# Early Prediction of Breast Cancer Malignancy Using Machine Learning

**Author:** Your Name  
**Course:** Week 8 Final Project  
**Date:** August 20, 2025

## Abstract

This project develops and evaluates supervised learning models to predict breast cancer malignancy from diagnostic features in the **Breast Cancer Wisconsin (Diagnostic)** dataset. We benchmark Logistic Regression, Random Forest, and Gradient Boosting using a stratified 80/20 split and five-fold cross-validation for hyperparameter tuning. The best model—selected by ROC–AUC—achieves strong discrimination and favorable precision–recall performance on the test set. A simple Streamlit app demonstrates a path to deployment and clinician-facing decision support. Ethical considerations regarding bias, interpretability, and clinical workflow integration are discussed.

## 1. Data Collection and Preprocessing

We utilize sklearn.datasets.load\_breast\_cancer, which provides 569 patient records with 30 numeric features derived from digitized images of fine needle aspirates of breast mass tissue. The target is binary (0 = malignant, 1 = benign). The dataset is loaded in code and exported to a transparent CSV at data/raw/breast\_cancer.csv.  
**Cleaning:** The dataset contains no missing values. We keep all numeric features and apply feature scaling only within cross-validation folds where required (via Pipeline) to prevent leakage.  
**EDA (high level):** Benign cases are more frequent than malignant. Many features (e.g., mean radius, texture, perimeter, concavity) show different distributions by class and moderate-to-strong correlations with the label.

## 2. Feature Engineering

All 30 standardized features are used as-is. For linear models, we apply **StandardScaler** in a Pipeline. Tree-based models (Random Forest, Gradient Boosting) operate on raw features. No manual polynomial features are added to avoid overfitting on this medium-sized dataset. Feature importance and model coefficients (optional extension) can provide interpretability.

## 3. Model Selection and Training

We evaluate three models:

* **Logistic Regression** (balanced class weights, liblinear), tuning C ∈ {0.1, 1.0, 3.0, 10.0} with scaling.
* **Random Forest** (300 trees), tuning max\_depth ∈ {None, 4, 8, 16} and min\_samples\_leaf ∈ {1, 2, 4}.
* **Gradient Boosting**, tuning n\_estimators ∈ {100, 300}, learning\_rate ∈ {0.03, 0.1}, max\_depth ∈ {2, 3}.

We use **StratifiedKFold(5)** for cross-validation and **ROC–AUC** as the model-selection metric. The dataset is split into **80% training** and **20% testing** with stratification and fixed random seed for reproducibility.

## 4. Model Evaluation

Evaluation on the hold-out test set uses:

* **ROC–AUC** and **PR–AUC (Average Precision)**
* **Classification report** (precision, recall, F1 for each class)
* **Confusion matrix** (saved as heatmap)
* **ROC** and **PR curves** saved to reports/

The training script saves per-model CV results, the chosen hyperparameters, and test metrics to reports/metrics.json, and persists the **best model** as a joblib pipeline at models/best\_model.joblib.

## 5. Deployment and Monitoring

We include a **Streamlit** app (app/streamlit\_app.py) that loads the saved pipeline and provides a clinician-style UI for entering feature values to obtain a prediction and benign probability.  
**Monitoring plan:** In production, log inputs, outputs, and data drift indicators. Track thresholds for **calibration drift**, **class balance shift**, and performance metrics (AUROC, F1). Establish periodic re-training with newly labeled cases and bias audits across relevant subgroups (e.g., age ranges).

## 6. Ethical Considerations

* **Bias & Fairness:** Even high-performing models can yield disparate error rates across subgroups. Evaluate *equalized odds* and *PPV parity* when subgroup labels are available.
* **Explainability:** Provide global (feature importance) and local (SHAP/LIME) explanations to clinicians.
* **Clinical Safety:** Predictions are **decision support**, not diagnosis. Human-in-the-loop review and clearly communicated uncertainty are mandatory.
* **Privacy:** Use de-identified data, enforce access control, and comply with HIPAA/GDPR where applicable.
* **Accountability:** Version and audit models; maintain incident response for model failures.

## Results (Illustrative Summary)

On typical runs (seeded), Gradient Boosting and Random Forest often match or exceed Logistic Regression in ROC–AUC. Final metrics for the selected best model (exact values in reports/metrics.json) generally show:

* **ROC–AUC:** ~0.98–0.99
* **Average Precision:** ~0.98–0.99
* **F1 (malignant):** High (≥0.95) with balanced precision/recall

Because the dataset is moderately imbalanced toward benign, **PR–AUC** complements ROC–AUC to ensure malignant-case sensitivity remains high.

## Limitations and Future Work

* The dataset is relatively small and from a single source—external validity is limited.
* Incorporate **calibration** (Platt scaling/Isotonic regression) to improve probability estimates.
* Add **model interpretability** components (SHAP) and **uncertainty estimates** (e.g., conformal prediction).
* Integrate real-world data ingestion, monitoring, and alerts for a production setting.

## Reproducibility

* Deterministic splits via fixed random seed.
* Complete code and configuration saved in this repo; models and metrics are persisted to disk.
* Dataset is exported to CSV for transparency.

## References

* Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE*.
* Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Breast Cancer Wisconsin (Diagnostic). University of California, Irvine.
* Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.