World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Insights into breast cancer: A simple machine learning method for early disease detection

Anthony Sowah [1], Titus Santigie-Sankoh [2], Vero Bai-Anku [2,*] and Eric Jhessim [3]

[1] *Department of Electrical Engineering, Centrale Nantes University, France.*
[2] *Department of Technology, Njala University, Sierra Leone.*
[3] *Department of Electrical and Computer Engineering, University of Delaware, USA.*

World Journal of Advanced Research and Reviews, 2025, 25(01), 1357-1360

Publication history: Received on 01 December 2024; revised on 13 January 2025; accepted on 15 January 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.1.0179

Abstract

Breast cancer remains a significant global health challenge, where accurate prediction plays a vital role in early diagnosis and effective treatment, ultimately saving lives. This study evaluates the performance of three machine learning models; Support Vector Machine (SVM), Random Forest Classifier, and XGBoost for breast cancer prediction. Using the Wisconsin Breast Cancer Dataset, the models were assessed based on their accuracy. The experimental results demonstrated that SVM outperformed the other models, while both XGBoost and the Random Forest Classifier achieved just slightly lower accuracies. This research underscores the potential of machine learning models in enhancing breast cancer prediction and highlights their importance in advancing early detection and treatment strategies.

Keywords: Support Vector Machines; Random Forest Classifier; XGBoost; Machine Learning; Breast Cancer

1. Introduction

Breast cancer remains a critical global health challenge, and the ability to accurately predict diagnoses is vital for early detection and effective treatment (Adir et al., 2020; Hu et al., 2018). It occurs when abnormal cells in the breast grow uncontrollably. While some breast lumps are benign and do not pose a direct threat, others are malignant and can spread to other parts of the body. Although benign lumps do not metastasize, they may increase the likelihood of developing breast cancer over time. Therefore, distinguishing between benign and malignant tumors is essential for timely medical intervention.

Among women worldwide, breast cancer is one of the most common cancers, underscoring the importance of early diagnosis and precise prediction in improving treatment outcomes. Recent data highlights over 250,000 new cases of invasive breast cancer and over 50,000 cases of non-invasive breast cancer are anticipated annually in the United States (Adrah et al., 2023; Alam et al., 2015; Chow & Ho, 2013). The rise of machine learning in recent years has provided promising tools for breast cancer prediction (Adrah et al., 2024.; Agboklu et al., 2024). By analyzing complex patterns and features within medical data, these computational models assist in identifying malignancies and supporting healthcare professionals in making well-informed decisions (Aminizadeh et al., 2023; Bhardwaj et al., 2017; Hu et al., 2018). Breast cancer is a menace, and has lead families and people to consult even alternative medicine as a resource, similar to other chronic diseases (Denu et al., 2024; Jazieh et al., 2012; Tascilar et al., 2006).

This study aims to leverage machine learning models to accurately classify breast tumors as benign or malignant, thereby contributing to improved diagnostic processes. The structure of this paper is as follows: Section 2 reviews related work, Section 3 outlines the methodology, and Sections 4 and 5 present the results and conclusions.

* Corresponding author: Vero Bai-Anku

2. Methodology

The primary objective of this research is to comprehensively evaluate various machine learning models for breast cancer prediction. This study focuses on comparing the performance of three widely used algorithms: Support Vector Machine (SVM), Random Forest Classifier, and XGBoost.

2.1. Machine Learning Models

- **Support Vector Machine (SVM)**: SVM is a robust classification algorithm that determines the optimal hyperplane to separate data into distinct classes using the nearest data points. Renowned for its versatility, SVM handles both linear and non-linear classification tasks with high precision, making it a reliable method for predictive modeling (Alanazi, 2022; Aminizadeh et al., 2023).
- **Random Forest Classifier**: Random Forest is an ensemble learning technique that constructs multiple decision trees to improve classification accuracy and reduce the risk of overfitting. During the training phase, it combines the outputs of individual trees, using majority voting (for classification) or averaging (for regression) to make predictions. Random Forest has demonstrated exceptional performance in handling complex datasets while maintaining robustness and reliability (Bhardwaj et al., 2017).

- **XGBoost (Extreme Gradient Boosting)**: XGBoost is an advanced gradient boosting algorithm designed to enhance predictive accuracy by minimizing errors through optimization of an objective function. It sequentially builds an ensemble of weak learners, such as decision trees, and has gained popularity for its ability to handle complex datasets with high efficiency. XGBoost is particularly effective for breast cancer prediction, offering exceptional performance and reliability in identifying patterns within data (Bhardwaj et al., 2017; Lartey et al., 2023).

By evaluating and comparing these three machine learning models, this study aims to highlight their strengths, limitations, and overall effectiveness in predicting breast cancer. The findings will guide the selection of optimal algorithms for accurate and reliable predictions, contributing to early detection and improved patient outcomes.

3. Data collection

The research leverages the Wisconsin Breast Cancer Dataset, a widely recognized and extensively studied dataset in the field of breast cancer research. It offers a rich set of features that are instrumental in facilitating accurate predictions of breast cancer outcomes.

4. Results and discussion

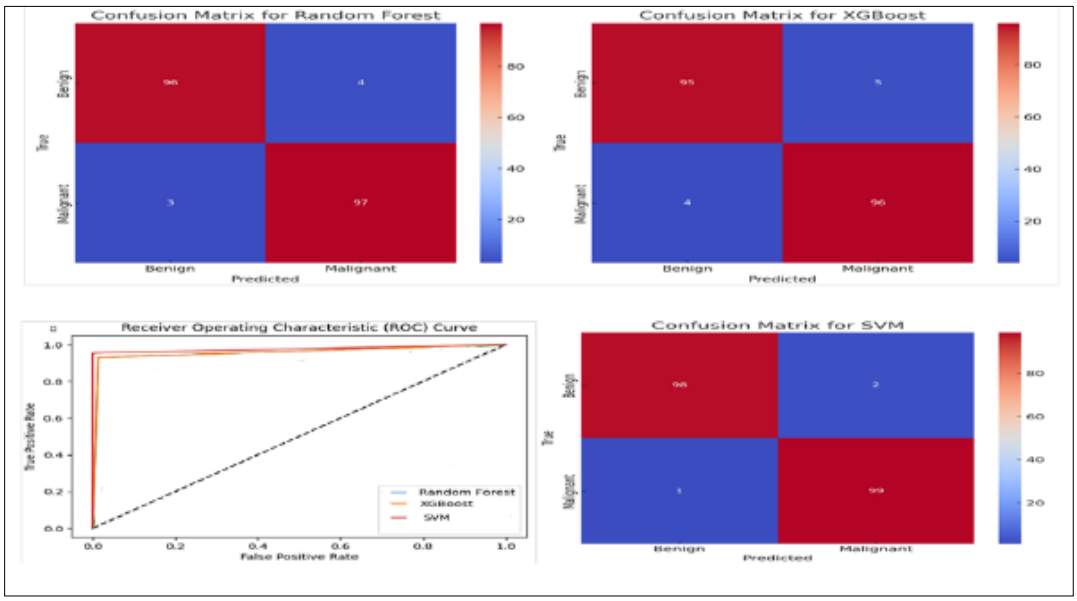

Figure 1 Confusion Matrices and ROC curve

This dataset comprises samples derived from breast masses, with each sample characterized by multiple features extracted from digitized images. These features are critical in identifying key patterns and attributes that differentiate benign tumors from malignant ones. The target variable in the dataset indicates the classification of each sample as either "Benign" or "Malignant," with these labels determined by expert medical diagnoses and serving as the ground truth for training and evaluating machine learning models.

To ensure data quality and reliability, the dataset underwent preprocessing to address missing values and outliers, thereby reducing potential biases and inaccuracies. Additionally, the data was split into training and testing sets using an 80:20 ratio, enabling robust evaluation and validation of the models. This structured approach ensures that the dataset provides a solid foundation for developing effective p*redictive algorithm

5. Conclusion

These results allow us to draw the conclusion that SVM, Random Forest Classifier, and XGBoost are good machine learning models for predicting breast cancer. These models can help medical practitioners correctly categorize cases of breast cancer, which is essential for prompt detection and effective treatment planning. These models will be useful in clinical situations, as indicated by their excellent accuracy. These results demonstrate significant potential, particularly in the realm of global health. Chronic diseases account for nearly 74% of all deaths worldwide, according to the World Health Organization, with conditions such as cardiovascular diseases, cancer, diabetes, and chronic respiratory diseases causing immense strain on families and economies globally. Machine learning techniques like XGBoost as demonstrated in this research paper have the potential to revolutionize early disease detection, thereby saving millions of lives. Studies have shown that XGBoost models can achieve up to 95+% accuracy in diagnosing conditions such as breast cancer, making them invaluable tools for preventative healthcare and reducing the global burden of chronic diseases.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Adir, O., Poley, M., Chen, G., Froim, S., Krinsky, N., Shklover, J., Shainsky-Roitman, J., Lammers, T., & Schroeder, A. (2020). Integrating artificial intelligence and nanotechnology for precision cancer medicine. *Advanced Materials*, *32*(13), 1901989.

[2] Adrah, F. A., Denu, M. K., & Buadu, M. A. E. (2023). Nanotechnology applications in healthcare with emphasis on sustainable covid-19 management. *Journal of Nanotechnology Research*, *5*(2), 6–13.

[3] Adrah, F. A., Mottey, B. E., & Nyavor, H. (n.d.). The Landscape of Artificial Intelligence Applications in Health Information Systems. *International Journal of Computer Applications*, *975*, 8887.

[4] Agboklu, M., Adrah, F. A., Agbenyo, P. M., & Nyavor, H. (2024). From bits to atoms: Machine learning and nanotechnology for cancer therapy. *Journal of Nanotechnology Research*, *6*(1), 16–26.

[5] Alam, F., Naim, M., Aziz, M., & Yadav, N. (2015). Unique roles of nanotechnology in medicine and cancer-II. *Indian Journal of Cancer*, *52*(1), 1–9.

[6] Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, *30*, 100924.

[7] Aminizadeh, S., Heidari, A., Toumaj, S., Darbandi, M., Navimipour, N. J., Rezaei, M., Talebi, S., Azad, P., & Unal, M. (2023). The applications of machine learning techniques in medical data processing based on distributed computing and the Internet of Things. *Computer Methods and Programs in Biomedicine*, 107745.

[8] Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). *A study of machine learning in healthcare*. *2*, 236–241.

[9] Chow, E. K.-H., & Ho, D. (2013). Cancer nanomedicine: From drug delivery to imaging. *Science Translational Medicine*, *5*(216), 216rv4-216rv4.

[10] Denu, M. K., Buadu, M. A. E., Adrah, F., Normeshie, C. A., & Berko, K. P. (2024). Traditional complementary and alternative medicine (TCAM) use among PLHIV on antiretroviral medication. *AIDS Research and Therapy*, *21*(1), 84.

[11] Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., & Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis− A survey. *Pattern Recognition*, *83*, 134–149.

[12] Jazieh, A. R., Al Sudairy, R., Abulkhair, O., Alaskar, A., Al Safi, F., Sheblaq, N., Young, S., Issa, M., & Tamim, H. (2012). Use of complementary and alternative medicine by patients with cancer in Saudi Arabia. *The Journal of Alternative and Complementary Medicine*, *18*(11), 1045–1049.

[13] Lartey, B., Adrah, K., Adrah, F., & Isichei, J. (n.d.). Application of Machine Learning for Predicting the Occurrence of Nephropathy in Diabetic Patients. *International Journal of Computer Applications*, *975*, 8887.

[14] Tascilar, M., de Jong, F. A., Verweij, J., & Mathijssen, R. H. (2006). Complementary and alternative medicine during cancer treatment: Beyond innocence. *The Oncologist*, *11*(7), 732–741.