Machine Learning Approaches to Tumor Diagnostics: A Comparative Study of Classification Models

Author: Åsa Hellstrand

Date: April 26, 2025

Abstract

This study investigates the application of machine learning methods for tumor diagnostics, with the specific aim of identifying the most suitable classification algorithm for distinguishing between benign and malignant tumors. Using the Breast Cancer Wisconsin dataset from the UCI Machine Learning Repository, several classification models were trained and evaluated, including *Logistic Regression*, *K-Nearest Neighbors*, *Support Vector Machine*, and *Random Forest*. The models were compared on *accuracy*, *precision*, *recall*, *F1-score*, and the *area under the ROC curve (AUC)*. Results indicate that Random Forest achieved the strongest overall performance, though Support Vector Machine also demonstrated competitive results under alternative train-test splits. These findings highlight the value of ensemble methods while underlining the influence of data partitioning in small, easily separable datasets.

Introduction

Early and accurate diagnosis of tumors is critical in clinical decision-making, as it directly affects treatment outcomes. Machine learning (ML) has emerged as a valuable tool in medical diagnostics, enabling the automated analysis of complex clinical datasets. The present study focuses on comparing four widely used classification algorithms to determine which is most effective in predicting tumor malignancy based on cytological features with a particular dataset.

Materials and Methods

Dataset

The analysis was based on the Breast Cancer Wisconsin dataset (UCI Machine Learning Repository). The dataset contains 683 records with 10 predictive attributes describing cellular characteristics such as clump thickness, cell size uniformity, bare nuclei, and mitoses. The dependent variable is the tumor classification, encoded as 2 for benign and 4 for malignant.

The attribute 'Sample code number' was excluded from analysis, as it functions solely as an identifier and carries no diagnostic relevance.

The dataset contains no missing values.

Models Evaluated

Four classification algorithms were tested: *Logistic Regression*, *K-Nearest Neighbors*, *Support Vector Machine*, and *Random Forest*. These models were selected for their established role in medical ML applications and their complementary methodological strengths.

Evaluation Procedure

The models were implemented in Python using Jupyter Notebook. Training and evaluation were conducted on stratified train-test splits of the dataset. Performance was assessed using *accuracy*, *precision*, *recall*, *F1-score*, and *AUC*. Particular emphasis was placed on recall, since minimizing false negatives is crucial in tumor diagnostics.

Results

The comparative performance of the models is presented in Table 1.

Table 1. Performance metrics of classification models

Method	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.956	0.940	0.940	0.940	0.99
K-Nearest Neighbors	0.956	0.923	0.960	0.941	0.99
Support Vector Machine	0.956	0.907	0.980	0.942	0.99
Random Forest	0.971	0.942	0.980	0.961	1.00

Random Forest outperformed the other models across most metrics, particularly in balancing precision and recall. While Support Vector Machine achieved comparable recall, Random Forest demonstrated a slightly stronger F1-score. The AUC of 1.00 for Random Forest indicates near-perfect separability, though this may also suggest overfitting. Inspection of the ROC curve, however, confirmed a robust generalization ability.

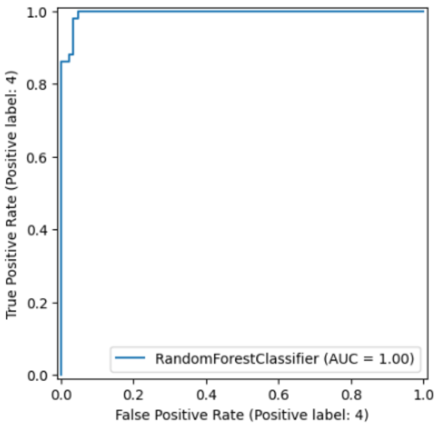

Figure 1. ROC curve for Random Forest model

Interestingly, when varying the random seed used for data partitioning, Support Vector Machine occasionally emerged as the best-performing model. This variability likely reflects the dataset's relatively straightforward separability, where small changes in train-test splits can influence model rankings.

Discussion

In tumor diagnostics, recall is of paramount importance, as missing a malignant case can have severe clinical consequences. Precision also matters, particularly in contexts where minimizing false positives reduces unnecessary interventions. The F1-score provides a balanced view, while the AUC offers a threshold-independent measure of model discrimination.

The findings suggest that ensemble methods such as Random Forest are highly suitable for this type of dataset. Nonetheless, Support Vector Machine remains a strong candidate, and the sensitivity of results to data partitioning underscores the importance of robust validation strategies such as cross-validation.

Conclusion

Random Forest emerged as the most reliable classifier in this initial study, combining high accuracy, strong recall, and robust AUC performance. However, the comparable results obtained with Support Vector Machine demonstrate that multiple models can achieve high diagnostic accuracy in easily separable datasets. Further research with larger, more complex medical datasets is necessary to validate these conclusions and support their translation into clinical practice.