Advanced Breast Cancer Prediction System Using Machine Learning & Automated Model Selection

Navanit Sharma

Project Presentation

February 24, 2026

Agenda

1. Problem Statement & Motivation
2. Dataset Overview
3. Methodology (Preprocessing + Models)
4. Training Strategy (Cross-Validation + Tuning)
5. Evaluation Metrics & Results
6. Deployment (Streamlit App)
7. Demo Flow
8. Conclusion & Future Work

Problem Statement & Motivation

Goal

Predict whether a breast tumor is:

- **Malignant (Cancer)** or
- **Benign (Non-cancerous)**

based on measured tumor features.

- Early detection improves treatment outcomes.
- ML can assist clinicians by providing fast decision support.
- Focus on reliability and interpretability for medical screening.

Dataset Overview

- Dataset: **Breast Cancer Wisconsin (Diagnostic)**
- Type: Tabular, numeric features
- Target:
  - $1 \to$ Malignant
  - $0 \to$ Benign
- Features (examples):
  - radius, texture, perimeter, area
  - smoothness, compactness, concavity
  - symmetry, fractal dimension

Data Preprocessing

- Removed unnecessary column(s) such as `id` (if present).
- Encoded diagnosis:

$$y = \begin{cases} 1, & \text{if diagnosis} = \text{M} \\ 0, & \text{if diagnosis} = \text{B} \end{cases}$$

- Train-test split (e.g., 80/20) with **stratification**.
- Standardization where needed:

$$x' = \frac{x - \mu}{\sigma}$$

Models Compared

Candidate Models

- **Support Vector Machine (SVM)** (with scaling)
- **Random Forest Classifier**
- **XGBoost Classifier**

- Objective: Select the best model using cross-validation and ROC-AUC.
- Save the final pipeline for consistent inference during deployment.

Training Strategy

- **Stratified K-Fold Cross-Validation** (k=5)
- **Hyperparameter Optimization** using RandomizedSearchCV
- Model selection criterion: **Highest ROC-AUC**

Why ROC-AUC?

Useful for medical classification where class separation matters and threshold can be adjusted.

Evaluation Metrics

- Confusion Matrix:

	Predicted Benign	Predicted Malignant
Actual Benign	TN	FP
Actual Malignant	FN	TP

- Accuracy, Precision, Recall, F1-score
- ROC Curve and ROC-AUC:

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR)$$

Results Summary

- Best-performing model selected from: SVM / RF / XGBoost
- Reported:
  - ROC-AUC (Cross-Validation)
  - ROC-AUC (Test Set)
  - Confusion matrix and classification report
- Medical focus: prioritize **Recall** for malignant cases when needed.

Key Outcome

An optimized model pipeline saved as `best_pipeline.pkl` for reliable deployment.

ROC Curve Visualization

- The ROC curve compares True Positive Rate vs False Positive Rate.
- Higher curve and larger AUC indicate better separability.

roc_breast_cancer.jpeg

Performance Evaluation

PE_breast_cancer.jpeg

Deployment: Streamlit Web App

Features

- Manual input prediction
- CSV batch prediction + downloadable results
- Probability output + threshold adjustment
- Performance visuals (ROC)
- Educational section: causes, effects, prevention

Artifact Used

`best_pipeline.pkl` contains model + feature columns + threshold.

Demo Flow

1. Open the Streamlit app
2. Manual input:
   - Enter feature values $\rightarrow$ Predict
   - View label + probability bar
3. CSV mode:
   - Upload CSV $\rightarrow$ Predict batch
   - View summary: total / malignant / benign
   - Download predictions

Limitations

- Model trained on a specific dataset; performance may vary on real-world clinical data.
- Not a substitute for medical diagnosis.
- Requires consistent feature format and measurement standards.

Ethical Note

Predictions should be used only as decision support and validated by clinicians.

Future Work

- Explainable AI: SHAP/Feature importance dashboards
- Threshold optimization for high-sensitivity screening
- Deep learning with histopathology images (CNN/Transfer Learning)
- Deploy as REST API (FastAPI) + Docker + cloud deployment

Conclusion

- Built an advanced ML pipeline for breast cancer prediction.
- Compared multiple models and selected the best using ROC-AUC.
- Deployed as a user-friendly Streamlit application for real-time and batch predictions.

Thank You!