A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR BREAST CANCER DIAGNOSIS

Ayşegül Yabacı Tak

Abstract

Breast cancer remains one of the most prevalent cancers among women worldwide, and early diagnosis significantly increases the chances of successful treatment. In this study, we evaluate and compare the performance of several supervised machine learning algorithms in classifying breast cancer cases as malignant or benign using the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository. The dataset comprises 569 instances, each with 30 real-valued input features derived from digitized images of fine needle aspirate (FNA) of breast masses. We implemented and assessed Logistic Regression, Support Vector Machines (SVM), XGBoost, Random Forest and Artificial Neural Networks (ANN). Each model was compared based on Accuracy, AUC (Area Under Curve), Precision, Recall, and F1 score. The results show that XGBoost consistently achieved the highest classification accuracy %99, while also demonstrating strong generalization across folds. The study also highlights the impact of preprocessing techniques and hyperparameter tuning on model performance. Our findings emphasize the potential of machine learning in enhancing diagnostic decision support for breast cancer and provide insights into the suitability of various classification algorithms in clinical data applications.

Key words: breast cancer, machine learning, diagnosis, classification, decision support systems

JEL Code: C38, C45

1 Introduction

As is well known, breast cancer has become one of the most common causes of cancer-related deaths among women. Early and accurate diagnosis of breast cancer is highly important in terms of increasing survival rates. In recent years, machine learning methods have emerged as a remarkable potential tool for clinical decision-making processes. In this study, the aim is to compare the diagnostic classification performance of five widely used machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machines (SVM), Artificial

Neural Network (ANN), and Extreme Gradient Boosting (XGBoost)—on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which was obtained from the open-access UCI Machine Learning Repository (Wolberg et al.,1995).

Artificial Neural Networks (ANNs) are based on the working principles of neurons in the human brain and consist of a highly structured and complex network. Each unit behaves similarly to a biological neuron. The concept was first introduced by psychologist Frank Rosenblatt (1958). One of the most widely used models is the Back Propagation Neural Network (BPNN), which is also referred to as a multi-layer feed-forward neural network or multi-layer perceptron (MLP) (Ojha, V. K. et al.,2017; Stattin, P. et al,2014).

Logistic regression is one of the most fundamental and widely used classification algorithms in supervised machine learning. Despite its statistical roots, it has been effectively adapted into the machine learning domain due to its simplicity, interpretability, and solid theoretical foundations. It is particularly useful in binary classification problems where the goal is to estimate the probability of class membership based on input features (Hosmer et al., 2013). Recent studies have emphasized logistic regression's effectiveness as a baseline model in medical diagnostics, credit scoring, and text classification tasks (Kotsiantis, 2007; Wu et al., 2008).

The Support Vector Machine (SVM) classifier is a supervised machine learning method grounded in statistical learning theory. Originally introduced by Cortes and Vapnik (1990) in their foundational research, the algorithm operates by identifying an optimal hyperplane that separates data points into distinct classes. It evaluates how effectively the hyperplane distinguishes between categories by maximizing the margin between them. In essence, SVM aims to achieve high classification accuracy by positioning the decision boundary in a way that best separates the classes on either side.

XGBoost (Extreme Gradient Boosting) is a tree-based model developed by Chen and Guestrin (2016). It operates very efficiently and demonstrates strong performance on large datasets. It is an optimized implementation of the gradient boosting technique.

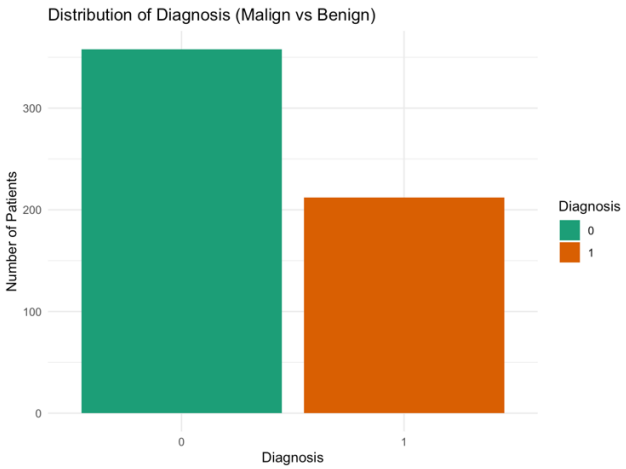
The Random Forest method was proposed by Breiman in 2001. It has high classification accuracy, is robust to outliers, and does not suffer from overfitting. Thanks to these properties, it has become one of the most widely used machine learning methods in big data mining as well as in medical and biological applications.

2 Material and Method

2.1 Dataset

The WDBC dataset consists of variables obtained from the examination of samples taken from breast tumors of 569 women. Each instance includes 30 numerical features such as radius, texture, perimeter, and smoothness. The target variable (diagnosis) is binary, representing malignant tumors as 1 (212(%37.26) and benign tumors as 0 (357(%62.74) *(Fig.1)*.

Fig.1: Distribution of Diagnosis (Based on authors' own calculations)



2.2 Preprocessing

The target variable was encoded as a binary factor (Malignant = 1, Benign = 0). The variables were standardized using Z-score normalization. For the application of machine learning methods, the dataset was split into 80% training and 20% testing sets. Missing values were imputed using an appropriate missing data estimation method.

2.3 Machine Learning Model Development

Five classification models were implemented:

- Logistic Regression
- Random Forest (ntree = 100)
- SVM with RBF kernel
- Artificial Neural Network (size = 5, max iterations = 500)
- XGBoost (max.depth = 3, eta = 0.1, nrounds = 100)

2.4 Evaluation Criteria

The following metrics were used for evaluation (Sokolova, M., & Lapalme, G. (2009)):

- **Accuracy**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where;

TP (True Positive): The case is actually positive (e.g., malignant) and the model correctly predicted it as positive.

TN (True Negative): The case is actually negative (e.g., benign) and the model correctly predicted it as negative.

FP (False Positive): The case is actually negative, but the model incorrectly predicted it as positive (false alarm).

FN (False Negative): The case is actually positive, but the model incorrectly predicted it as negative (missed case).

- **Area Under the Curve (AUC)**

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad (2)$$

Where;

TPR: True Positive Rate (Sensitivity or Recall)

FPR: False Positive Rate=$\frac{FP}{FP+TN}$

- **Precision**

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- **Recall (Sensitivity)**

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

- **F1-Score**

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

3 Results

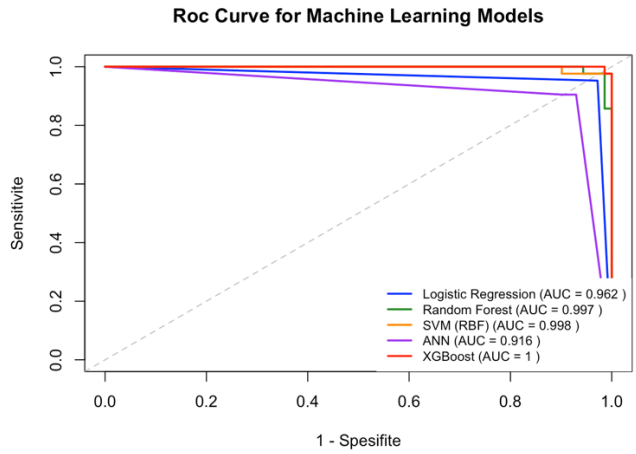
The results are summarized in the following table:

Tab.1: Evaluation Criteria of Models

Model	Accuracy	AUC	Precision (Specificity)	Recall (Sensitivity)	F1-Score
Logistic Regression	0.9646	0.9621	0.9524	0.9524	0.9524
Random Forest	0.9558	0.9970	0.9111	0.9762	0.9425
SVM (RBF)	0.9646	0.9977	0.9318	0.9762	0.9535
ANN	0.9204	0.9158	0.8837	0.9048	0.8941
XGBoost	**0.9912**	**0.9997**	**0.9767**	**1.0000**	**0.9882**

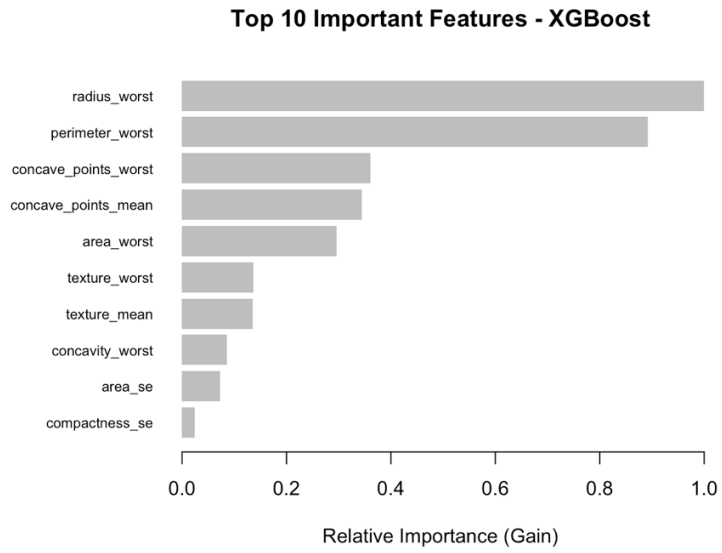
As shown in *Tab.1*, among the evaluated methods, the XGBoost algorithm achieved the highest accuracy on the WDBC dataset with 99%. In addition to the numerical results, ROC curves were plotted for all models, confirming the superior discriminative power of XGBoost *(Fig. 2)*. It was followed by Support Vector Machine (SVM) and Logistic Regression, which also demonstrated high classification accuracy.

Fig. 2: Roc Curve for Machine Learning Models (Based on authors' own calculations)



Feature importance analysis conducted to determine the most significant predictors in breast cancer classification revealed that, for the XGBoost model, "radius worst", "perimeter worst", and "concave points worst" were the most influential variables contributing to model performance *(Fig.3)*.

Fig.3: Top contributing features according to the XGBoost model (Based on authors' own calculations)



Conclusion

Accurate prediction of cancer is not limited to a diagnosis and prognosis process based solely on physical examination or biopsy. There is no single variable that determines cancer; instead, multiple variables and, more importantly, the relationships among them play a crucial role. These complex interactions are often overlooked in univariate statistical analyses. Machine learning techniques, however, allow for the consideration of multiple variables simultaneously, leading to more accurate results in distinguishing cancer cases.

As demonstrated in this study, the results emphasize the robustness and reliability of ensemble models such as XGBoost in differentiating between malignant and benign cases with high accuracy. For this reason, machine learning techniques have become increasingly popular not only in social and natural sciences but also in health sciences in recent years. In future studies, more advanced deep learning approaches and integration of diverse clinical data will be explored to further enhance diagnostic accuracy.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK).

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249–268.

Ojha, V. K., Abraham, A., & Snášel, V. (2017). Metaheuristic design of feedforward neural networks: A review of two decades of research. Engineering Applications of Artificial Intelligence, 60, 97–116.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386.

Sharma, A., Kulshrestha, S., & Daniel, S. B. (2018). Machine learning approaches for cancer detection. International Journal of Engineering and Manufacturing, 8(2), 45.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

Stattin, P., Carlsson, S., Holmström, B., Vickers, A., Hugosson, J., Lilja, H., & Jonsson, H. (2014). Prostate cancer mortality in areas with high and low prostate cancer incidence. Journal of the National Cancer Institute, 106(3), dju007.

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences of the United States of America, 87(23), 9193–9196.

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

Wu, X., Kumar, V., Quinlan, J. R., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37.

Contact

Ayşegül Yabacı Tak
Bezmialem Vakıf University, Faculty of Medicine, Department of Biostatistics,
Vatan Bulvarı No:113 34093 Fatih/İstanbul, Türkiye
ayabaci@bezmialem.edu.tr