Machine Learning and Feature Selection for Breast Cancer Prediction

Xinlei He

Shanghai Pinghe School, 201208, Shanghai, China
hexinlei@shphschool.com

Abstract. One of the most prevalent and fatal tumors that impact women globally is breast cancer. Traditional diagnostic methods, while effective, can be costly. The goal of this research is to improve the precision and effectiveness of breast cancer detection by combining feature selection techniques with machine learning models. Seven machine learning models were trained and assessed using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset in conjunction with three feature selection strategies: filter method, univariate selection (SelectKBest), and embedded method (Random Forest importance). Experimental results show that neural networks achieved the highest performance when using all features, while ensemble models performed best when used with filter feature selection. The study found that the choice of feature selection method should be aligned with the nature of the model, and combining suitable selection strategies with machine learning models can significantly enhance diagnostic performance. This approach can reduce misdiagnosis and improve early treatment outcomes.

Keywords: Machine Learning, Breast Cancer Diagnosis, Feature Selection, WDBC Dataset

1 Introduction

Breast cancer is a disease that is frequently encountered and has resulted in numerous fatalities among women worldwide. Over 2.3 million new cases and 670,000 fatalities were reported in 2022 [1]. Early diagnosis and prediction are essential to raising patient survival rates. Although mammograms and other traditional diagnostic methods are useful, they can also be costly and subject to human error [2]. The development of systems that are based on machine learning (ML) has been facilitated by these limitations, with the aim of improving the accuracy and efficiency of diagnosis [3]. The use of machine learning in breast cancer diagnosis has been prevalent over the past two decades. A wide range of ML algorithms have been applied to medical datasets with promising results [3–6]. The Wisconsin Breast Cancer Dataset (WBCD) is one of the benchmarks utilized in these studies. It comprises 30 numerical features that are derived from digitized images of FNA of breast masses.

Early work by Street et al. showed the effectiveness of linear models using features like cell radius and texture [7]. SVMs have demonstrated high accuracy (>95%) for

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_35

tumor classification, while Random Forest and Decision Trees are used for their interpretability and resistance to overfitting [4, 5].

However, as medical datasets grow in size and complexity, the role of feature selection has become increasingly important. According to Guyon and Elisseeff, eliminating redundant features enhances model performance and makes it easier to spot trends in the datasets [8]. Filter methods (like SelectKBest), wrapper methods (like Sequential Feature Selection), and embedding methods are the three general categories into which feature selection approaches are usually divided. Each has its strengths and weaknesses regarding computational cost, performance, and generalizability [9].

As datasets have grown, feature selection has become more important to remove redundant variables and improve both model performance and clarity [8]. Feature selection techniques include filter, wrapper, and embedded methods, each with their own advantages regarding efficiency and generalizability [9].

Despite these advancements, few studies have compared feature selection methods across multiple classifiers using the WBCD. This study aims to fill the gap by analyzing how different feature selection strategies impact the performance and simplicity of ML models for breast cancer diagnosis.

2 Methodology

2.1 Description of the Dataset

There are 569 cases of FNA from breast masses in the WDBC dataset [10]. The first property in each row is an identifying number; the second is a binary diagnostic (M for malignant and B for benign); the following 30 attributes are real-valued qualities derived from the properties of cell nuclei. Measures like radius, texture, area and perimeter are examples of these characteristics. Three values are given for each of these metrics: the mean, the standard error, and the worst value. Note that the WDBC dataset contains slightly unbalanced class (357 benign, 212 malignant), which may bias the models.

2.2 Dataset Preprocessing

Given the dataset's lack of missing attribute values, no data imputation is necessary. The column 'id' was discarded as it is unrelated to model training. The categorical "diagnosis" attribute was transformed into a binary numerical variable, where "M" was encoded as 1 (malignant) and "B" as 0 (benign) to simplify upcoming ML modeling. To ensure the validity of the experimental results, the dataset was randomly selected using a fixed random seed (42) to create an 80% training set and a 20% test set. This split allows for the evaluation of the model on general data, which is a critical step when validating the model's performance.

2.3. Machine Learning Models

A variety of ML algorithms was employed to evaluate classification performance across different modeling paradigms. Note that all random seeds used for ML models were set to 42 for consistent results.

In feature space, instances are grouped by the majority vote of their k-nearest neighbors using a non-parametric model known as K-Nearest Neighbors (KNN). KNN was chosen for its simplicity and its ability to capture local patterns in the data without making strong assumptions about their distribution. In this study, k was set to 5 to balance local and global similarity.

A sigmoid function is used in the linear model known as logistic regression to predict the target class's probability. It is suitable for binary classification and provides interpretable coefficients. Logistic Regression was chosen for its ability to serve as a strong baseline, and its simplicity helps mitigate overfitting in smaller datasets. The maximum training iteration was set to 10000 for quick model training.

A radial basis function kernel is used by the discriminative Support Vector Machine (SVM) model to map input into a high-dimensional space. SVMs typically achieve strong performance thanks to their flexibility in handling complex decision boundaries. The maximum training iteration was set to 500 for efficient model training.

The Classification and Regression Tree method serves as the foundation for the Decision Tree tree structure model, which constructs a binary tree by recursively separating the feature space to complete the classification of data. Decision trees are useful because they may represent intricate relationships, highlight significant characteristics, and keep the decision-making process transparent.

Random Forest, which ensembles a group of decision trees, can improve performance and resilience by lowering variance through bagging and random feature selection. Random Forests are favored for their strong empirical performance, robustness to noisy data, and interpretable feature importance scores. To balance performance and training duration, 100 estimators were used.

Gradient Boosting is also an ensemble method. It is based on decision trees and focuses on misclassified instances to improve prediction accuracy gradually. This model was selected for its high predictive accuracy, especially in handling tabular datasets with complex relationships and class imbalance. The XGBoost implementation was used for its efficiency and performance.

A feedforward neural network (NN) with one hidden layer, which has 16 nodes, and a rectified linear unit (ReLU) activation function is implemented using PyTorch. It was chosen for its ability to capture intricate, complex feature interactions that traditional models may miss. The model was trained with the Adam optimizer and binary cross-entropy loss for 100 epochs, allowing it to learn characteristics of the features.

2.4 Performance Metrics

Using a range of assessment measures, the model's performance was carefully examined. In medical diagnosis, precision measures the model's ability to minimize false positives, while accuracy shows the overall proportion of correct predictions. Recall

assesses the model's ability to detect genuine positive instances, which lowers the number of missed diagnoses. The F1 score, which takes the harmonic mean of accuracy and recall, is particularly useful when dealing with imbalanced data. Additionally, using the area under the receiver operating characteristic curve (AUC), the model's ability to distinguish between classes across a variety of classification criteria was assessed.

2.5 Feature Selection Methods

Filter method. To reduce complexity and improve model efficiency, a manual filter approach was employed. The filter method is chosen for its efficiency and its wide use over biomedical ML tasks [9].

First, a variance threshold was applied to remove features with low discriminative ability (variance < 0.01). This is because features with less variation are unlikely to contribute significantly to classification. Second, features with high pairwise correlations (Pearson's r > 0.6) were discarded to reduce redundant features.

This combined approach ensured that the remaining features were both informative and clean, which enhances the interpretability and performance of subsequent models.

Univariate feature selection. The SelectKBest method, with k=5, was used to identify the top 5 features. Using the Analysis of Variance (ANOVA) F-value as the scoring function, this approach evaluates each feature's statistical significance in distinguishing between benign and malignant classes without considering feature interactions [8]. This approach produced a number of features that are better suited for training machine learning models.

Embedded feature selection. To assess feature importances, the entire feature set was used to train a Random Forest classifier. This embedded method captures the relative importance of features in an ensemble model with a tree structure. This can account for interactions between features. The top 5 features were selected. In contrast to the filter technique and univariate selection, this methodology offers a supplementary strategy that may improve the model's ability to identify intricate patterns in the data.

3 Results

3.1 Feature Selection Outcomes

The filter feature selection approach reduced the 30 initial features to 18 attributes. The Univariate Selection (SelectKBest) method was implemented to identify the top 5 features. All of them showed significant univariate associations with the diagnosis (ANOVA F-values > 600). The top 5 features determined by the Random Forest model have importances ranging from 0.6 to 0.15.

3.2 Model Performance with all Features

After training and evaluating the models using all features (i.e., without feature selection), the performance metrics are shown in Fig. 1.

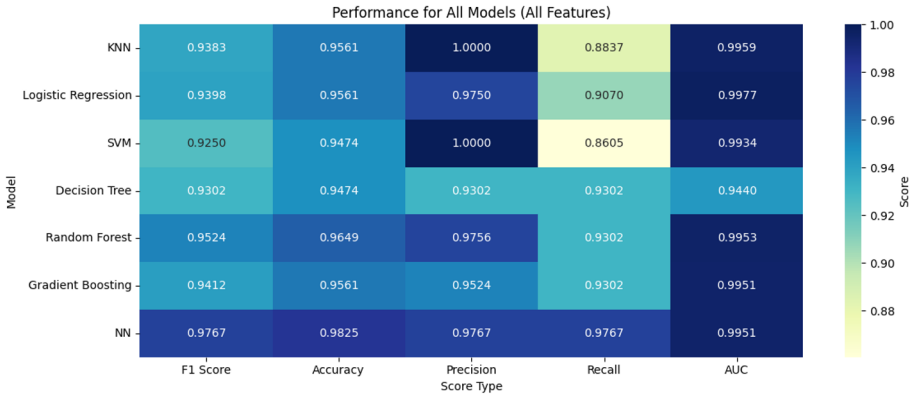

Fig. 1. Performance for all models (all features) (Photo/Picture credit: Original).

When evaluated on the full 30-feature set, the neural network demonstrated the highest performance with the highest F1 score, accuracy, and recall. Gradient Boosting and Random Forest both had respectable results across the board. However, KNN, Logistic Regression, and SVM all have relatively low recall, scores meaning that they reported many false positives on the testing dataset.

3.3 Impact of Feature Selection Methods

The percentage change in performance measures is calculated after feature selection techniques have been applied to every model. As F1 score is the most important metrics that the study focus on, due to its ability to represent a model's ability to balance recall and precision, only the percentage change in F1 scores is shown Fig. 2 illustrates the percentage change in F1 scores for various models after feature selection methods were applied. Fig. 2, the x-axis is the different models implemented while the y-axis represents the different feature selection methods, including filter methods, SelectKBest, and embedded methods. Each number, which represent the percentage change in F1 score, is color coded for readability (orange means improvement while blue indicates performance drop).

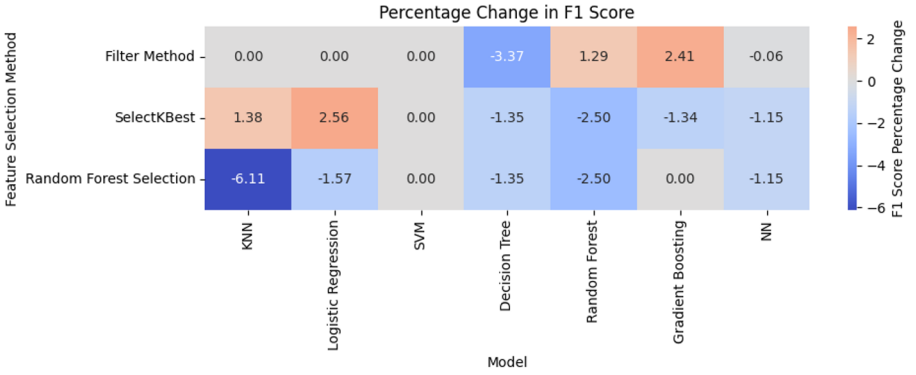

Fig. 2. Percentage change in F1 score (Photo/Picture credit: Original).

For the filter method, most models maintained constant F1 scores.There were no significant changes in the F1 scores of models like KNN, SVM, and Logistic Regression, while Random Forest and Gradient Boosting exhibited slight improvements, with their F1 scores increasing by approximately 1% and 2%, respectively.

In contrast, the SelectKBest method produced mixed outcomes. Models relying on simpler mechanisms, such as KNN and Logistic Regression, experienced minor gains in F1 scores of roughly 1% to 3%, likely due to better alignment with the statistically significant features selected. However, more complex models suffered performance declines ranging from 1% to 3%.

The random forest selection method exhibited a negative outcome. KNN, Logistic Regression, and Random Forest recorded considerable performance drops, with F1 scores decreasing by as much as 6%. Meanwhile, SVM, Decision Tree, Gradient Boosting, and the neural network showed either neutral changes or slight negative impacts. This trend probably means that the embedded feature selection method is less compatible with linear models and models that are based on distance.

4 Discussion

The neural network performed the best overall across all measures when all characteristics were included, which is consistent with other research showing how well deep learning captures high-dimensional and non-linear patterns in medical data [6]. The network's capacity to recognize intricate patterns among features is responsible for this outcome.

The implementation of feature selection methods showed varying effects depending on the model architecture.

The filter method had a particularly positive impact on ensemble models. This supports findings by Chandrashekar and Sahin, who noted that filtering irrelevant features benefits model performance when computational budget is limited, especially in models that aggregate multiple weak learners [9].

The SelectKBest univariate approach benefited simpler models like Logistic Regression and KNN. These models typically do not capture feature interactions well, so the selection of statistically significant features aligns with their mechanisms [3]. However, this approach caused performance drops in ensemble and neural models, possibly due to the loss of important relationships among features.

The Random Forest-based embedded selection did not improve performance despite its theoretical advantage of identifying important features. One reason could be that using the most important features derived from it alone may exclude feature combinations important to other classifiers. This is consistent with research that warns against blindly transferring feature importance across different model types [9].

Another observation is that feature selection did not benefit all models equally. The lack of improvement in neural network performance after feature selection suggests that neural networks might inherently learn to ignore irrelevant features through their internal mechanisms. This means that feature reduction is less needed for neural networks.

5 Conclusion

The WDBC dataset was used in this work to evaluate how three feature selection methods affected seven machine learning models for breast cancer diagnosis. Results showed that the filter method maintained stable performance for simpler models and slightly enhanced ensemble models like Gradient Boosting and Random Forest. SelectKBest delivered mixed outcomes, marginally benefitting simpler models while reducing the effectiveness of models that rely on feature interactions. The embedded method, based on Random Forest feature importance, performed poorly for all experimented models, likely due to the exclusion of critical feature combinations.

Neural networks achieved the best overall performance when trained on all features, making them suitable for clinical deployment, while Gradient Boosting paired with the filter method offered a strong balance of accuracy and interpretability.

Future studies should address class imbalance, refine neural network architectures, and explore deep learning techniques for direct image analysis to further enhance diagnostic tools. This work emphasizes how crucial it is to match feature selection tactics with model attributes in order to maximize machine learning performance in the detection of breast cancer.

References

1. World Health Organization: Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, last accessed 2025/08/05
2. U.S. Preventive Services Task Force: Screening for breast cancer: Recommendation statement. Journal of the American Medical Association 331(22), 1896–1907 (2024)
3. Mao, L., Wang, H., Hu, L.S., Tran, N.L., Canoll, P.D., Swanson, K.R., Li, J.: Knowledge informed machine learning for cancer diagnosis and prognosis: A review. (2024, in review or preprint – please confirm publication info)

4. El Kenawy, E.-S.M., et al.: HHO SVM: A hybrid support vector machine with Harris Hawks Optimization for breast cancer diagnosis. Mathematics 11(14), 3251 (2023)
5. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine 34(2), 113–127 (2005)
6. Lee, H., Kim, J., Park, E., Kim, M., Kooi, T.: Enhancing breast cancer risk prediction by incorporating prior mammographic images. (2023, please confirm source – journal/conference info missing)
7. Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology, vol. 1905, pp. 861–870. SPIE, Bellingham (1993)
8. Buyukkececi, M., Okur, M.C.: A comprehensive review of feature selection and feature selection stability in machine learning. Journal of Science 36(4), 1506–1520 (2023)
9. Hopf, K., Reifenrath, S.: Filter methods for feature selection in supervised machine learning applications – review and benchmark. (2021, please provide journal or conference name)
10. Dua, D., Graff, C.: Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic, last accessed 2025/08/05