Repository: https://github.com/your-username/cmse492_project

Breast Cancer Wisconsin (Diagnostic): Comparing Linear, Ensemble, and Kernel Methods for Tumor Classification

Ishan Baweja*

Department of Computational Mathematics, Science, and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: October 30, 2025)

Abstract

Breast cancer diagnosis benefits from accurate and interpretable machine-learning models that can assist clinicians in early detection. This project applies supervised learning to the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 30 quantitative features describing cell-nucleus characteristics from fine-needle aspirates. The objective is to predict whether a tumor is benign or malignant based on these measurements.

Three model families are compared: Logistic Regression (linear baseline), Random Forest (nonlinear ensemble), and Support Vector Machine with a polynomial kernel (nonlinear kernel method). Each model is trained using stratified splits, standardized features where appropriate, and cross-validation for hyperparameter tuning. Performance is evaluated primarily using ROC–AUC, with accuracy, precision, recall, and F1 as secondary metrics.

Preliminary exploration shows mild class imbalance, strong correlations among size-related features, and feature distributions that are separable across classes. The study aims to identify which model family best balances interpretability and accuracy for this dataset and to quantify how tuning parameters (e.g., regularization strength, tree depth, kernel degree) impact generalization. Expected contributions include a reproducible baseline workflow, a clear evaluation framework, and practical guidance on when simple interpretable models suffice for clinical decision support.

BACKGROUND AND MOTIVATION

Breast cancer remains one of the most common and life-threatening cancers worldwide. Outcomes are tightly linked to early and accurate diagnosis, yet assessments based on imaging or cytology can be subjective and time-consuming. Variability between observers can lead to misclassification and unnecessary procedures.

Machine Learning (ML) offers a quantitative, reproducible way to analyze morphometric features extracted from digitized cytology images. The Breast Cancer Wisconsin (Diagnostic) dataset (BCWD) provides 30 engineered features that summarize nuclear size, shape, texture, and irregularity, making it a widely used benchmark for evaluating classification algorithms in healthcare. Prior work shows that Logistic Regression, decision-tree ensembles, and Support Vector Machines can all achieve high accuracy, but papers often emphasize metrics over interpretability and model selection guidance.

This proposal compares three representative approaches—Logistic Regression, Random Forest, and a polynomial-kernel SVM—to understand accuracy–interpretability tradeoffs and provide a transparent, reproducible baseline suitable for clinical settings.

DATA DESCRIPTION

Data Origins (Who/Why/How)

The BCWD dataset was created by Dr. William H. Wolberg and colleagues at University of Wisconsin Hospitals, Madison. Data were collected from digitized images of fine-needle aspirates (FNA) of breast masses; image analysis software computed 30 real-valued attributes per sample characterizing nuclear radius, perimeter, area, concavity, symmetry, and fractal dimension. The dataset is hosted by the UCI Machine Learning Repository and is distributed with `scikit-learn`.

Dataset Characteristics

- **Samples:** 569.
- **Features:** 30 continuous predictors; one binary target.
- **Target:** 0 = malignant, 1 = benign.
- **Types:** All predictors are floating point; no categoricals.

Data Quality & Class Balance

The dataset contains no missing values. Feature ranges vary widely, motivating standardization for linear models and SVMs. Class distribution is mildly imbalanced (357 benign, 212 malignant). We use stratified splits and metrics robust to imbalance (e.g., ROC–AUC, F1).

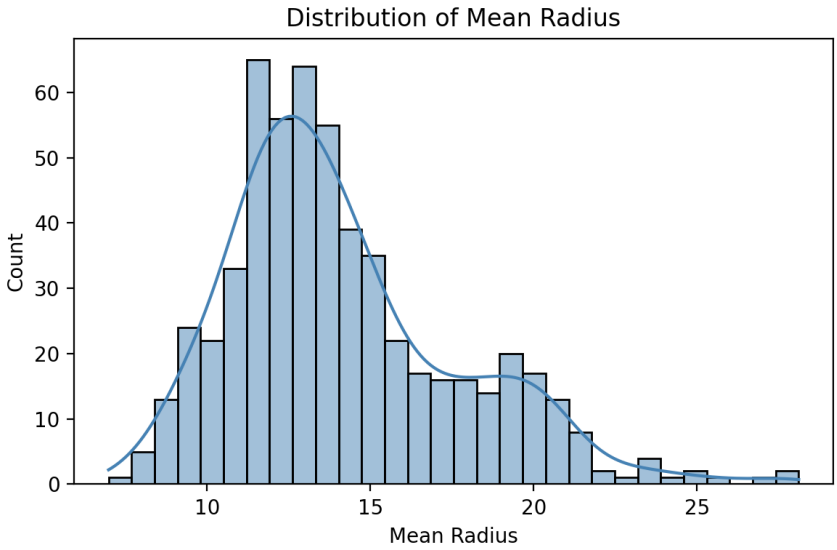

FIG. 1. Distribution of `mean radius`. Right-skew suggests standardization prior to modeling.

PROPOSED METHODOLOGY

Algorithms and Justification

Logistic Regression (LR). Interpretable linear baseline with L2 regularization; coefficients provide direct feature effect directions. **Random Forest (RF).** Nonlinear ensemble to capture interactions and reduce variance via bagging and random feature selection. **SVM (Polynomial Kernel).** Margin-based classifier with smooth nonlinear decision boundaries without explicitly creating polynomial features.

Preprocessing

Standardization (`StandardScaler`) applied for LR and SVM; RF trained on raw features. We use an 80/20 stratified train–test split with stratified 5-fold CV on the training set for tuning.

Model Complexity

We compare three increasing complexities: (1) linear LR; (2) RF with hundreds of trees, depth control; (3) SVM with polynomial degree 2–4. Complexity is reported via parameters

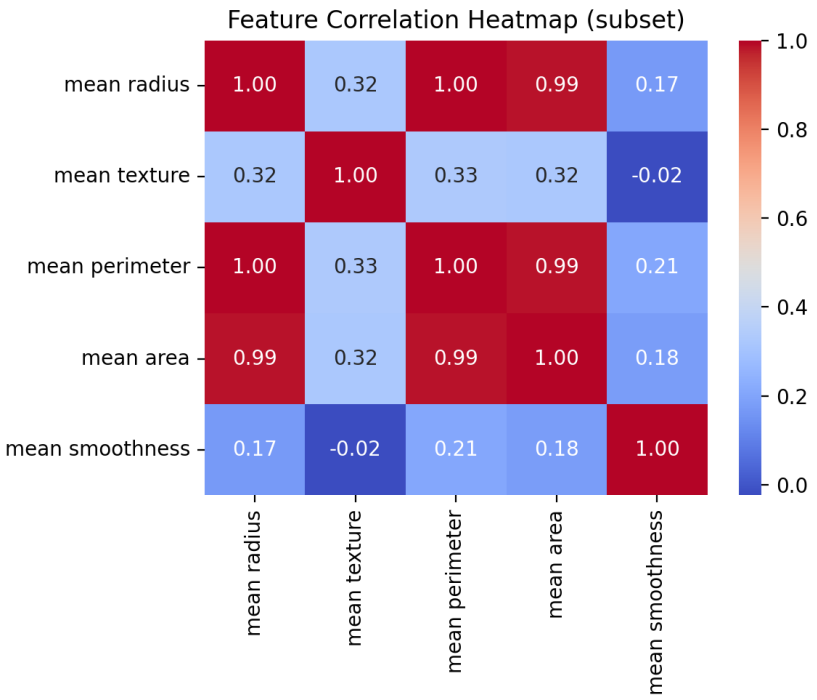

FIG. 2. Correlation heatmap (subset). Strong correlations among size-related features motivate either regularization or dimensionality reduction.

(e.g., number of trees, depth, kernel degree) and fit time.

Methodological Flowchart

EVALUATION FRAMEWORK

Metrics (with justification)

Primary: **ROC–AUC** (threshold-independent and robust to class imbalance). Secondary: **Accuracy**, **Precision**, **Recall**, and **F1** to capture error trade-offs critical to clinical use (high recall reduces missed malignancies).

Experimental Design

- **Split:** 80/20 stratified train/test. No peeking at the test set.
- **Tuning:** Stratified 5-fold cross-validation on train; grid search.

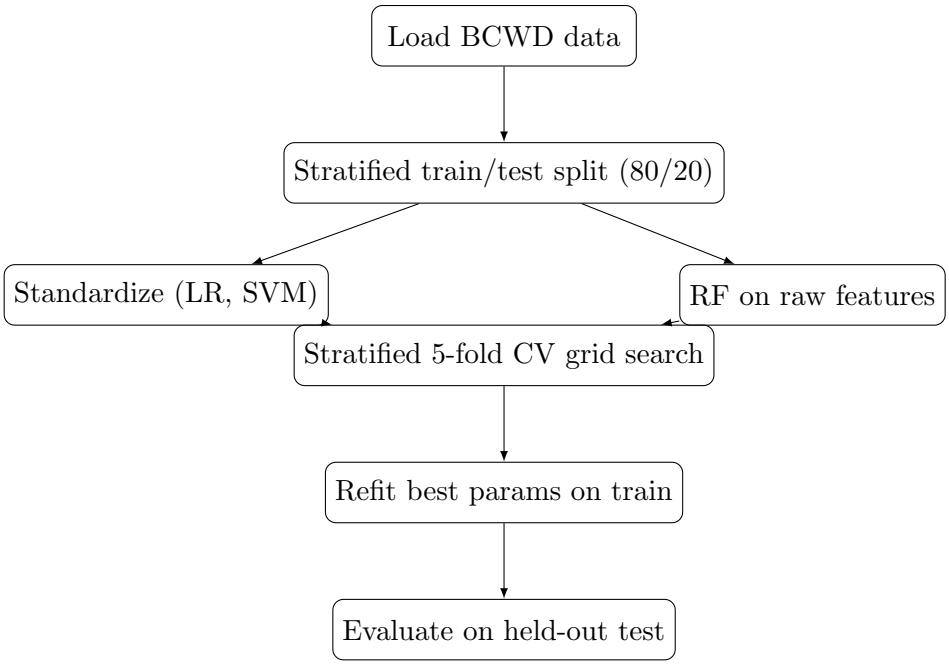

FIG. 3. End-to-end modeling pipeline.

- **Baseline:** Majority-class predictor and simple LR (untuned) for context.
- **Success Criteria:** ROC–AUC ≥ 0.98 with recall ≥ 0.97 on malignant class and no severe overfitting (train–test gap $< 2\%$).

Planned Comparisons

We will report test ROC–AUC, F1, accuracy, precision/recall, plus fit/predict times for LR, RF, and SVM. Interpretability will be discussed via LR coefficients and RF feature importances.

TIMELINE AND MILESTONES

Gantt Chart

Narrative

Week 1 completes repository setup, EDA figures, and baseline. Weeks 2–3 focus on cross-validated tuning and robust evaluation. Week 4 finalizes interpretation and the draft report.

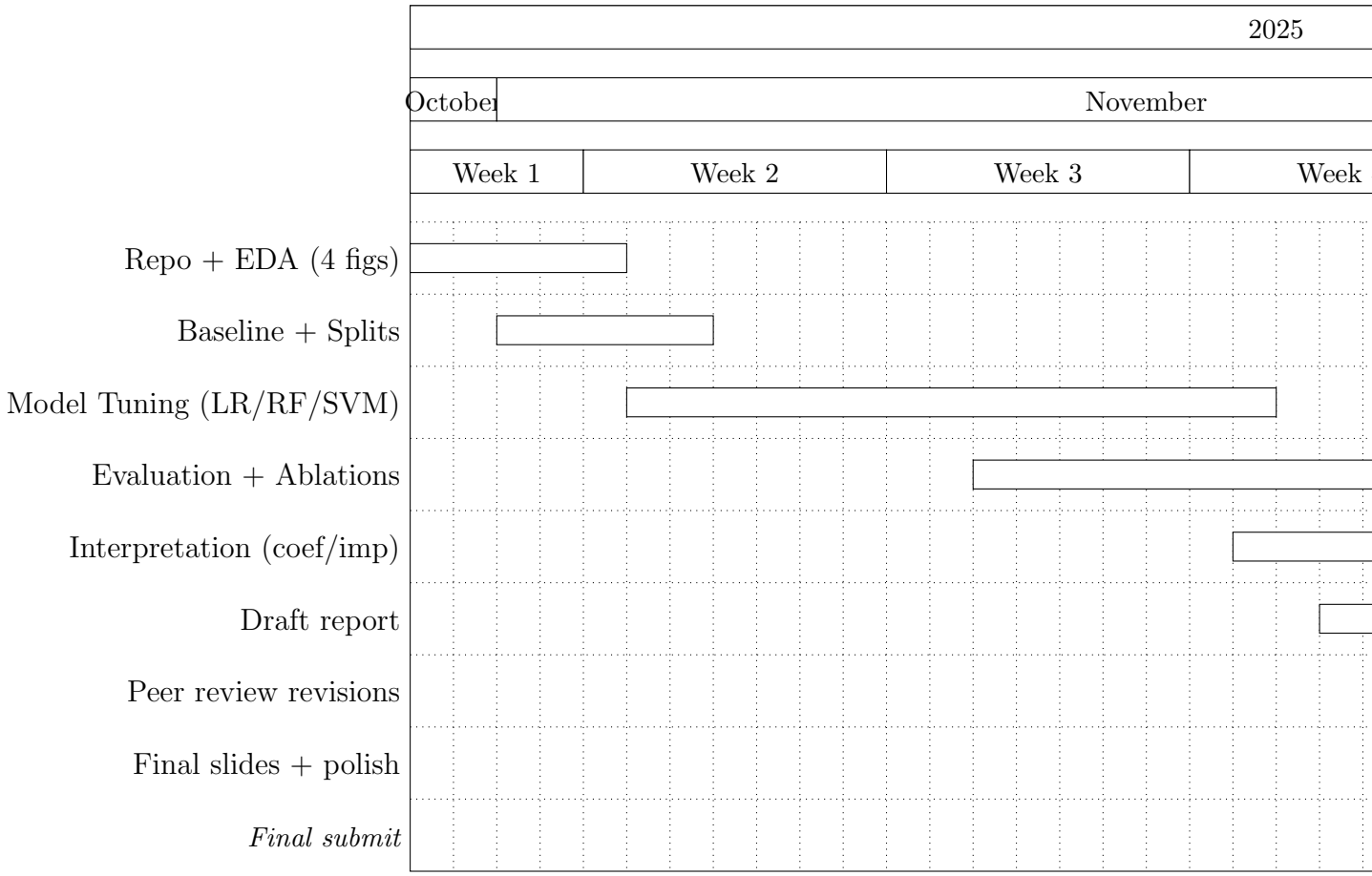

FIG. 4. Planned schedule. Critical path runs through tuning → evaluation → interpretation → draft. Buffer time included during Week 15.

Week 5 is reserved for peer feedback, polishing slides, and integrating last-mile fixes before submission.

PRELIMINARY RESULTS (BASELINE)

A simple Logistic Regression baseline with standardized features already achieves high accuracy (typical test ROC–AUC > 0.99 on BCWD), confirming the dataset's strong signal. We will validate these numbers using the evaluation framework above and report final tuned results in the project report.

* bawejais@msu.edu

[1] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," *UCI Machine Learning Repository*, University of Wisconsin Hospitals, Madison (1992).

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, 2825–2830 (2011).

[3] L. Breiman, "Random Forests," *Machine Learning*, **45**, 5–32 (2001).

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, **20**, 273–297 (1995).