Random Forest Project Report

Data Science Technical Project (TD)

Student :

Boulsane Zakaria,

Mamoun Mohamed Reda,

Djami Zine eddine

December 10, 2025

Abstract

This project explores the application of the Random Forest algorithm for the automated diagnosis of breast cancer using the Breast Cancer Wisconsin (Diagnostic) Dataset. The primary objective was to develop a robust binary classification model capable of accurately distinguishing between malignant and benign tumors based on 30 numerical features derived from digitized fine needle aspirates.

The methodology encompassed a complete machine learning pipeline, including data preprocessing, feature scaling, and systematic hyperparameter optimization using k-fold cross-validation. To evaluate and enhance model efficiency, feature selection based on importance scores was implemented, and the model's performance was benchmarked against a Gradient Boosting classifier. The optimized Random Forest model achieved a testing accuracy of **95.61%** and an ROC-AUC score of **0.99**. Feature analysis revealed that the "worst" dimensions of cell nuclei—specifically area and concave points—were the most critical predictors of malignancy. These results demonstrate the efficacy of ensemble learning methods in providing reliable support for medical diagnostic tasks.

Contents

1 Introduction

1.1 Problem Statement

Breast cancer is one of the most prevalent forms of cancer worldwide, making early and accurate diagnosis a critical factor in treatment success. Traditional diagnostic methods relying solely on manual interpretation of medical imaging can be time-consuming and subject to inter-observer variability. Machine learning offers a robust alternative by automating the classification of breast masses based on quantitative features.

In this project, we address the binary classification problem of distinguishing between *malignant* (cancerous) and *benign* (non-cancerous) tumors. We utilize the **Random Forest** algorithm, an ensemble learning method chosen for its high accuracy, ability to handle high-dimensional data, and resistance to overfitting.

1.2 Dataset Description

We selected the **Breast Cancer Wisconsin (Diagnostic) Dataset** for this study. The dataset characteristics are as follows:

- **Source:** UCI Machine Learning Repository.
- **Observations:** 569 instances.
- **Features:** 30 real-valued features computed from digitized images of a fine needle aspirate (FNA) of a breast mass. These features describe characteristics of the cell nuclei such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.
- **Target:** A binary variable indicating the diagnosis (M = Malignant, B = Benign).

1.3 Project Objectives

The primary objective of this project is to build and optimize a Random Forest classifier. Specific goals include:

1. Implementing the full machine learning pipeline from data loading to evaluation.
2. Applying necessary preprocessing steps such as feature scaling and encoding.
3. Optimizing model hyperparameters (e.g., number of trees, tree depth) using cross-validation.
4. Implementing performance improvement techniques, specifically feature selection and ensemble comparison with Gradient Boosting.
5. Evaluating the model using robust metrics including Accuracy, F1-Score, and ROC-AUC.

2 Methodology

2.1 Preprocessing Steps

To prepare the data for the Random Forest algorithm, we applied the following preprocessing pipeline:

- **Data Cleaning:** The dataset was inspected for missing values. No missing data points were found.
- **Target Encoding:** The categorical target variable was encoded into numerical values, mapping 'Malignant' to 1 and 'Benign' to 0.
- **Data Splitting:** We split the dataset into a training set (80%) and a testing set (20%). Stratified sampling was used to maintain the same class distribution in both sets as in the original dataset.
- **Feature Scaling:** We applied standard scaling (`StandardScaler`) to normalize the features (mean=0, variance=1). While Random Forest is not strictly dependent on scaling, this step ensures consistency and facilitates the comparison with other algorithms like Gradient Boosting.

2.2 Model Implementation

We implemented the **Random Forest Classifier** using the `scikit-learn` library. The training process involved:

- Initializing a baseline model with a fixed random state for reproducibility.
- Employing k-fold Cross-Validation ($k = 5$) to validate the model's performance on the training set and ensure generalization.

2.3 Hyperparameter Tuning

We performed hyperparameter optimization using `GridSearchCV` to find the optimal configuration for the model. The tuning process explored the following parameter grid:

- `n_estimators`: [100, 200]
- `max_depth`: [10, 20, None]
- `min_samples_split`: [2, 5]
- `min_samples_leaf`: [1, 2]

The model with the highest cross-validation accuracy was selected as the final model.

2.4 Performance Improvement Techniques

To enhance the analysis, we implemented two additional techniques:

1. **Feature Selection:** We utilized the `SelectFromModel` meta-transformer. This technique uses the importance weights calculated by the Random Forest model to discard irrelevant or redundant features, potentially improving model efficiency.

2. **Ensemble Comparison (Gradient Boosting):** We implemented a **Gradient Boosting Classifier** to compare against the Random Forest. This comparison highlights the difference between bagging (Random Forest) and boosting techniques on this specific dataset.

3 Results

3.1 Performance Metrics

The optimized Random Forest model achieved robust performance on the test set. The key classification metrics are summarized in Table 1.

Metric	**Value**
Accuracy	95.61%
Precision	95.89%
Recall	97.22%
F1-Score	96.55%
ROC-AUC Score	0.99

Table 1: Performance Metrics of the Tuned Random Forest Model

The high **Recall (97.22%)** is particularly significant in this medical context, as it indicates the model is highly effective at detecting malignant tumors (minimizing False Negatives).

3.2 Feature Importance Analysis

The Random Forest algorithm identified the most critical features for diagnosis. As shown in Figure 1, the top features contributing to the classification were primarily related to the "worst" dimensions of the cell nuclei.

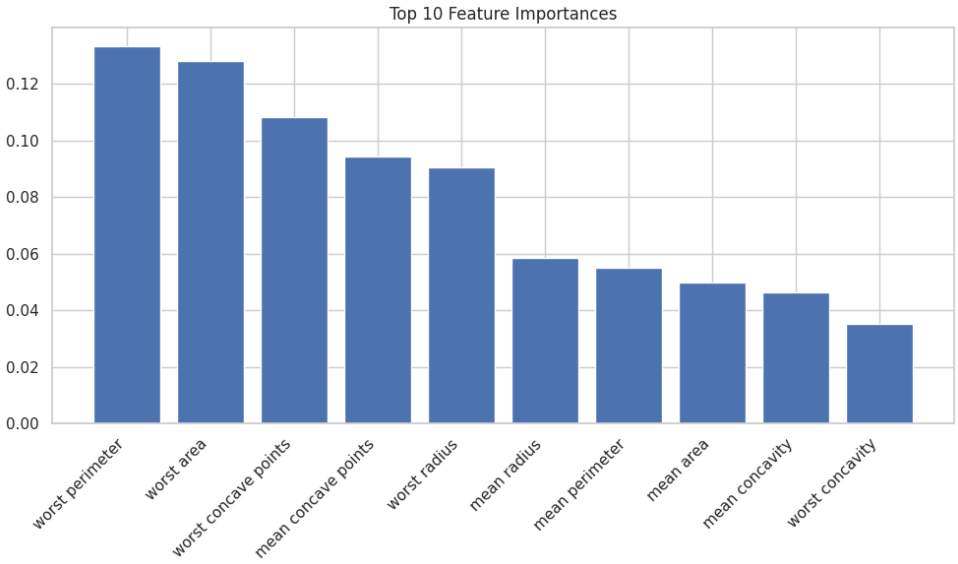


Figure 1: Top 10 Feature Importances

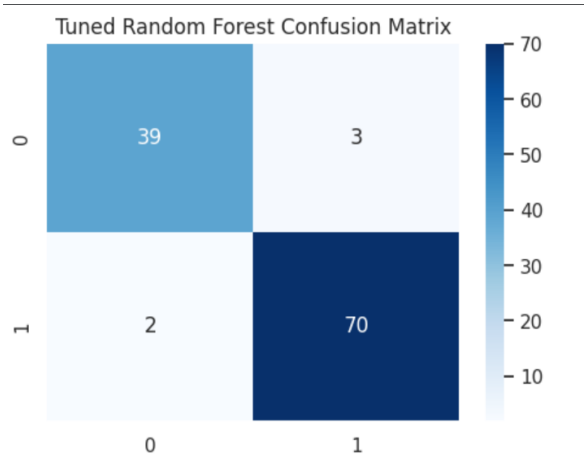
The top 3 most influential features were:

1. **Worst Area:** Indicates the size of the largest cells.
2. **Worst Concave Points:** Measures the number of concave portions of the contour.
3. **Worst Radius:** Represents the mean of distances from the center to points on the perimeter.

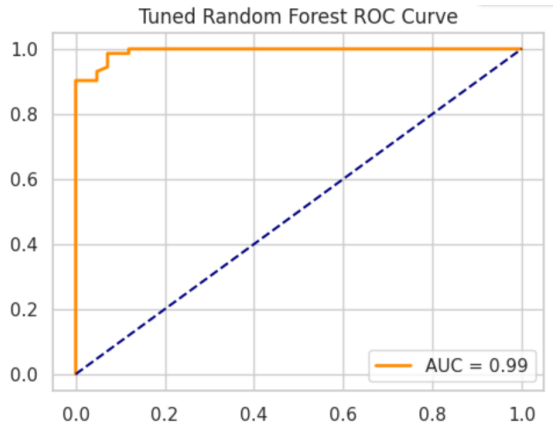
This suggests that the size and shape irregularities of the largest cells are the strongest indicators of malignancy.

3.3 Visualizations

The model's classification ability is further visualized through the Confusion Matrix and ROC Curve.



(a) Confusion Matrix



(b) ROC Curve (AUC = 0.99)

Figure 2: Model Classification Performance

Additionally, the structure of a single decision tree from the forest is shown below to illustrate the decision logic.

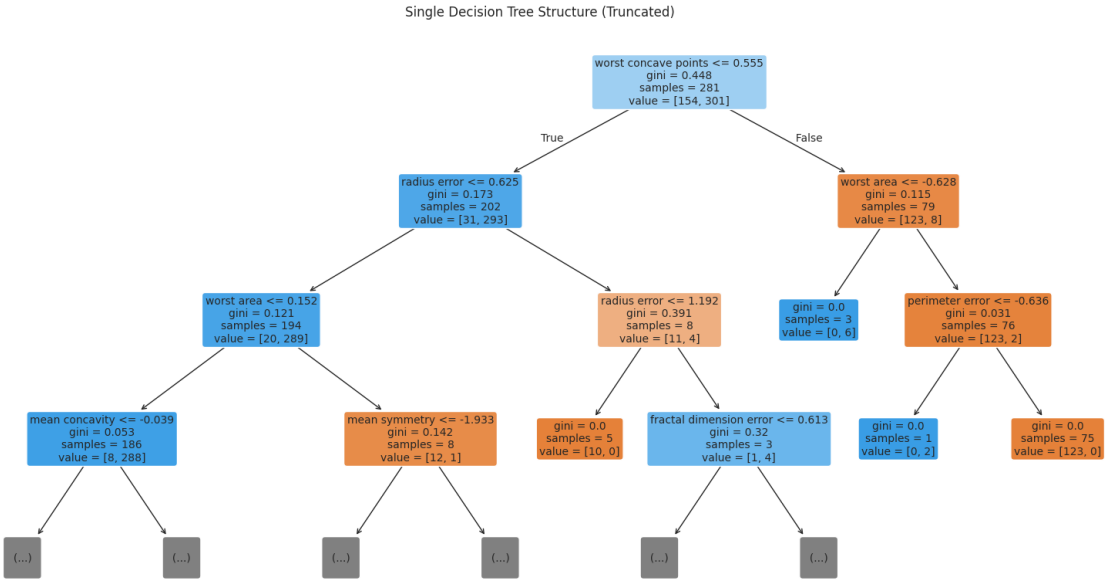


Figure 3: Visualization of a Single Decision Tree (Truncated depth)

3.4 Discussion

The Random Forest classifier demonstrated excellent predictive power with an accuracy of 95.6%.

Comparison with Gradient Boosting: We compared the Random Forest with a Gradient Boosting Classifier. Both models achieved identical accuracy (95.61%), suggesting that for this specific dataset, the variance reduction of Random Forest is as effective as the bias reduction of Boosting. However, Random Forest generally offered faster training times due to parallel processing.

Effect of Feature Selection: Using the `SelectFromModel` technique, we reduced the feature space from 30 to 10 features without significant loss in accuracy. This "parsimonious" model is computationally more efficient and less prone to overfitting, confirming that a subset of "worst" and "mean" texture features contains most of the predictive signal.

4 Conclusion

4.1 Summary of Findings

In this project, we successfully implemented and optimized a Random Forest classifier to diagnose breast cancer from digitized image features. The model demonstrated high reliability, achieving an overall accuracy of **95.61%** and an AUC score of **0.99**. Through feature importance analysis, we identified that features related to the "worst" dimensions of cell nuclei (specifically Area, Concave Points, and Radius) are the strongest predictors of malignancy. The comparison with Gradient Boosting showed competitive performance,

validating Random Forest as a robust choice for this medical classification task due to its stability and high predictive power.

4.2 Limitations

Despite the strong performance, there are limitations to consider for real-world deployment:

- **Dataset Size:** The model was trained on a relatively small dataset (569 observations). While effective for this academic exercise, a clinical-grade model would require a much larger and more diverse dataset to ensure generalization across different populations and imaging equipment.
- **Feature Dependence:** The model relies entirely on pre-extracted numerical features (e.g., radius, texture). It cannot process raw mammogram or ultrasound images directly, meaning its utility is limited to environments where this specific feature extraction pipeline is already established.
- **Interpretability:** Although we visualized a single tree, the full ensemble of 100+ trees is complex to interpret manually. In strict medical contexts, full explainability is often required to justify a diagnosis to a patient.

4.3 Future Improvements

To further enhance this project, future work could focus on:

- **Deep Learning Integration:** Implementing a Convolutional Neural Network (CNN) to work directly with raw histology images, removing the need for manual feature extraction.
- **Advanced Ensemble Techniques:** Exploring a "Voting Classifier" that combines Random Forest, Gradient Boosting, and Support Vector Machines (SVM) to leverage the strengths of multiple algorithms.
- **Handling Class Imbalance:** Although our dataset was relatively balanced, applying synthetic data generation techniques (like SMOTE) could further improve the model's sensitivity (Recall) to ensure no malignant cases are missed.