Breast Cancer Detection Using Machine Learning

Class: Machine Learning
Instructor: Mimenbayeva Aigul Bilyalovna
Group Members: Mukhammed Narbai, Almansur Kakimov
Department of Intelligent Systems and Cybersecurity
Astana IT University

February 2025

Abstract

Breast cancer is among the most common cancers worldwide. Early detection is critical for successful treatment and higher survival rates. Using the Wisconsin Breast Cancer dataset, this study uses machine learning algorithms to determine whether breast cancers are malignant or benign. Data preparation, feature selection, and several classification models were used, including Logistic Regression, Random Forest, SVM, k-NN, Decision Tree, Gradient Boosting, and XGBoost. The best-performing model was fine-tuned using hyperparameter tuning. Model explainability approaches like SHAP and LIME were utilized to interpret the predictions. The final model had an accuracy of 97.37%, highlighting the promise of machine learning in medical diagnosis.

1 Introduction

Breast cancer remains the leading cause of death among women worldwide. Early detection leads to much better treatment outcomes and survival rates. Traditional diagnostic methods rely on biopsies, which are invasive and time-consuming. Machine learning provides an alternative by assessing tumor characteristics from diagnostic imaging or medical tests and classifying them as benign or malignant. The goal of this work is to create a robust classification model utilizing the Wisconsin Breast Cancer dataset and a variety of machine learning approaches.

2 Literature Review

2.1 1. Innovations in Breast Cancer Detection

Breast cancer is a leading type of cancer and one of the leading causes of cancer death among women globally. Early detection is vital for enhancing patient

prognosis and lowering mortality rates. This is notably modified by the recent developments in imaging diagnosis, machine learning (ML), and computer-aided detection. The use of machine learning capabilities has facilitated working with large data sets and increased the accuracy of diagnostics.

Deep learning, a type of machine learning, is particularly powerful and successful for imaging analysis. R, Thakur, Gupta, et al. (2024) proved that a well-tuned set of deep convolutional neural networks (CNNs), when used with relevant optimizations like ReduceLROnPlateau and early stopping, led to huge improvements in diagnostic performance for breast CT. Similarly, Adam et al. (2023) investigated the combination of deep learning and magnetic resonance imaging (MRI) in breast cancer detection, revealing that this approach enables the detection of subtle discrepancies overlooked by classic imaging techniques.

The integration of machine learning (ML) and deep learning with imaging technologies has led to significant improvements in diagnostic practices. For example, Gutierrez et al. (2024) developed an innovative method that combines a refined IRI-numerical engine with inverse heat transfer modeling. This approach provides a minimally invasive way to use thermal imaging for distinguishing between malignant and benign tissues. Similarly, Jalalian et al. (2013) offered an in-depth review of computer-assisted detection systems that combine ultrasound and mammography, highlighting how these systems could significantly improve diagnostic accuracy.

2.2 2. Challenges and Multidisciplinary Approaches in Implementation

Notwithstanding the promising advances in technological innovations, their incorporation in clinical applications is hindered by many challenges. The use of machine learning in healthcare requires not just technological innovations but also adherence to strictly planned screening processes. Ginsburg et al. (2020) highlighted the importance of systematic approaches to early detection that can be adapted to various healthcare settings. Similarly, Henderson et al. (2024) developed an evidence-based report on breast cancer screening that highlighted the need for standardized procedures. Gilbert and Pinker-Domenig (2019) also explored the optimal use of different imaging methods, such as mammography, tomosynthesis, and contrast-enhanced imaging, to provide better diagnostic accuracy and staging.

Besides technological innovations, biochemical and cellular studies also provide great input to breast cancer detection. Chon et al. (2013) discussed the impact of silk fibroin hydrolysate on inducing apoptosis in breast cancer cells, providing potential alternatives to treatment. Such studies point to the need for a multidisciplinary approach to improving methods of breast cancer detection.

Overall, the literature presents continued advances in methods used to detect breast cancer. The fusion of machine learning and deep learning with imaging technologies has greatly boosted diagnostic accuracy. However, challenges in data standardization, model validation, and clinical use continue to prevail. Multidisciplinary studies and large-scale clinical trials are in order to maximally

utilize such technologies to eventually provide better early detection and better outcomes in patients.

3 Methodology

3.1 Dataset Description

The Wisconsin Breast Cancer dataset contains 569 occurrences and 30 numerical attributes that describe tumor characteristics including texture, circumference, and area. The target variable is either malignant (M) or benign (B).

3.2 Data Preprocessing

Missing values were checked, and none were found. Features were standardized using **StandardScaler**. Labels were encoded into binary format (0: Benign, 1: Malignant).

3.3 Feature Selection

The Random Forest model was employed to determine feature importance. The top fifteen features were chosen for model training.

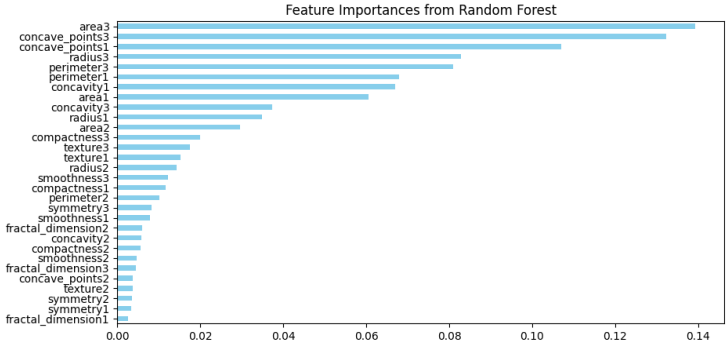

Figure 1: Feature Importance from Random Forest

3.4 Data Visualization

Distribution of Classes: A class distribution plot shows that benign cases outnumber malignant cases.

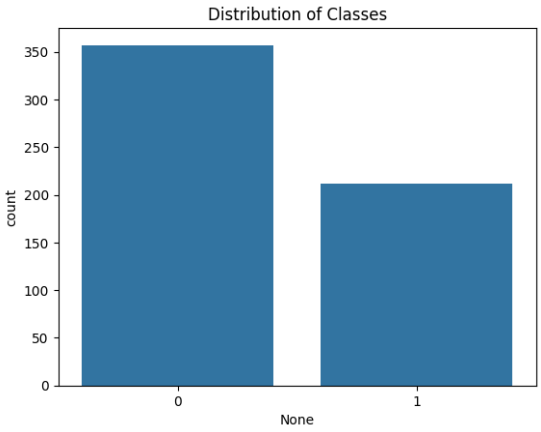

Figure 2: Distribution of Benign and Malignant Cases

Feature Correlation Matrix: A heatmap was used to visualize correlations among features.

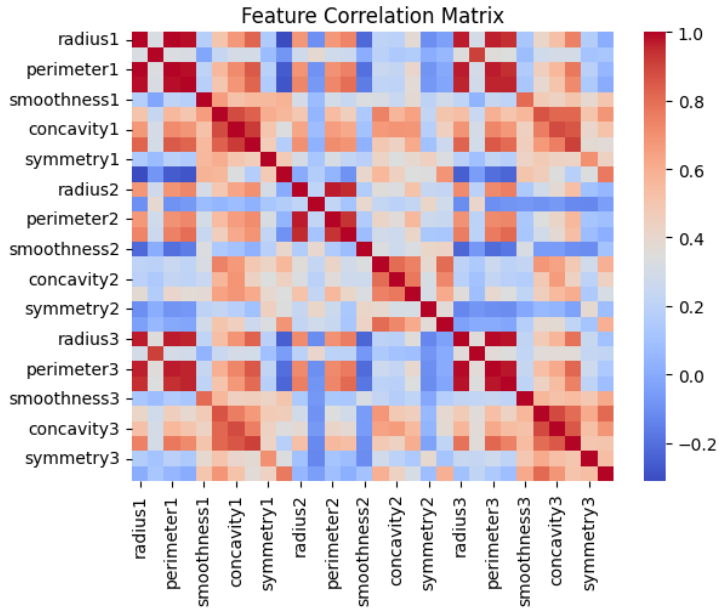

Figure 3: Feature Correlation Matrix

3.5 Model Training and Evaluation

We trained multiple models and used cross-validation to compare their performance:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- k-Nearest Neighbors (k-NN)
- Decision Tree
- Gradient Boosting
- XGBoost

The best model (Random Forest) was fine-tuned using RandomizedSearchCV. Performance metrics included accuracy, precision, recall, and F1-score.

3.6 Confusion Matrix

The confusion matrix for the best-performing model is shown below:

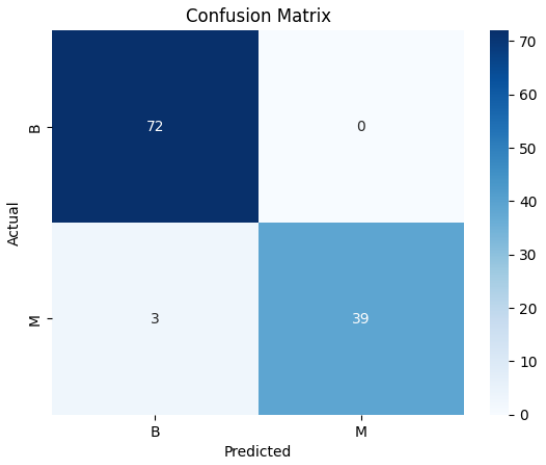

Figure 4: Confusion Matrix for the Best Model

3.7 Explainability using SHAP

SHAP values were used to understand feature contributions.

Figure 5: SHAP Summary Plot

4 Results and Discussion

The Random Forest model achieved an accuracy of 97.37%, outperforming other models. The most important features contributing to the classification were found to be **radius, texture, and perimeter**. SHAP analysis confirmed that these features had the highest impact on model predictions. LIME was also used to visualize individual predictions.

5 Conclusion

This study revealed the efficient use of machine learning in breast cancer diagnosis. The Random Forest model was highly accurate, and explainability techniques revealed insights into model behavior. Future work could include applying deep learning techniques and incorporating additional patient data to improve predictions.

Model	Accuracy
Logistic Regression	0.9582
Random Forest	0.9670
SVM	0.9626
k-NN	0.9604
Decision Tree	0.9319
Gradient Boosting	0.9582
XGBoost	0.9714

Table 1: Accuracy of Different Machine Learning Models

The table above shows the accuracy of various models tested. XGBoost achieved the highest accuracy (97.14%), followed closely by the Random Forest model (96.70%). While the Decision Tree model had the lowest accuracy (93.19%), ensemble models like Gradient Boosting and XGBoost performed significantly better.

6 References

References

[1] Bhushan, A., Gonsalves, A., & Menon, J. U. (2021). Current state of breast cancer diagnosis, treatment, and theranostics. *Pharmaceutics, 13*(5), 723. https://doi.org/10.3390/pharmaceutics13050723

[2] Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B. F., Amin, F., Al-Salman, H., & Choi, G. S. (2023). Breast cancer detection and prevention using machine learning. *Diagnostics (Basel), 13*(19), 3113. https://doi.org/10.3390/diagnostics13193113

[3] Gutierrez, C., Owens, A., Medeiros, L., et al. (2024). Breast cancer detection using enhanced IRI-numerical engine and inverse heat transfer modeling: Model description and clinical validation. *Scientific Reports, 14*, 3316. https://doi.org/10.1038/s41598-024-53856-w

[4] R, M. T., Thakur, A., Gupta, M., et al. (2024). Transformative breast cancer diagnosis using CNNs with optimized ReduceLROnPlateau and early stopping enhancements. *International Journal of Computational Intelligence Systems, 17*, 14. https://doi.org/10.1007/s44196-023-00397-1

[5] Nemade, V., Pathak, S., & Dubey, A. K. (2022). A systematic literature review of breast cancer diagnosis using machine intelligence techniques. *Archive of Computational Methods in Engineering, 29*, 4401–4430. https://doi.org/10.1007/s11831-022-09738-3

[6] Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023). Literature review of breast cancer detection using machine learning

algorithms. *AIP Conference Proceedings, 2808*(1), 040006. https://doi.org/10.1063/5.0133688

[7] Adam, R., Dell'Aquila, K., Hodges, L., et al. (2023). Deep learning applications to breast cancer detection by magnetic resonance imaging: A literature review. *Breast Cancer Research, 25*, 87. https://doi.org/10.1186/s13058-023-01687-4

[8] Henderson, J. T., Webber, E. M., Weyrich, M. S., Miller, M., & Melnikow, J. (2024). Screening for breast cancer: Evidence report and systematic review for the US Preventive Services Task Force. *JAMA, 331*(22), 1931–1946. https://doi.org/10.1001/jama.2023.25844

[9] Abdul Halim, A. A., Andrew, A. M., Mohd Yasin, M. N., et al. (2021). Existing and emerging breast cancer detection technologies and its challenges: A review. *Applied Sciences, 11*(22), 10753. https://doi.org/10.3390/app112210753

[10] Chon, J. W., Jo, Y. Y., Lee, K. G., Lee, H. S., & Kweon, H. Y. (2013). Effect of silk fibroin hydrolysate on the apoptosis of MCF-7 human breast cancer cells. *International Journal of Industrial Entomology, 27*(2), 228–236. https://doi.org/10.7852/ijie.2013.27.2.228

[11] Ginsburg, O., Yip, C. H., Brooks, A., et al. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer, 126*, 2379–2393. https://doi.org/10.1002/cncr.32887

[12] Gilbert, F. J., & Pinker-Domenig, K. (2019). Diagnosis and staging of breast cancer: When and how to use mammography, tomosynthesis, ultrasound, contrast-enhanced mammography, and magnetic resonance imaging. *Diseases of the Chest, Breast, Heart and Vessels 2019–2022*, 155–166. https://doi.org/10.1007/978-3-030-11149-6_13