Breast Cancer Detection Using Machine Learning: A Comparative Study

DANISH SHEIKH

Introduction

The purpose of this project is to develop a machine learning model capable of accurately classifying breast tumors as either benign or malignant. This classification task leverages the **Breast Cancer Wisconsin Dataset**, a widely recognized resource in medical diagnostics. The dataset contains **569 instances**, each characterized by **30 features** derived from digitized images of fine needle aspirates (FNA) of breast masses. These features include measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The primary objectives of this project are:

- To implement four distinct classification algorithms: **Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM).**
- To evaluate the performance of these models using a range of metrics.
- To determine the most effective model for this medical diagnostic task, with a focus on its applicability in real-world scenarios.

A key emphasis is placed on achieving high recall, as correctly identifying malignant cases (minimizing false negatives) is critical in medical diagnostics.

Methodology

Algorithms

Four classification algorithms were selected for this study, each chosen for its unique strengths in binary classification tasks:

- **Logistic Regression:** A linear model serving as a baseline due to its simplicity and interpretability. It performs well when the relationship between features and the target variable is approximately linear.
- **Decision Tree:** A non-linear model that splits the data based on feature thresholds to capture complex relationships. While intuitive, it is susceptible to overfitting.

- **Random Forest:** An ensemble method that aggregates multiple decision trees to enhance robustness and reduce overfitting. It excels in high-dimensional datasets and captures feature interactions effectively.
- **Support Vector Machine (SVM):** A powerful algorithm that identifies the optimal hyperplane to separate classes. It is particularly suited to high-dimensional spaces and can adapt to non-linear data with appropriate kernel functions.

Preprocessing Steps

The dataset required minimal preprocessing due to its completeness, but the following steps were applied to ensure optimal model performance:

- **Feature Scaling**: Features were standardized using the `StandardScaler` to normalize their scales. This step is crucial for algorithms like Logistic Regression and SVM, which are sensitive to feature magnitudes.
- **Data Splitting**: The dataset was divided into training and testing sets using an **80-20 split**, with stratification to preserve the class distribution (benign vs. malignant) in both subsets.

Hyperparameter Tuning and Cross-Validation

For this initial analysis, **default hyperparameters** were used across all models to maintain simplicity and ensure a fair comparison. No hyperparameter tuning or cross-validation was performed. However, these techniques could be explored in future iterations to potentially enhance model performance.

Results

The performance of each model was assessed using five key metrics: **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**. These metrics provide a comprehensive view of model effectiveness, particularly in the context of medical diagnostics where recall is paramount. The results are presented in the table below:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.965	0.975	0.929	0.951	0.996
Decision Tree	0.930	0.905	0.905	0.905	0.925
Random Forest	0.974	1.000	0.929	0.963	0.993
Support Vector Machine (SVM)	0.974	1.000	0.929	0.963	0.995

Metric Definitions

- **Accuracy**: The proportion of correctly classified instances out of the total.
- **Precision**: The proportion of positive (malignant) predictions that are correct.
- **Recall**: The proportion of actual malignant cases correctly identified (critical in this context).
- **F1-Score**: The harmonic mean of precision and recall, balancing the two metrics.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.

Discussion

The evaluation of four classification models reveals **Random Forest** and **Support Vector Machine (SVM)** as the top performers, each achieving an accuracy of 0.974, perfect precision (1.000), and identical F1-scores of 0.963. Both models recorded a recall of 0.929, ensuring robust detection of malignant cases. SVM's slightly higher ROC-AUC (0.995) compared to Random Forest (0.993) suggests a marginal advantage in distinguishing between classes across various thresholds. The perfect precision of both models eliminates false positives, a critical attribute in medical diagnostics to prevent unnecessary patient anxiety or interventions..

Logistic Regression performs admirably, with an accuracy of 0.965, precision of 0.975, and recall of 0.929, matching the recall of Random Forest and SVM. Its ROC-AUC of 0.996 is competitive, and its interpretability is a significant advantage in clinical settings, where understanding the basis of predictions enhances trust and applicability. However, its slightly lower accuracy and precision place it just behind the leading models.

Decision Tree demonstrates the weakest performance, with an accuracy of 0.930, precision, recall, and F1-score all at 0.905, and a ROC-AUC of 0.925. These metrics indicate challenges in generalizing effectively, likely due to overfitting or insufficient capture of the dataset's complexity compared to ensemble methods like Random Forest.

In breast cancer detection, where maximizing recall is paramount to minimize false negatives, **Random Forest** and **SVM** offer an optimal balance of high recall, perfect precision, and superior accuracy. Random Forest's ensemble approach enhances its robustness, while SVM excels in high-dimensional spaces, making either model suitable for this critical task, with Random Forest potentially preferred for its implementation simplicity.

Conclusion

This project effectively implemented and evaluated four machine learning models for breast cancer detection using the Breast Cancer Wisconsin Dataset. **Random Forest** and **Support Vector Machine (SVM)** emerged as the most effective, each achieving an accuracy of 0.974, precision of 1.000, recall of 0.929, F1-score of 0.963, and near-perfect ROC-AUC scores (0.993

and 0.995, respectively). These models demonstrate exceptional performance, making them well-suited for this vital diagnostic application.

Key findings include:

- Random Forest and SVM excel in precision and recall, ensuring no false positives and reliable malignant case detection.
- Logistic Regression provides strong interpretability and competitive performance, with a recall of 0.929.
- Recall is a critical metric, and Random Forest and SVM achieve a robust balance of all performance indicators.

Suggestions for Improvement

To enhance this work, the following strategies are recommended:

- **Hyperparameter Optimization**: Tuning parameters, such as Random Forest's number of estimators or SVM's kernel settings, using grid search could further improve recall.
- **Feature Engineering**: Applying Principal Component Analysis or feature selection may streamline the model by reducing dimensionality while preserving predictive power.
- **Advanced Algorithms**: Investigating models like Gradient Boosting or Neural Networks could yield additional performance improvements.
- **Cross-Validation**: Employing k-fold cross-validation would provide a more robust estimate of model performance, minimizing potential bias from a single train-test split.

This project highlights the transformative role of machine learning in medical diagnostics, emphasizing the importance of selecting models that align with clinical priorities, such as maximizing recall while maintaining high precision and accuracy.
