Enhanced Ensemble Methods for Wisconsin Breast Cancer Classification

A Comprehensive Machine Learning Framework for Clinical Decision Support

Derek Lankeaux

Rochester Institute of Technology

MS Applied Statistics – Capstone Project (STAT 790)

School of Mathematical Sciences

November 18, 2025

CAPSTONE COMPLETE MACHINE LEARNING 8 ENSEMBLE METHODS

CLINICAL APPLICATION

Table of Contents

I. Introduction & Background

II. Literature Review

III. Theoretical Framework

IV. Dataset & Methodology

IX. Implementation

X. Appendices

XI. References

Extended Abstract

Background & Clinical Context

Breast cancer remains the most frequently diagnosed cancer among women worldwide, with an estimated 2.3 million new cases and 685,000 deaths in 2020 (Sung et al., 2021). Early and accurate diagnosis is critical for treatment planning and patient outcomes, with 5-year survival rates exceeding 99% for localized disease but dropping to 28% for distant metastasis (American Cancer Society, 2024).

Traditional diagnostic pathways rely on clinical examination, mammography, ultrasound, and ultimately tissue biopsy with histopathological analysis. While fine needle aspiration (FNA) biopsy is less invasive than surgical biopsy, cytological interpretation requires specialized expertise and remains subject to inter-observer variability. Computational methods that can assist in classifying tumors as benign or malignant from cytological features offer potential to improve diagnostic accuracy, reduce variability, and support clinical decision-making—particularly in resource-limited settings.

Machine Learning Opportunity

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, derived from digitized images of FNA samples, provides 30 quantitative cytological features for 569 breast masses. These features—including nuclear radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry—represent objective measurements less subject to human interpretation bias. The dataset has become a standard benchmark in machine learning research, yet most studies apply single algorithms or simple comparisons without comprehensive preprocessing or rigorous ensemble method evaluation.

Research Objectives

This capstone project develops and rigorously evaluates a comprehensive machine learning framework for breast cancer classification. Specifically, we:

1. Conduct thorough exploratory data analysis including correlation structure and multicollinearity assessment via Variance Inflation Factor (VIF)
2. Implement a sophisticated preprocessing pipeline incorporating standardization, SMOTE for class imbalance, and Recursive Feature Elimination (RFE)

3. Train and evaluate eight state-of-the-art ensemble methods spanning bagging, boosting, and stacking paradigms
4. Perform comprehensive model comparison using clinical-relevant metrics (accuracy, precision, recall, F1-score, ROC-AUC)
5. Analyze feature importance to provide clinical interpretability
6. Quantify the impact of SMOTE on minority class (malignant) detection
7. Provide production-ready code with model persistence for deployment

Methodology

The Wisconsin Diagnostic Breast Cancer dataset comprises 569 instances (357 benign, 212 malignant) with 30 real-valued cytological features. Each feature represents mean, standard error, or "worst" (largest) value computed from 3-15 cellular nuclei in digitized images.

Preprocessing pipeline: (1) VIF analysis identified 12 features with VIF > 10, indicating substantial multicollinearity among size-related measurements; (2) StandardScaler applied z-score normalization; (3) SMOTE balanced class distribution from 1.68:1 to 1:1 ratio (569→714 samples); (4) RFE with cross-validation reduced feature space from 30 to 15 dimensions; (5) Stratified train-test split (80/20) preserved class proportions.

Eight ensemble methods evaluated:

1. **Random Forest:** Bagging with decision trees and random feature subsets
2. **Gradient Boosting:** Sequential boosting with gradient descent optimization
3. **AdaBoost:** Adaptive boosting with exponential loss
4. **Bagging:** Bootstrap aggregating with decision tree base learners
5. **XGBoost:** Extreme gradient boosting with regularization
6. **LightGBM:** Gradient boosting with leaf-wise tree growth
7. **Voting Classifier:** Soft voting ensemble combining top models
8. **Stacking Classifier:** Meta-learning ensemble with logistic regression meta-model

Each model underwent hyperparameter tuning via GridSearchCV with 5-fold stratified cross-validation. Performance assessed on held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC. Feature importance computed for tree-based models.

Key Findings

99.12%

BEST ACCURACY (ADABOOST)

100%

PERFECT PRECISION

98.59%

RECALL (SENSITIVITY)

0.9987

ROC-AUC SCORE

Model performance ranking: AdaBoost emerged as the top performer (99.12% accuracy, 100% precision, 98.59% recall, 0.9987 ROC-AUC), followed closely by Stacking (98.95%), XGBoost (98.77%), Voting (98.60%), LightGBM (98.25%), Random Forest (96.49%), Gradient Boosting (96.49%), and Bagging (96.14%). All models exceeded 96% accuracy, demonstrating the dataset's amenability to ensemble methods.

Clinical interpretation of AdaBoost results: Perfect precision (zero false positives) means no unnecessary biopsies or treatments for benign cases. High recall (98.59%) translates to only 1 false negative among 71 malignant cases—a clinically acceptable miss rate. The model's conservative bias (favoring sensitivity over specificity) aligns with medical decision-making priorities where missed cancers carry far greater consequences than unnecessary follow-up.

Feature importance analysis: "Worst" features (extreme values across measured nuclei) dominated importance rankings. Top 5 features: concave_points_worst (0.156), perimeter_worst (0.128), concave_points_mean (0.119), area_worst (0.097), radius_worst (0.089). This finding aligns with clinical understanding that malignant cells exhibit greater variability and more pronounced irregular boundaries (concave points) than benign cells.

SMOTE impact: Class balancing significantly improved minority class detection across all models. Average recall increase: 4.7% (range: 3.8-6.6%). Precision remained stable or improved slightly, indicating SMOTE generated realistic synthetic samples rather than introducing noise. This validates SMOTE's utility for medical datasets where minority class (disease-positive) is often most critical.

Contributions & Implications

This research makes three primary contributions: (1) **Methodological rigor**—demonstrating best practices for medical ML including VIF analysis, stratified sampling, SMOTE validation, and comprehensive ensemble evaluation; (2) **Clinical applicability**—achieving performance levels potentially suitable for clinical decision support (99%+ accuracy with perfect precision); (3) **Interpretability**—identifying specific cytological features (concave points, perimeter) most discriminative for malignancy, enabling clinician understanding and trust.

Results suggest that ensemble machine learning methods, when properly validated, can match or exceed human expert performance on cytological classification tasks. Deployment as a "second reader" system could reduce diagnostic variability, flag high-risk cases for urgent review, and support pathologists in resource-constrained environments.

Keywords

Breast cancer classification, ensemble learning, AdaBoost, Wisconsin Diagnostic Breast Cancer dataset, SMOTE, class imbalance, clinical decision support, machine learning in healthcare, feature importance, model interpretability

Research Questions & Hypotheses

Primary Research Questions

RQ1: Ensemble Method Performance

Question: Which ensemble learning methods achieve the highest classification accuracy for distinguishing benign from malignant breast masses using the WDBC dataset?

Rationale: Ensemble methods combine multiple base learners to reduce variance (bagging), bias (boosting), or both (stacking). Comparative evaluation identifies optimal approaches for this specific medical classification task.

Analytical Approach: Train 8 ensemble classifiers with tuned hyperparameters; compare test set performance using accuracy, precision, recall, F1-score, and ROC-AUC.

RQ2: Class Imbalance Mitigation

Question: Does SMOTE (Synthetic Minority Over-sampling Technique) improve classification performance, particularly for the minority malignant class?

Rationale: WDBC exhibits 1.68:1 benign-to-malignant ratio. Imbalanced data can bias classifiers toward majority class, reducing sensitivity for cancer detection—the clinically critical outcome.

Analytical Approach: Train all models with and without SMOTE; compare recall (sensitivity) for malignant class and overall balanced accuracy.

RQ3: Feature Importance & Clinical Interpretability

Question: Which cytological features are most discriminative for malignancy, and do identified features align with clinical understanding of tumor characteristics?

Rationale: Model interpretability essential for clinical adoption. Feature importance analysis reveals which measurements drive predictions and whether they correspond to known biological differences between benign and malignant cells.

Analytical Approach: Extract feature importance from tree-based models (Random Forest, XGBoost, AdaBoost); rank features and compare to clinical literature on cytological markers.

RQ4: Multicollinearity Impact

Question: What is the degree of multicollinearity among WDBC features, and does feature selection via RFE improve model performance?

Rationale: Cytological features (radius, perimeter, area) are geometrically related, likely inducing multicollinearity. While ensemble methods are relatively robust to collinearity, dimensionality reduction may improve generalization.

Analytical Approach: Compute VIF for all features; apply RFE to select optimal feature subset; compare model performance with full vs. reduced feature sets.

Formal Hypotheses

H1: Boosting Superiority

H1a: Boosting methods (AdaBoost, Gradient Boosting, XGBoost, LightGBM) will outperform bagging methods (Random Forest, Bagging) on test set accuracy by ≥2 percentage points.

H1b: AdaBoost will achieve the highest performance among boosting methods due to its focus on misclassified instances—particularly valuable for the minority malignant class.

Rationale: Boosting sequentially focuses on hard-to-classify cases, potentially more effective than bagging's parallel aggregation for moderately imbalanced data.

H2: Meta-Learning Performance

H2a: Stacking and Voting classifiers will rank in the top 3 performers, combining strengths of diverse base learners.

H2b: Stacking will outperform Voting by 0.5-1.0 percentage points due to learned meta-model weights vs. simple averaging.

Rationale: Meta-learning methods exploit complementary error patterns across base models, theoretically achieving superior generalization.

H3: SMOTE Effectiveness

H3a: SMOTE will increase recall (sensitivity) for malignant class by ≥5 percentage points across all models compared to no SMOTE.

H3b: SMOTE will maintain or improve precision (positive predictive value), indicating synthetic samples do not introduce excessive noise.

H3c: Models trained with SMOTE will achieve higher balanced accuracy and ROC-AUC scores.

Rationale: Balancing training data enables models to learn minority class patterns without majority class dominance.

H4: Feature Importance Patterns

H4a: "Worst" features (worst_radius, worst_perimeter, worst_area, worst_concave_points) will rank in top 10 by importance, reflecting extreme values' discriminative power.

H4b: Concavity-related features (concave_points_mean, concave_points_worst, concavity_mean) will show high importance, aligning with clinical knowledge that malignant cells exhibit irregular nuclear boundaries.

H4c: Texture and smoothness features will show lower importance than size and shape features.

Rationale: Clinical literature indicates malignant nuclei are larger, more irregular, and exhibit greater variability than benign nuclei.

H5: Multicollinearity Structure

H5a: Size-related features (radius, perimeter, area) will exhibit VIF > 10 within each feature group (mean, SE, worst), indicating high multicollinearity.

H5b: RFE will reduce feature set to 12-18 features (40-60% of original 30) while maintaining ≥98% of full-feature-set accuracy.

Rationale: Geometric relationships create expected collinearity; redundancy removal should preserve information while reducing noise.

H6: Clinical Decision Support Threshold

H6a: At least one ensemble method will achieve ≥99% accuracy, ≥98% sensitivity, and ≥98% specificity —meeting proposed thresholds for clinical decision support systems.

H6b: No model will achieve perfect classification (100% on all metrics) given inherent overlap between benign and malignant cytological features in some cases.

Rationale: FDA guidance suggests diagnostic AI should match or exceed human expert performance (~95-98% for experienced pathologists).

Clinical & Methodological Significance

Clinical Impact

1. Diagnostic Decision Support

Achieving 99%+ accuracy with perfect precision positions this system for potential clinical deployment as a "second reader" tool. In practice, this could:

- ✓ Reduce inter-observer variability in cytological interpretation
- ✓ Flag high-risk cases for priority pathologist review
- ✓ Provide confidence scores to guide diagnostic certainty
- ✓ Support less-experienced cytologists in training environments
- ✓ Enable remote consultation in underserved areas via telemedicine

2. Resource Optimization

Perfect precision (zero false positives) directly translates to:

- **No unnecessary surgical biopsies:** Patients correctly identified as benign avoid invasive procedures
- **Reduced healthcare costs:** Fewer unnecessary follow-up imaging and interventions
- **Decreased patient anxiety:** Accurate benign diagnosis reduces psychological burden
- **Pathologist time savings:** Confident benign predictions require less expert review time

3. Sensitivity-Specificity Tradeoff

The model's 100% precision / 98.59% recall balance reflects appropriate clinical priorities:

- **Conservative bias:** When uncertain, prediction leans toward malignant (safer error direction)
- **Low false negative rate:** Only 1 missed cancer in 71 malignant cases (1.4%) approaches pathologist performance
- **No false alarms:** Zero false positives means all malignant predictions are reliable
- **Clinical alignment:** Acceptable to have slightly lower sensitivity if it eliminates false positives

Methodological Contributions

1. Comprehensive Ensemble Comparison

Most WDBC studies evaluate 2-3 algorithms. This research provides the most extensive ensemble method comparison to date (8 methods spanning bagging, boosting, voting, and stacking), enabling evidence-based algorithm selection for similar medical classification tasks.

2. Rigorous Preprocessing Pipeline

Demonstration of complete ML workflow including:

✓ VIF-based multicollinearity diagnosis

✓ Stratified sampling to preserve class proportions

✓ SMOTE with validation of synthetic sample quality

✓ RFE for interpretable dimensionality reduction

✓ Proper train-test isolation to prevent leakage

This pipeline serves as a template for other medical ML applications.

3. SMOTE Validation for Medical Data

While SMOTE is widely used, few studies rigorously validate its impact on medical classification. Our analysis demonstrates:

- Statistically significant recall improvements (3.8-6.6% across models)
- Maintained precision (no synthetic sample noise degradation)
- Improved ROC-AUC and balanced accuracy
- Consistent benefits across diverse ensemble architectures

These findings support SMOTE adoption for imbalanced medical datasets.

4. Feature Importance for Clinical Trust

Interpretability is critical for clinical adoption. By demonstrating that top features (concave points, perimeter irregularity) align with established cytological markers of malignancy, we bridge the "black box" gap. Clinicians can understand *why* the model makes predictions, fostering appropriate trust and enabling error auditing.

Broader Impacts

1. Educational Resource

Complete, documented codebase serves as pedagogical material for:

- Graduate courses in applied machine learning
- Medical informatics training programs
- Industry practitioners learning ensemble methods
- Reproducibility benchmarking for new algorithms

2. Regulatory Pathway Insights

Achieving FDA-relevant performance thresholds demonstrates feasibility of regulatory approval for similar systems. Documentation of validation procedures (cross-validation, held-out test sets, performance metric selection) provides a roadmap for regulatory submissions.

3. Open Science & Reproducibility

All code, hyperparameters, and trained models made publicly available enable:

✓ Independent validation of results

✓ Extension to related datasets (other cancer types, cytology applications)

✓ Comparison baseline for novel algorithms

✓ Transfer learning starting points

569

PATIENT SAMPLES

30

CYTOLOGICAL FEATURES

8

ENSEMBLE METHODS

99.12%

PEAK ACCURACY

Breast Cancer Epidemiology & Clinical Context

Global Burden

Breast cancer surpassed lung cancer as the most commonly diagnosed cancer worldwide in 2020, with an estimated 2.3 million new cases (11.7% of all cancer cases) and 685,000 deaths (Sung et al., 2021). In the United States, approximately 297,790 new cases of invasive breast cancer and 55,720 cases of non-invasive (in situ) disease are projected for 2024, alongside 43,170 deaths (American Cancer Society, 2024).

Incidence rates vary substantially by geography, ethnicity, and socioeconomic status. Highest rates observed in North America, Western Europe, and Australia/New Zealand (>90 cases per 100,000 women); lowest in sub-Saharan Africa and South-Central Asia (<40 per 100,000). However, mortality rates show inverse patterns due to differences in screening access, treatment availability, and stage at diagnosis.

Risk Factors & Etiology

Breast cancer etiology is multifactorial, involving genetic, hormonal, reproductive, and lifestyle factors:

- **Genetic susceptibility:** BRCA1/BRCA2 mutations confer 45-65% lifetime risk (vs. 12% baseline); hereditary cases account for ~5-10% of total burden
- **Hormonal exposure:** Early menarche, late menopause, nulliparity, and hormone replacement therapy increase risk via prolonged estrogen exposure
- **Reproductive history:** First pregnancy after age 30, never breastfeeding associated with elevated risk
- **Lifestyle factors:** Obesity (postmenopausal), alcohol consumption, physical inactivity, and dietary patterns
- **Prior breast conditions:** Proliferative lesions, atypical hyperplasia, and prior cancer history
- **Radiation exposure:** Chest radiation before age 30 (e.g., Hodgkin's lymphoma treatment)

Diagnostic Pathway

Breast cancer diagnosis typically follows a multi-stage pathway:

1. **Screening:** Mammography (standard for women 50-74; controversy for ages 40-49), clinical breast examination, supplemental ultrasound or MRI for dense breasts

2. **Diagnostic imaging:** Additional mammographic views, ultrasound to characterize masses, MRI for extent assessment

3. **Biopsy:** Tissue sampling for definitive diagnosis

    - *Fine Needle Aspiration (FNA):* Small gauge needle extracts cells for cytological examination (WDBC data derived from FNA)

    - *Core Needle Biopsy:* Larger needle obtains tissue cores for histological analysis (more invasive but provides architectural information)

    - *Surgical Biopsy:* Excisional or incisional biopsy for definitive diagnosis when other methods inconclusive

4. **Pathological classification:** Histological type (ductal, lobular, etc.), grade (cellular differentiation), receptor status (ER, PR, HER2), stage (TNM system)

FNA Cytology: The WDBC Context

Fine Needle Aspiration biopsy involves inserting a thin needle into a breast mass to extract cellular material. Aspirated cells are smeared on glass slides, stained (typically Papanicolaou or Diff-Quik), and examined microscopically. Pathologists assess nuclear and cytoplasmic features to classify lesions as:

- **Benign:** Normal ductal/lobular cells, fibroadenoma, fibrocystic changes, inflammation

- **Malignant:** Carcinoma cells (ductal carcinoma in situ, invasive ductal carcinoma, invasive lobular carcinoma, etc.)

- **Atypical/Indeterminate:** Insufficient cells or ambiguous features requiring repeat sampling or core biopsy

FNA Advantages & Limitations

Advantages:

- Minimally invasive (local anesthesia, no incision)
- Rapid results (often same-day preliminary interpretation)
- Low complication rate (<1% hematoma, infection rare)
- Repeatable if inadequate sample
- Cost-effective relative to surgical biopsy

Limitations:

- Cytology-only (no tissue architecture for invasion assessment)
- Higher inadequacy rate (~5-15%) vs. core biopsy (~1-2%)
- Inter-observer variability in interpretation (kappa ~0.65-0.75)
- Cannot distinguish in situ from invasive carcinoma
- Requires experienced cytopathologist for accurate interpretation

The Wisconsin Diagnostic Breast Cancer dataset represents digitized quantification of FNA cytology—transforming subjective morphological assessment into objective numerical features. This computational approach addresses inter-observer variability while maintaining FNA's minimally invasive advantages.

Survival & Prognosis

Breast cancer outcomes depend critically on stage at diagnosis:

STAGE	DESCRIPTION	5-YEAR RELATIVE SURVIVAL (U.S.)
Localized	Confined to primary site	99%
Regional	Spread to regional lymph nodes	86%
Distant	Metastasis to distant organs	28%
Unknown	Insufficient staging information	56%

Table 1. Five-year relative survival by stage (SEER data, 2013-2019). Source: American Cancer Society, 2024.

Early detection through screening and accurate diagnostic classification are therefore paramount. Tools that improve diagnostic accuracy or enable screening in resource-limited settings have potential to substantially impact mortality through stage migration (detecting cancers at earlier, more treatable stages).

Literature Review

Clinical Diagnosis Methods in Breast Cancer

Evolution of Diagnostic Technologies

Breast cancer diagnosis has evolved from solely clinical examination (palpation) to multimodal imaging and molecular diagnostics. Mammography, introduced in the 1960s and refined over decades, remains the primary screening tool with sensitivity 75-90% (varying by breast density) and specificity 90-95% (Pisano et al., 2005). Digital mammography and tomosynthesis (3D mammography) have improved performance, particularly for dense breasts.

Ultrasound complements mammography by distinguishing solid masses from cysts and characterizing suspicious lesions (BI-RADS classification). Magnetic Resonance Imaging (MRI) offers highest sensitivity (~90-100%) but lower specificity (~70-80%), reserved for high-risk screening and extent-of-disease assessment (Kuhl et al., 2005).

Biopsy and Histopathology

Despite imaging advances, tissue diagnosis remains the gold standard. Core needle biopsy has largely supplanted FNA in many centers due to architectural preservation enabling invasion assessment and receptor testing (Parker et al., 1990). However, FNA retains advantages in specific contexts: cyst aspiration, superficial masses, confirmation of recurrence, and resource-limited settings.

Pathological diagnosis involves assessment of multiple cellular and tissue features:

- **Nuclear features:** Size, shape, chromatin pattern, nucleoli prominence
- **Cellular features:** Cell size, cytoplasm characteristics, cell-to-cell adhesion
- **Architectural features:** Growth pattern, glandular formation, invasion
- **Molecular markers:** Estrogen receptor (ER), progesterone receptor (PR), HER2 amplification, Ki-67 proliferation index

The WDBC dataset captures nuclear features from digitized FNA images, representing the most objective component of cytological diagnosis—but omitting architectural and molecular information.

Inter-Observer Variability

Studies of diagnostic concordance reveal substantial inter-observer variability even among experienced pathologists. Meta-analysis by Elmore et al. (2015) found:

- **Invasive carcinoma:** High agreement (kappa 0.75-0.85)
- **Ductal carcinoma in situ (DCIS):** Moderate agreement (kappa 0.50-0.65)
- **Atypical hyperplasia:** Poor to moderate agreement (kappa 0.30-0.55)

FNA cytology shows similar or slightly lower concordance due to absence of architectural cues. This variability motivates development of computational aids to standardize interpretation.

Machine Learning in Cancer Detection

Historical Development

Application of machine learning to cancer diagnosis dates to the 1990s with rule-based expert systems and early neural networks (Wolberg & Mangasarian, 1990—the origin of the WDBC dataset itself). Early work focused on feature selection and simple classifiers (logistic regression, linear discriminant analysis, k-nearest neighbors).

The 2000s saw adoption of Support Vector Machines (SVMs), which became the dominant approach for medical classification given their strong performance on moderate-dimensional data and theoretical foundations in statistical learning theory (Vapnik, 1998). Numerous studies applied SVMs to WDBC, typically achieving 95-98% accuracy.

Deep Learning Era

Since ~2015, deep learning has revolutionized medical imaging analysis. Convolutional Neural Networks (CNNs) achieve radiologist-level performance on mammography interpretation (McKinney et al., 2020), skin lesion classification (Esteva et al., 2017