

Breast Cancer Classification and Prediction System: Technical Report

May 2025

Abstract

This report summarizes the development and performance of machine learning models for breast cancer classification and a web-based prediction system using the scikit-learn breast cancer dataset. It covers classification results, the Random Forest-based prediction system, and key observations.

1 Classification Observations

1.1 Wine Dataset

- Logistic Regression: Highest accuracy (~ 0.9815) due to linear separability of features.
- SVM: Lower accuracy (~ 0.7593) with default parameters; requires kernel tuning.
- Decision Tree: Good performance with `random_state=42` (~ 0.9630); variable without (~ 0.9444).

1.2 Breast Cancer Dataset

- Random Forest: Best performer (~ 0.9591), leveraging ensemble learning.
- Logistic Regression: Strong accuracy (~ 0.9474), suited to dataset's structure.
- SVM: Moderate accuracy (~ 0.9123); needs optimization.
- Decision Tree: Stable with `random_state = 42` (~ 0.9181); *variable otherwise* (~ 0.9064).

Table 1: Classification Performance

Model	Wine Accuracy	Breast Cancer Accuracy
Logistic Regression	0.9815	0.9474
SVM	0.7593	0.9123
Decision Tree (fixed)	0.9630	0.9181
Decision Tree (variable)	0.9444	0.9064
Random Forest	—	0.9591

2 Prediction System Observations

2.1 System Overview

- Model: Random Forest Classifier (~0.9591 accuracy), integrated into Flask web app.
- Input: Top 5 features selected via feature importance (e.g., worst perimeter, mean concave points).
- Output: Predicts Malignant (0) or Benign (1) with confidence score.

2.2 Design and Functionality

- Frontend: Responsive form in home page, styled with Tailwind CSS, validated via JavaScript.
- Backend: Flask processes inputs, scales features, and serves predictions.
- Feature Importance: Reduced input from 30 to 5 features, balancing simplicity and accuracy.

2.3 Performance and Limitations

- Performance: Fast, reliable predictions; highly user-friendly with minimal inputs.
- Limitations: Requires accurate feature inputs; no guidance on feature ranges.
- Future Work: Add feature range validation, visualize feature importance.