Breast Cancer Diagnosis Analysis

Predict breast cancer diagnosis (malignant vs benign) from digitized cell nuclei images using machin

Jotty SwarmMLComprehensive

February 05, 2026

Contents

Contents

Executive Summary

Predict breast cancer diagnosis (malignant vs benign) from digitized cell nuclei images using machine learning.

Key Results

Best Model: Logistic Regression

Performance Metrics:

Metric	Value
Accuracy	0.9825
Precision	0.9861
Recall	0.9861
F1	0.9861
Auc Roc	0.9954

Dataset: 30 features analyzed

Data Profile

Dataset Overview

- **Total Samples:** 569
- **Total Features:** 30

Data Types

Data Type	Count
float64	30

EDA Recommendations

- All features are numeric - no encoding needed
- Features describe cell nuclei measurements
- Consider feature scaling for distance-based models

Feature Importance Analysis

Feature importance measures how much each feature contributes to the model's predictions. Higher values indicate more influential features.

Top 20 Features

Rank	Feature	Importance
1	worst area	0.1400
2	worst concave points	0.1295
3	worst radius	0.0977
4	mean concave points	0.0909
5	worst perimeter	0.0722
6	mean perimeter	0.0696
7	mean radius	0.0687
8	mean concavity	0.0576
9	mean area	0.0492
10	worst concavity	0.0343
11	area error	0.0331
12	worst compactness	0.0186
13	worst texture	0.0186
14	radius error	0.0168
15	worst smoothness	0.0124
16	mean compactness	0.0117
17	perimeter error	0.0096
18	mean texture	0.0096
19	worst symmetry	0.0083
20	compactness error	0.0060

Feature Importance Visualization

Model Benchmarking

Multiple machine learning algorithms were evaluated using 5-fold cross-validation. The table below shows the performance of each model.

Model Comparison

Model	CV Score	Std Dev	Test Score	Time (s)
Logistic Regression	0.9802	±0.0128	0.9825	4.99
SVM	0.9714	±0.0179	0.9825	0.11
Random Forest	0.9538	±0.0235	0.9561	1.37
Gradient Boosting	0.9560	±0.0139	0.9561	2.73

Performance Visualization

Classification Performance

Classification Report

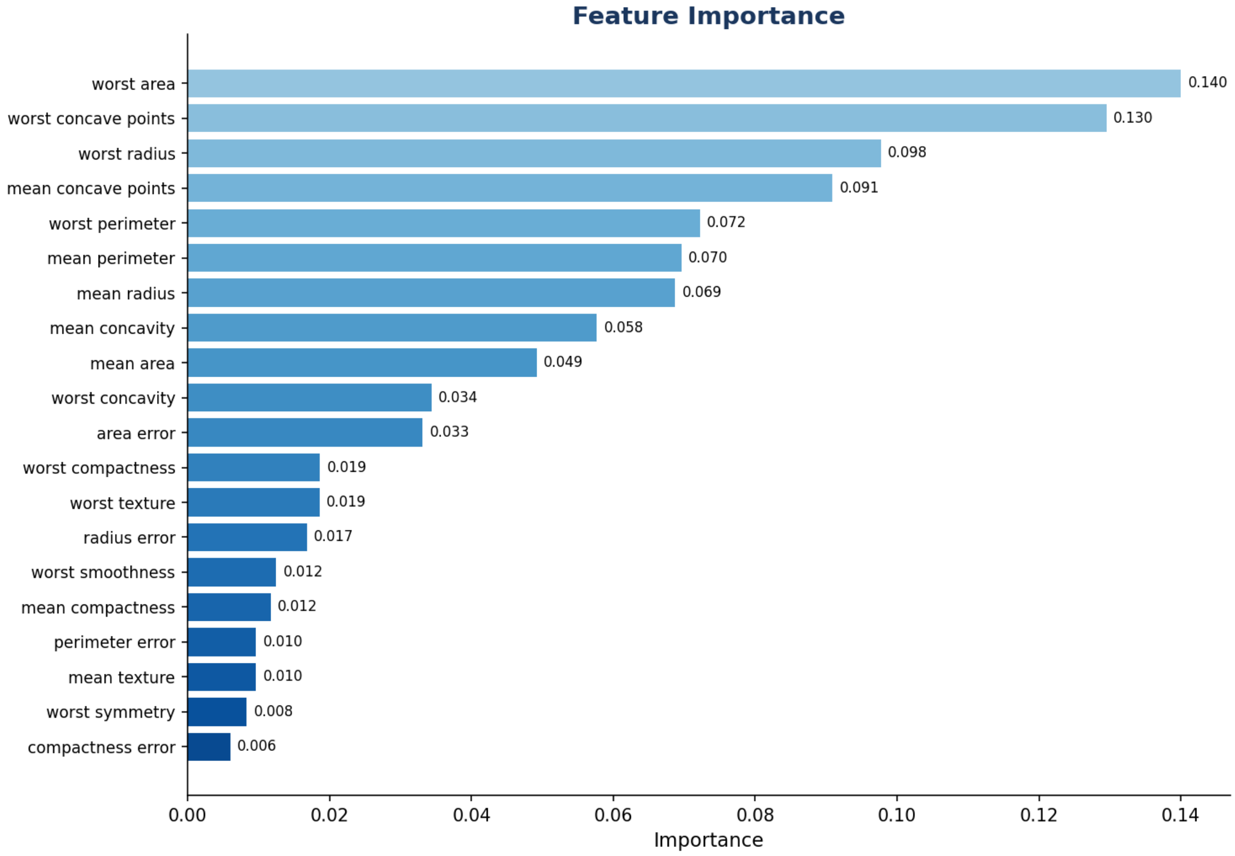


Figure 1: Feature Importance

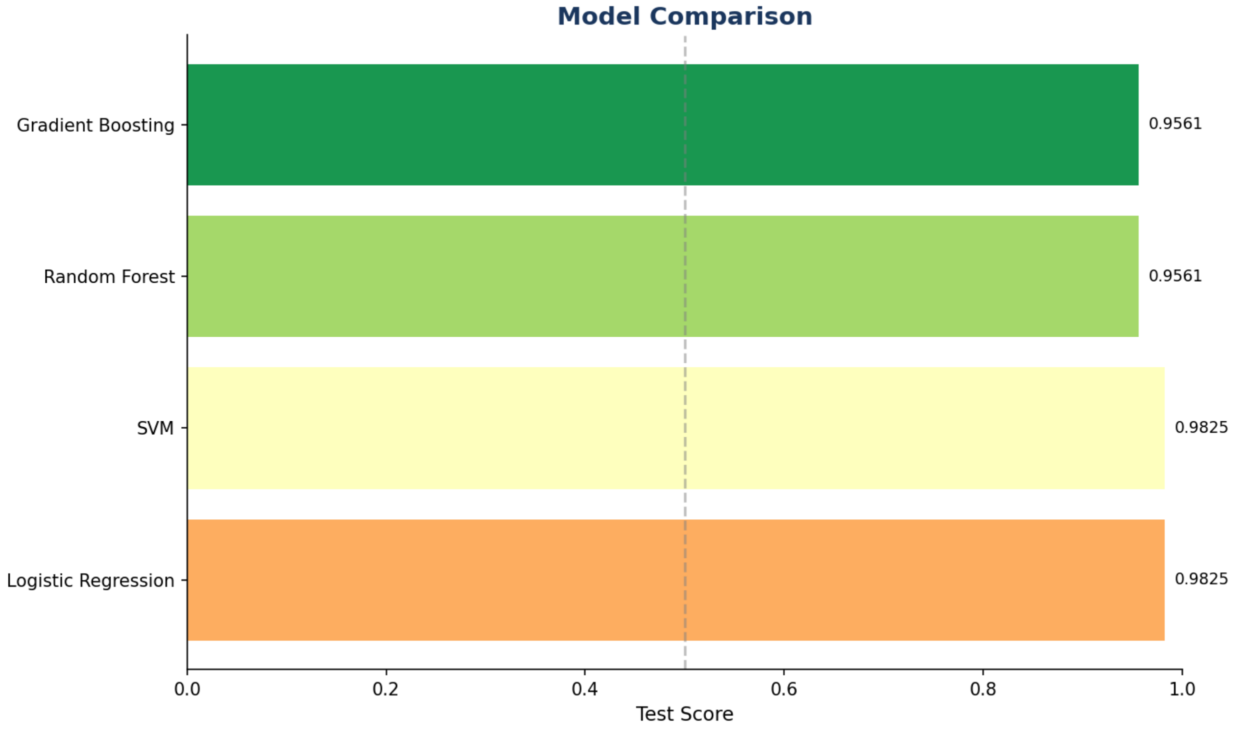
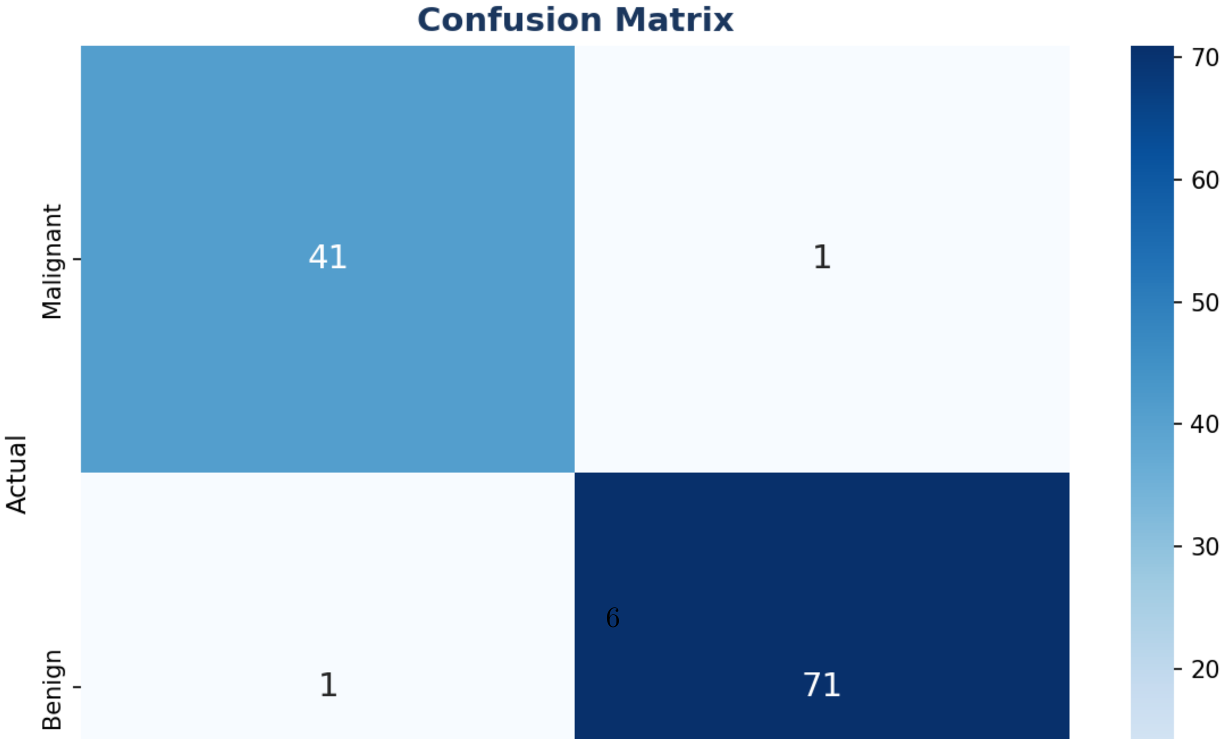


Figure 2: Model Benchmarking

Class	Precision	Recall	F1-Score	Support
Malignant	0.976	0.976	0.976	42
Benign	0.986	0.986	0.986	72
Accuracy			**0.982**	

Confusion Matrix



Key Metrics

- **AUC-ROC:** 0.9954
- **Optimal Threshold:** 0.3659

ROC Curve

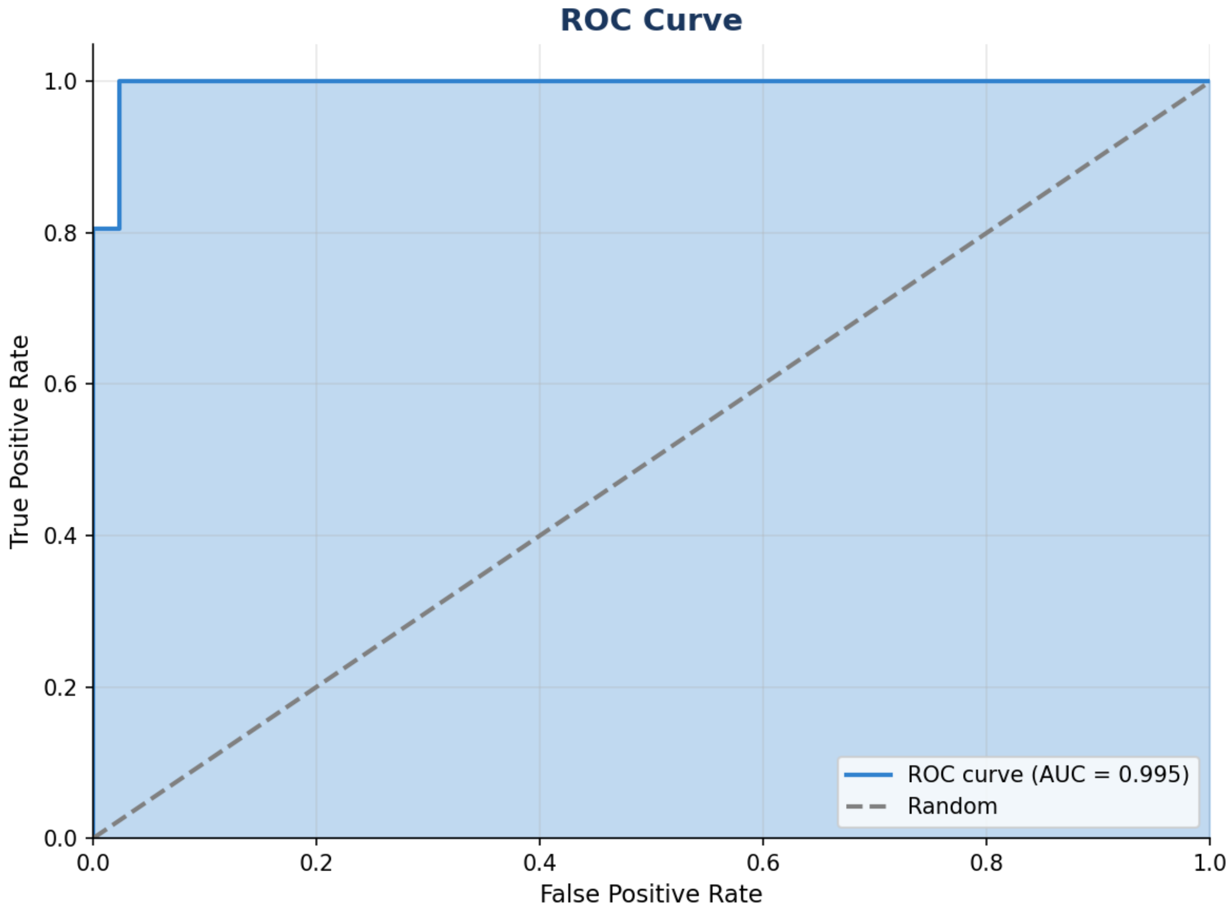


Figure 4: ROC Curve

Precision-Recall Analysis

The Precision-Recall curve is especially useful for imbalanced datasets, showing the trade-off between precision and recall.

Key Metrics

- **Average Precision:** 0.9971

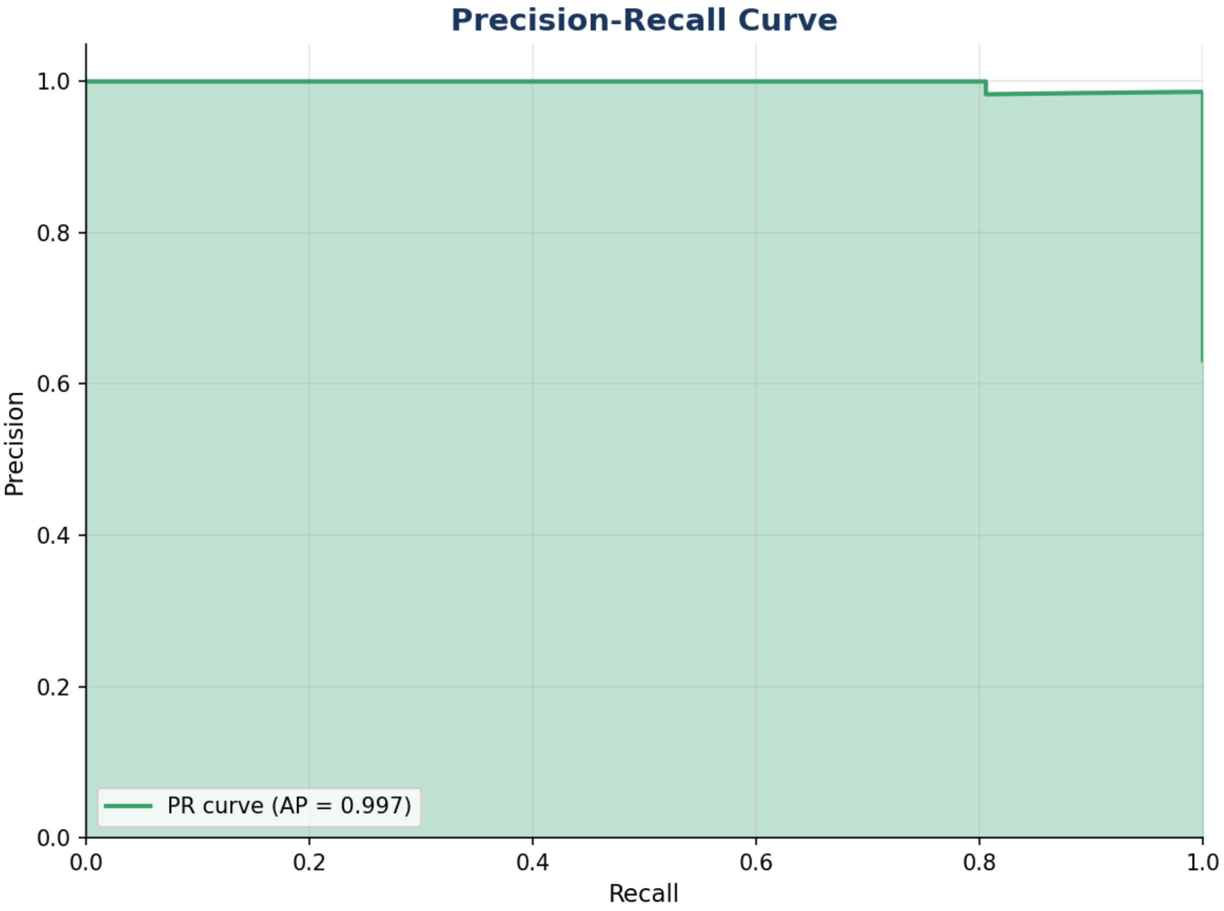


Figure 5: Precision-Recall Curve

Precision-Recall Curve

SHAP Feature Analysis

SHAP (SHapley Additive exPlanations) values provide model-agnostic explanations showing how each feature contributes to individual predictions.

SHAP Feature Importance

Feature	Mean
worst concave points	0.0681
worst area	0.0681
worst radius	0.0495
worst perimeter	0.0429
mean concave points	0.0411
mean concavity	0.0286
mean perimeter	0.0278
mean radius	0.0264
worst concavity	0.0229
mean area	0.0219
area error	0.0215
worst texture	0.0131
radius error	0.0091
worst smoothness	0.0088
worst compactness	0.0081

SHAP Summary Plot

Baseline Comparison

Performance Improvement

Model	Score	Improvement
Baseline	0.6270	-
Best Model	**0.9825**	**+0.3555 (+56.7%)**

The final model achieves a **56.7%** improvement over the baseline.

Recommendations & Next Steps

1. Best model Logistic Regression achieves 98.2% accuracy

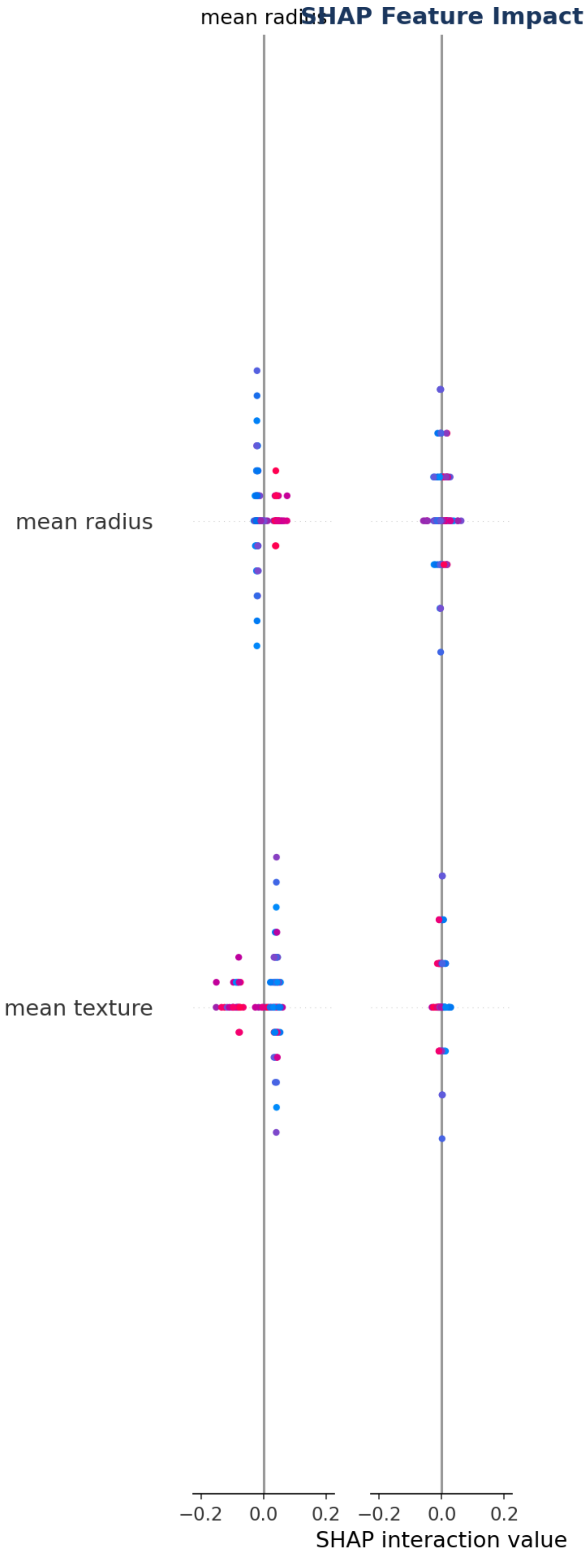


Figure 6: SHAP Analysis

2. High AUC-ROC (0.995) indicates excellent discrimination
3. Top features: worst perimeter, worst concave points, worst area
4. Model suitable for clinical decision support
5. Regular validation on new data recommended

Report generated by Jotty SwarmMLComprehensive on 2026-02-05 02:57:25