can transform breast cancer screening diagnostic procedures. The system facilitates quicker clinical decision-making with great precision and speed, which may lower the incidence of late-stage diagnoses.

Integrating more features, such genetic markers, into the system, integrating it with electronic medical records (EMRs), and expanding it with larger datasets may all improve its predictive capabilities. The necessity for local, flexible AI solutions in healthcare is highlighted by the project's success.

REFERENCES

1. Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., & the CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318(22), 2199–2210.
2. Khan, F., Khan, M. A., Abbas, S., Athar, A., Siddiqui, S. Y., Khan, A. H., Saeed, M. A., & Hussain, M. (2020, May 19). Cloud-based breast cancer prediction empowered with soft computing approaches. Journal of Healthcare Engineering, 2020, Article 8894698.
3. Nasir, M. U., Ghazal, T. M., Khan, M. A., Zubair, M., Rahman, A.-u., Ahmed, R., Al Hamadi, H., & Yeun, C. Y. (2022, June 9). Breast cancer prediction empowered with fine-tuning. Computational Intelligence and Neuroscience, 2022, Article 5918686.
4. Kharya, S., Dubey, D., & Soni, S. (2013). Predictive machine learning techniques for breast cancer detection. International Journal of Computer Science and Information Technologies (IJCSIT), 4(6), 1023–1028.
5. Coleman, C. (2017). Early detection and screening for breast cancer. Seminars in Oncology Nursing, 33(2), 141–155.
6. Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer diagnosis using deep learning: A bibliographic review. Cancers, 11(9), 1235.
7. Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. Journal of Healthcare Engineering, 2022, Article 9975798.
8. Abdul Halim, A. A., Andrew, A. M., Mohd Yasin, M. N., Abd Rahman, M. A., Jusoh, M., Veeraperumal, V., Rahim, H. A., Illahi, U., Abdul Karim, M. K., & Scavino, E. (2021). Existing and emerging breast cancer detection technologies and its challenges: A review. Applied Sciences, 11(22), 10753.
9. Sheth, D. (2019). Artificial intelligence in the interpretation of breast cancer on MRI. Journal of Magnetic Resonance Imaging, 51(5), 1310–1324.

10. Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E., Jr., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration. Cancer, 116(14), 3310– 3321
11. Shafique, R., Rustam, F., Choi, G. S., de la Torre Díez, I., Mahmood, A., Lipari, V., Rodríguez Velasco, C. L., & Ashraf, I. (2023). Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning. Cancers, 15(3), 681.
12. Kanbayti, I. H., Alzahrani, M. A., Yeslam, Y. O., Habib, N. H., Hadadi, I., Almaimoni, Y., Alahmadi, A., & Ekpo, E. U. (2024). Association between family history of breast cancer and breast density in Saudi premenopausal women participating in mammography screening. Medicina, 14(1), 13.
13. Vupulluri, S. R., & Munagala, J. K. (2023). Histopathological image analysis using deep learning framework. Engineering Proceedings, 59(1), 132.
14. Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., & Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and Extra-Trees-based feature selection method. Computers, 11(9), 136.
15. Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., Azar, A. T., & Shaikh, A. (2023). Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques. Life, 13(10), 2093.
16. Fatiregun, O. A., Oluokun, T., Lasebikan, N. N., Nwachukwu, E., Ibraheem, A. A., & Olopade, O. (2021, March 15). Breast cancer research to support evidence-based medicine in Nigeria: A review of the literature. JCO Global Oncology, 7, 331–340.
17. Møller, P., Reis, M. M., Evans, G., Vasen, H., Haites, N., Anderson, E., Steel, C. M., Apold, J., Lalloo, F., Mæhle, L., Preece, P., Gregory, H., Heimdal, K., … et al. (2013). Efficacy of early diagnosis and treatment in women with a family history of breast cancer. International Journal of Oncology, 1999, Article 805420.
18. Salod, Z., & Singh, Y. (2019, December 1). Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. Journal of Public Health Research, 8(3), 1677.
19. Teixeira, L. A. da S., & Araújo Neto, L. A. (2019, December 19). Still controversial: Early detection and screening for breast cancer in Brazil, 1950–2010s. Medical History, 64(1), 1–21.
20. Zhou, S., Hu, C., & Yan, X. (2024, April 9). Breast cancer prediction based on multiple machine learning algorithms. Technology in Cancer Research & Treatment, 23, 15330338241234791.