Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 12 March 2026
Revised 22 April 2026
Accepted 24 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 333–344

KEYWORDS:
breast cancer; cytology; machine learning; repeated cross-validation; calibration; permutation importance; decision curve analysis

Breast Cancer Classification from Cytomorphometric Features: A Leakage-Controlled Explainable Machine Learning Framework with Decision Curve Analysis

Ahmed A. F Osman

Applied College , King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia.
afadol@kfu.edu.sa , orcid.org/0009-0001-1362-4942

ABSTRACT:

Background: Certain metrics can be used to artificially inflate reported performance for medical machine learning models. This can happen, for example, with preprocessing leakage. Discrimination alone does not guarantee decision utility or probability reliability.

Objective: This study assesses a leakage-controlled, calibrated, and globally explainable machine-learning framework for breast cancer cytology classification.

Methods: The Wisconsin Diagnostic Breast Cancer dataset (569 observations; 212 malignant and 357 benign; 30 numerical predictors) was evaluated using Gradient Boosting, HistGradientBoosting, and Random Forest. Repeated stratified five-fold cross-validation with 10 repeats generated 50 outer test folds. Imputation, threshold selection, and sigmoid or isotonic recalibration were confined to training data. Performance was summarized from pooled patient-level out-of-fold predictions with 2,000 stratified bootstrap resamples. Held-out permutation importance quantified global predictor relevance, and decision curve analysis assessed potential net benefit.

Results: HistGradientBoosting achieved the highest pooled PR-AUC (0.9915; 95% CI, 0.9847–0.9969) and ROC-AUC (0.9933; 95% CI, 0.9874–0.9978), with sensitivity of 0.9670, specificity of 0.9664, Brier score of 0.0228, and ECE of 0.0135. Its PR-AUC exceeded Gradient Boosting (paired-bootstrap $p = 0.013$), whereas the ROC-AUC difference was not statistically supported ($p = 0.060$). HistGradientBoosting also outperformed Random Forest for ROC-AUC and PR-AUC (both $p = 0.020$). Worst area, worst concave points, worst texture, area error, and worst perimeter ranked highest by held-out permutation importance. Sigmoid-based decision curves showed higher net benefit than both default strategies across thresholds of 0.01–0.60 for Gradient Boosting and HistGradientBoosting.

Conclusion: Leakage-controlled repeated cross-validation produced strong internal benchmark performance and potential decision utility. These findings require external validation and prospective clinical-impact assessment before clinical use.

1. INTRODUCTION

Early and accurate medical diagnosis is paramount for improving patient survival rates, reducing healthcare costs, and optimizing therapeutic interventions across clinical specialties [1]. Studies in the fields of modern diagnostic oncology and pathology show that physical quantitative pathology measurements give complex sets of data describing cellular nuclear morphology. These data sets may describe dimensions such as size, texture, perimeter, and concavity. However, manual

interpretation by human pathologists of these multi-variate data sets is time consuming, has significant inter-observer variability, and is vulnerable to cognitive fatigue due to high throughput environments [2][3].

One growing response to these challenges has been the use of clinical decision support systems (CDSS) as built on AI and ML technologies. When trained on clinical, laboratory, and morphological health data, ML techniques can depict non-linear pathophysiological processes as well as provide initial risk assessment to clinicians for those patients of the highest concern [4][5].

The enormous promise these tools show in academic literature is unfulfilled due to three main methodological errors in the development of the models.:

1. Data Leakage in Evaluation Pipelines: Many published models perform global preprocessing—such as missing value imputation or feature scaling—across the entire dataset prior to cross-validation [6]. This leaks target distribution statistics into training folds, yielding artificially inflated performance metrics [7].
2. "Black-Box" Interpretability Barrier: Because opaque ensemble models don't provide clear explanations as to why a certain risk score was given, clinicians can't give an ethical account for the model outputs [8].
3. Misalignment Between Accuracy and Clinical Utility: Traditional metrics like ROC-AUC measure diagnostic discrimination but ignore probability calibration and net clinical benefit [9].

To overcome these shortcomings this study considers a globally explainable learning method which has leakage-controllability and a calibrated nature to classify fine-needle aspirate measures of breast cytology.

2. RELATED WORK

Traditionally,clinical risk scoring used linear logistic regression and score indexes [10]. However linear models are poor at handling the non-linear nature of the patient's underlying data despite their interpretability.

With the rise of ensemble modeling technologies, tree-based models such as XGBoost,LightGBM andCatBoosthave shown superior performance for tabular data medical data compared to deep neural networks [11]and[12]. The ability of tree ensembles to efficiently handle mixtures of numerical data and missingphysiological valuesat a low computation cost has led to widespread usage[13]but external cross-validation preprocessing transformations can be a main issue and a cause of reproducibility challenges in clinical AI [6], [7].

Because the "black-box" nature of models poses the "explainability" concern to be addressed in medical informatics. Featureattribution methods like SHapley Additive exPlanations (SHAP)[14]and permutationbased imports[15]have been recently used. The combination of featureattributionswithclinical expertise[16]provides evidence that AIs utilize pathophysiologically valid and important features which increase doctor confidence.

Moreover, in literature from the past five years, it is noted that the clinical usefulness of risk scoring models should not be judged only based on discriminativeperformance(e.g., ROC-AUC [9]). Assuch, both theExpected calibration error(ECE)[17],and Decision curve analysis(DCA)[18]have now been adopted as relevant quality measures to evaluate the calibration of probabilities and net clinical benefits.

Table I. Comparison of methodological features in clinical AI literature.

Literature Focus Area	Leakage-Free Preprocessing	Advanced Tree Ensembles	Model Calibration (ECE)	Clinician Explanations (XAI)	Decision Curve Analysis (DCA)
Traditional Risk Scores [10]	Yes	No	Partial	High	Rarely
Standard Clinical ML Studies [5]	Frequently Lacking	Yes	Rarely	Partial	Rarely
Tabular Deep Learning Papers [12]	Frequently Lacking	No	Rarely	Partial	Rarely
Proposed Framework	Strictly Enforced	Yes	Yes	Yes	Yes

Natural Resources for Human Health

eISSN: 2583-1194

3. MATERIALS AND METHODS

3.1 Dataset and Outcome Definition

The study used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset distributed through the UCI Machine Learning Repository and scikit-learn [2], [22]. The dataset contains 569 observations derived from digitized fine-needle aspirate images of breast masses, comprising 212 malignant (37.26%) and 357 benign cases. Thirty continuous nuclear morphometry predictors describe radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension as mean, standard-error, and worst-value summaries. The identifier and empty unnamed CSV column were excluded. Malignancy was encoded as the positive class (1), and benign status as 0.

The data audit identified no missing predictor values, no duplicate predictor rows, and no constant features. A distance-weighted five-neighbor KNN imputer was nevertheless retained inside every training pipeline to define a deployable missing-data pathway; because the benchmark dataset was complete, it performed no value replacement in this analysis. No feature scaling was applied because all evaluated estimators were tree based and scale invariant.

3.2 Leakage-Controlled Repeated Nested Validation

Internal validation used repeated stratified five-fold cross-validation with 10 repeats, producing 50 outer folds. In each outer fold, all preprocessing, threshold selection, model fitting, and recalibration operations were estimated using the outer-training partition only. The untouched outer-test partition was used once for performance estimation within each repeat. Each observation therefore contributed 10 held-out predictions per model, which were averaged to obtain pooled patient-level out-of-fold probabilities.

Classification thresholds were selected within each outer-training partition using four-fold stratified inner cross-validation. The threshold maximizing Youden's J was identified from inner out-of-fold probabilities and then applied to the corresponding outer-test predictions:

$$J(t) = Sensitivity(t) + Specificity(t) - 1.$$

Threshold-dependent predictions were aggregated across the 10 held-out appearances by majority vote. This design prevented information from outer-test observations from influencing preprocessing, threshold selection, or model fitting.

3.3 Evaluated Models and Fixed Hyperparameters

Three scikit-learn tree ensembles were evaluated. GradientBoostingClassifier used 150 estimators, learning rate 0.05, maximum tree depth 2, minimum leaf size 3, and subsampling rate 0.90. HistGradientBoostingClassifier used learning rate 0.06, 250 boosting iterations, 15 maximum leaf nodes, minimum leaf size 10, and L2 regularization 0.5. RandomForestClassifier used 600 trees, square-root feature sampling, minimum leaf size 1, and balanced class weights. Random seed 42 governed reproducible splitting and estimation. No data-driven hyperparameter search was performed; these settings were fixed before the reported evaluation.

For the two boosting estimators, binary log loss was minimized through stage-wise additive trees, represented generically as:

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} h_m(x),$$

where η is the learning rate and $h_m(x)$ is the tree added at iteration m. Random Forest estimated malignant probability by averaging class probabilities across B bootstrap trees:

$$\hat{p}(x) = (1/B) \sum_{b=1}^{B} \hat{p}_b(x).$$

3.4 Performance, Uncertainty, and Paired Comparisons

PR-AUC was the prespecified primary endpoint because malignant cases were the less prevalent class. Secondary measures were ROC-AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, balanced accuracy, Matthews correlation coefficient (MCC), Brier score, ECE, calibration intercept, and calibration slope. Pooled patient-level estimates were accompanied by percentile 95% confidence intervals from 2,000 stratified patient-level bootstrap resamples that preserved malignant and benign counts.

Pairwise differences in ROC-AUC, PR-AUC, and Brier score were calculated within the same stratified bootstrap resamples, preserving patient-level pairing between models. Two-sided empirical p-values were derived from the proportion of bootstrap differences on either side of zero. Values below 0.001 are reported as $p < 0.001$ rather than $p = 0$.

3.5 Calibration Assessment and Recalibration

Raw probability calibration was evaluated using Brier score, calibration intercept and slope, and ECE with 10 equal-width probability bins. ECE was calculated as:

$$ECE = \Sigma_{m=1}^{M} (|B_m|/N) \; |obs(B_m) - pred(B_m)|, \quad M = 10,$$

where $obs(B_m)$ and $pred(B_m)$ are the observed malignant proportion and mean predicted probability in bin B_m. Sigmoid and isotonic recalibration were also evaluated using three-fold CalibratedClassifierCV fitted entirely within each outer-training partition. Recalibration was treated as a comparative assessment and was not assumed to improve every model.

3.6 Held-Out Permutation Feature Importance

Global explainability was quantified using held-out permutation feature importance. For every fitted raw model and outer fold, one feature was randomly permuted in the untouched outer-test data 25 times while the remaining features were unchanged. Importance was the decrease in outer-test ROC-AUC relative to the unpermuted data. For feature j across F = 50 outer folds:

$$PFI_j = (1/F) \Sigma_{f=1}^{f} [AUC_f - AUC_{f,j}^{perm}].$$

Mean importance, between-fold standard deviation, and empirical 2.5th and 97.5th percentiles were reported. These percentiles describe fold-to-fold variability and were not interpreted as formal confidence intervals. The main importance table and figure correspond to the numerical pooled PR-AUC leader. A single permutation importance does measure the dependence of predictions, but neither its direction nor the causation, and collinear morphometric factors may compete with or divide their respective importance.

3.7 Decision Curve Analysis

Decision curve analysis used pooled patient-level out-of-fold probabilities over decision thresholds from 0.01 to 0.60 in increments of 0.01. The primary figure used leak-free sigmoid-recalibrated probabilities. Net benefit was computed as:

$$NB(p_t) = TP/N - FP/N \times p_t/(1 - p_t).$$

A model was considered potentially useful at a threshold only when its net benefit exceeded both Treat All and Treat None. DCA was interpreted as internal evidence of potential decision utility, not proof of prospective safety, biopsy reduction, or clinical effectiveness.

4. RESULTS

4.1 Dataset Audit and Validation Completeness

All 569 observations and 30 predictors were retained. The complete analysis comprised 50 outer folds, 150 model-specific inner thresholds, 17,070 outer-test prediction records, 1,707 pooled patient–model records, 66,000 bootstrap metric records, and 4,500 held-out fold–model–feature permutation records. Every patient had exactly 10 held-out predictions per model.

4.2 Discrimination and Overall Performance

Table II. Pooled patient-level out-of-fold discrimination and calibration with 2,000-bootstrap confidence intervals.

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Brier (95% CI)	ECE (95% CI)
Gradient Boosting	0.992 (0.985–0.997)	0.989 (0.982–0.995)	0.029 (0.020–0.038)	0.016 (0.015–0.034)
HistGradientBoosting	0.993 (0.987–0.998)	0.992 (0.985–0.997)	0.023 (0.014–0.033)	0.013 (0.011–0.030)
Random Forest	0.991 (0.982–0.997)	0.989 (0.981–0.996)	0.031 (0.023–0.040)	0.038 (0.030–0.053)

HistGradientBoosting was the numerical leader, with ROC-AUC 0.9933 (95% CI, 0.9874–0.9978) and PR-AUC 0.9915 (95% CI, 0.9847–0.9969). Gradient Boosting and Random Forest also showed high discrimination, but both had higher Brier scores than HistGradientBoosting. Figure 1 shows the pooled ROC curves and Figure 2 shows the corresponding precision–recall curves against the malignant-class prevalence baseline of 0.3726.

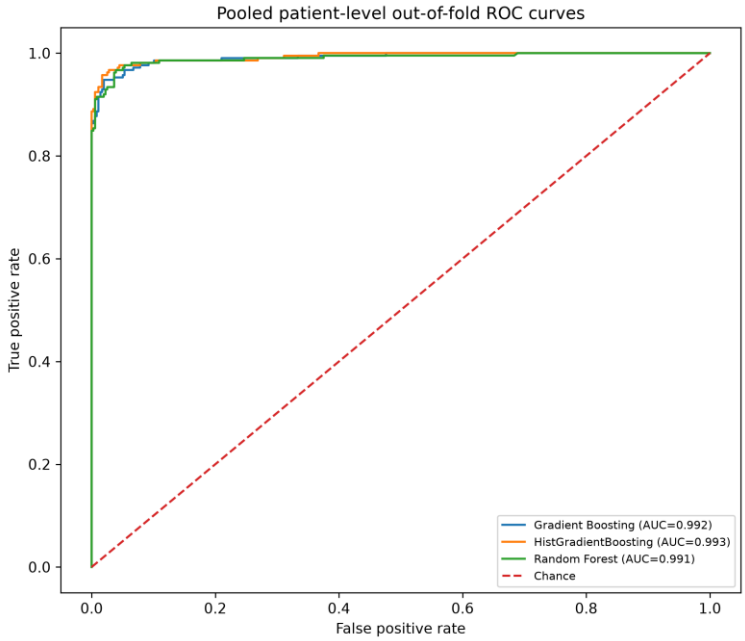

Figure 1. Pooled patient-level out-of-fold ROC curves. HistGradientBoosting achieved the highest numerical ROC-AUC; all curves are based on probabilities averaged across 10 held-out appearances per patient.

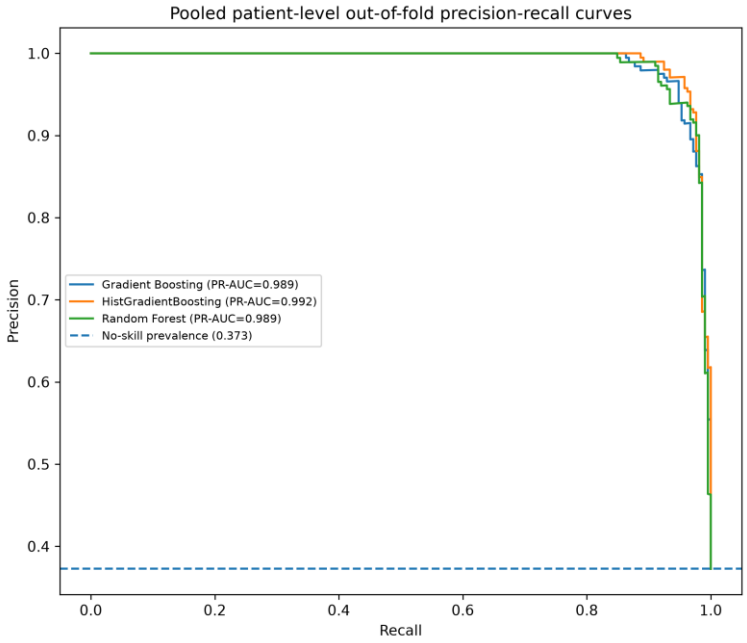

Figure 2. Pooled patient-level out-of-fold precision–recall curves. The dashed reference line represents malignant prevalence (0.3726).

4.3 Threshold-Dependent Diagnostic Performance

Table III. Threshold-dependent diagnostic performance using majority-voted nested training-side Youden thresholds.

Model	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Gradient Boosting	0.953 (0.925–0.981)	0.958 (0.936–0.978)	0.931 (0.899–0.962)	0.972 (0.954–0.988)
HistGradientBoosting	0.967 (0.939–0.991)	0.966 (0.947–0.983)	0.945 (0.915–0.972)	0.980 (0.964–0.994)
Random Forest	0.967 (0.943–0.991)	0.958 (0.936–0.978)	0.932 (0.899–0.962)	0.980 (0.966–0.994)

Table IV. Summary classification metrics and pooled confusion-matrix counts.

Model	F1 (95% CI)	Balanced accuracy (95% CI)	MCC (95% CI)	TN/FP/FN/TP
Gradient Boosting	0.942 (0.919–0.963)	0.955 (0.937–0.972)	0.907 (0.870–0.940)	342/15/10/202
HistGradientBoosting	0.956 (0.935–0.974)	0.967 (0.950–0.980)	0.929 (0.895–0.959)	345/12/7/205
Random Forest	0.949 (0.928–0.970)	0.962 (0.946–0.978)	0.918 (0.885–0.951)	342/15/7/205

HistGradientBoosting correctly classified 205 of 212 malignant observations and 345 of 357 benign observations, corresponding to sensitivity 0.9670 and specificity 0.9664. It achieved the highest F1-score (0.9557), balanced accuracy (0.9667), and MCC (0.9291). Random Forest matched its sensitivity but produced three additional false-positive classifications.

Table V. Paired patient-level bootstrap comparisons. Differences are first model minus second model; lower Brier scores are better.

Paired comparison	Metric	Difference	95% bootstrap interval	p-value
Gradient Boosting − HistGradientBoosting	ROC-AUC	-0.0017	-0.0040 to 0.0000	0.060
Gradient Boosting − HistGradientBoosting	PR-AUC	-0.0023	-0.0046 to -0.0005	0.013
Gradient Boosting − HistGradientBoosting	Brier	0.0058	0.0025 to 0.0093	<0.001
Gradient Boosting − Random Forest	ROC-AUC	0.0011	-0.0018 to 0.0054	0.610
Gradient Boosting − Random Forest	PR-AUC	0.0004	-0.0019 to 0.0032	0.749
Gradient Boosting − Random Forest	Brier	-0.0029	-0.0069 to 0.0012	0.168
HistGradientBoosting − Random Forest	ROC-AUC	0.0028	0.0002 to 0.0064	0.020
HistGradientBoosting − Random Forest	PR-AUC	0.0027	0.0004 to 0.0056	0.020
HistGradientBoosting − Random Forest	Brier	-0.0087	-0.0134 to -0.0040	0.001

HistGradientBoosting exceeded Gradient Boosting for PR-AUC (difference 0.0023, $p = 0.013$) and Brier score (difference −0.0058 when expressed as HistGradientBoosting minus Gradient Boosting, $p < 0.001$), while their ROC-AUC difference was not statistically supported ($p = 0.060$). HistGradientBoosting exceeded Random Forest for ROC-AUC and PR-AUC (both $p = 0.020$) and had a lower Brier score ($p = 0.001$). Gradient Boosting and Random Forest did not differ significantly for these endpoints.

4.4 Calibration and Recalibration

Table VI. Calibration assessment before and after leak-free sigmoid or isotonic recalibration.

Model	Probability	Brier	ECE	Intercept	Slope
Gradient Boosting	Raw	0.0286	0.0160	0.092	1.066
Gradient Boosting	Sigmoid	0.0287	0.0273	0.157	1.397
Gradient Boosting	Isotonic	0.0282	0.0142	-0.066	0.999
HistGradientBoosting	Raw	0.0228	0.0135	0.152	0.776
HistGradientBoosting	Sigmoid	0.0262	0.0314	0.131	1.433
HistGradientBoosting	Isotonic	0.0249	0.0103	-0.073	1.046
Random Forest	Raw	0.0315	0.0375	-0.201	1.498
Random Forest	Sigmoid	0.0291	0.0253	0.046	1.267
Random Forest	Isotonic	0.0293	0.0116	-0.125	0.904

HistGradientBoosting had the best raw Brier score (0.0228) and raw ECE (0.0135). Isotonic recalibration lowered its ECE to 0.0103 and produced a calibration slope of 1.046, but increased the Brier score to 0.0249. Sigmoid recalibration did not improve HistGradientBoosting calibration. For Random Forest, both recalibration methods reduced Brier score and ECE relative to raw probabilities. Accordingly, recalibration effects were model dependent rather than uniformly beneficial.

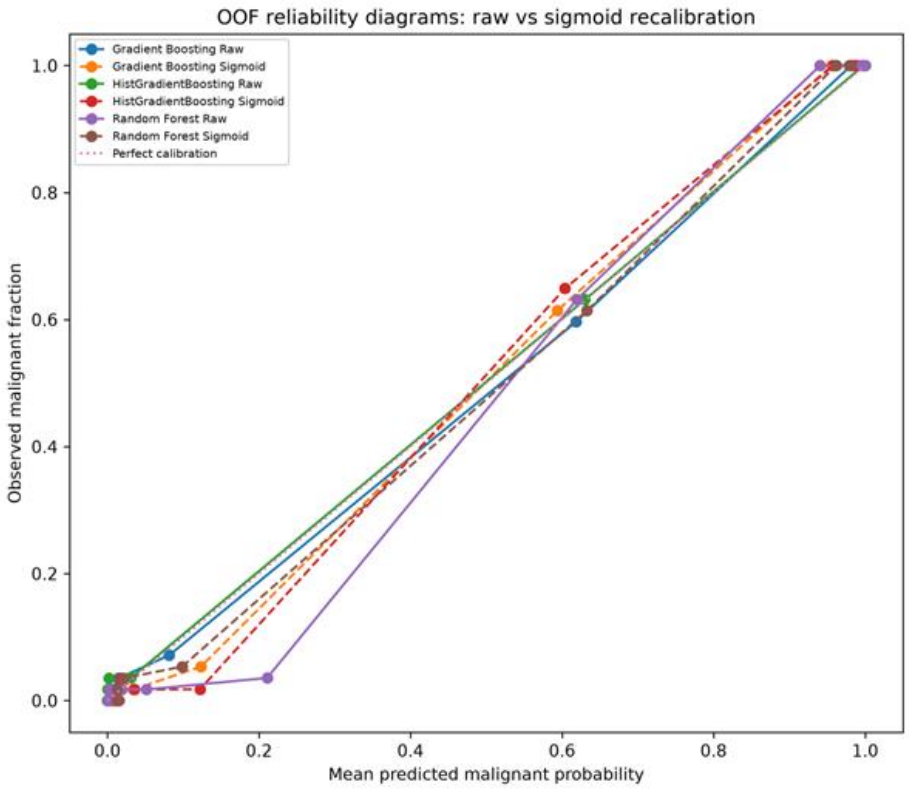

Figure 3. Out-of-fold reliability diagrams comparing raw and sigmoid-recalibrated probabilities. Quantile-based plotting groups are used for visualization; ECE was calculated with 10 equal-width bins.

Natural Resources for Human Health

eISSN: 2583-1194

4.5 Global Explainability

Table VII. Held-out permutation feature importance for HistGradientBoosting, the numerical pooled PR-AUC leader.

Rank	Feature	Mean ROC-AUC decrease	Empirical 2.5th–97.5th percentile
1	worst area	0.00655	0.00091 to 0.01932
2	worst concave points	0.00590	-0.00020 to 0.01629
3	worst texture	0.00523	0.00111 to 0.01279
4	area error	0.00321	-0.00064 to 0.00879
5	worst perimeter	0.00281	0.00005 to 0.00931
6	worst concavity	0.00204	-0.00071 to 0.00631
7	mean texture	0.00198	-0.00043 to 0.00648
8	mean concave points	0.00191	-0.00075 to 0.00604
9	worst radius	0.00065	-0.00055 to 0.00302
10	compactness error	0.00063	-0.00052 to 0.00263

Worst area, worst concave points, worst texture, area error, and worst perimeter were the five leading predictors by mean held-out ROC-AUC decrease. Worst area, worst texture, and worst perimeter had empirical 2.5th percentiles above zero, indicating more consistent importance across the 50 outer folds. The ranking emphasizes nuclear size, boundary irregularity, and texture variation, but permutation importance does not indicate whether larger values increase or decrease malignant probability and does not establish causal biological relevance.

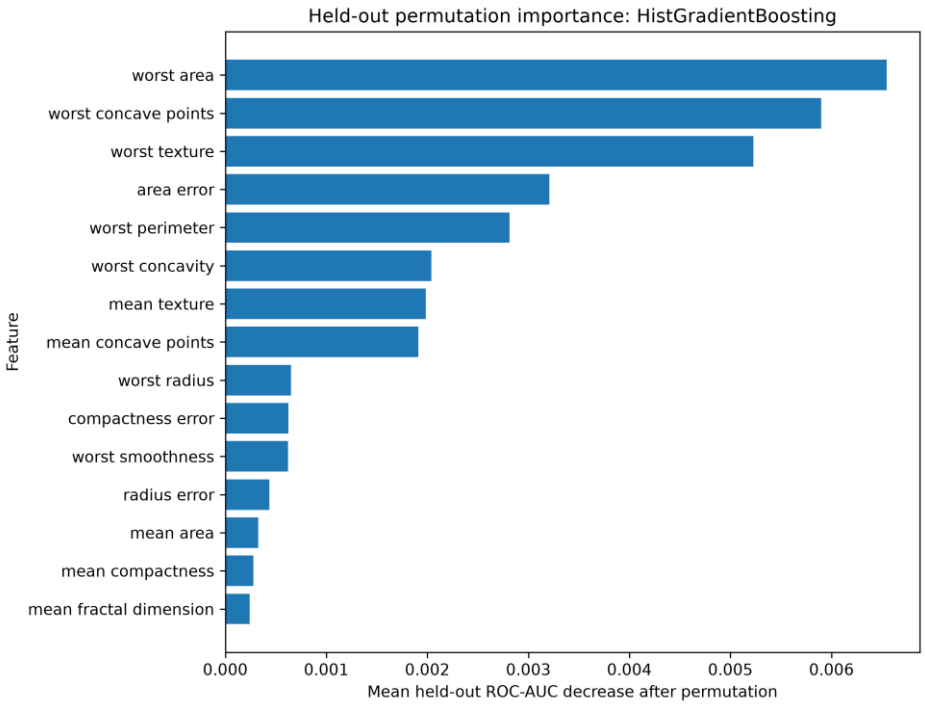

Figure 4. Held-out permutation importance for HistGradientBoosting. Values represent mean decreases in outer-test ROC-AUC after feature permutation; correlated predictors can redistribute importance.

Natural Resources for Human Health

eISSN: 2583-1194

4.6 Decision Curve Analysis

When utilizing leak-free sigmoid-recalibrated pooled out-of-fold probabilities, both Gradient Boosting and HistGradientBoosting outperform both Treat All and Treat None for all threshold levels evaluated (0.01 to 0.60). Random Forest out performs Treat All and Treat None for the threshold levels 0.02 to 0.60 and ties with treat all instead of out-performing Treat All at 0.01. At the 0.20 threshold, all model net benefits are approximately 0.35 while Treat All has a net benefit of approximately 0.216 and Treat None has a net benefit of 0. This suggest internal decision utility, but do not give quantitative measures of the reduction in number of biopsy or the safety thereof.

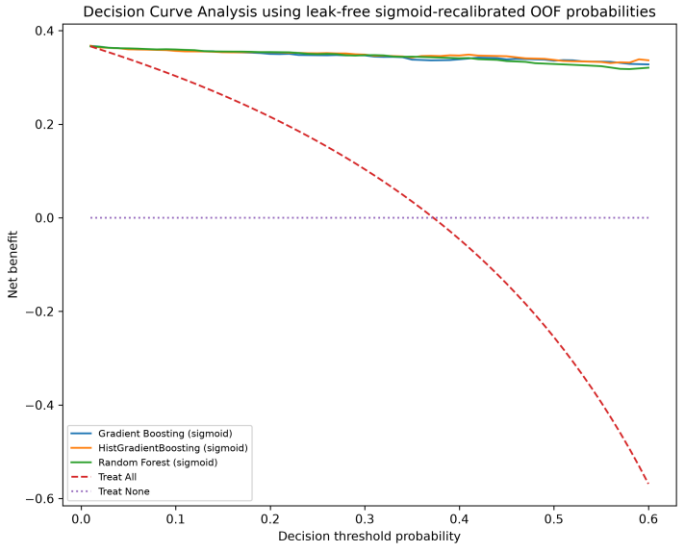

Figure 5. Decision curve analysis using leak-free sigmoid-recalibrated pooled out-of-fold probabilities. Net benefit is compared with Treat All and Treat None over thresholds from 0.01 to 0.60.

4.7 Confusion Matrices

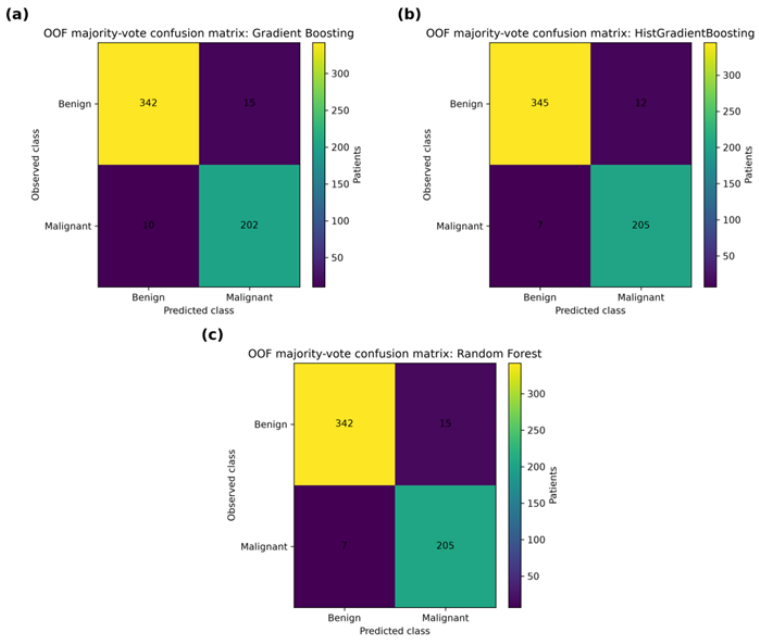

Figure 6. Pooled patient-level majority-vote confusion matrices: (a) Gradient Boosting, (b) HistGradientBoosting, and (c) Random Forest. Malignancy is the positive class.

5. DISCUSSION

5.1 Principal Findings

In this internally validated WDBC benchmark, all three tree ensembles achieved high discrimination, but HistGradientBoosting provided the strongest overall combination of PR-AUC, ROC-AUC, threshold-dependent performance, and probability accuracy. The paired analysis adds necessary nuance: its PR-AUC and Brier score were better than Gradient Boosting, whereas the ROC-AUC comparison narrowly crossed zero and did not meet the conventional 0.05 threshold. HistGradientBoosting also outperformed Random Forest for both area-under-curve endpoints and Brier score. These comparisons support selection of HistGradientBoosting as the numerical leader without implying that every difference was statistically established.

Calibration findings also argue against treating post-hoc recalibration as automatically beneficial. HistGradientBoosting's raw probabilities already had the best Brier score and ECE. Isotonic recalibration improved ECE and slope but worsened the Brier score, while sigmoid recalibration worsened both measures. This trade-off reinforces the need to report calibration methods and metrics separately rather than labeling a model as calibrated solely because a transformation was fitted.

5.2 Explainability and Morphometric Relevance

Held-out permutation importance focused on the worst region and worst concave region of points as well as worst texture, area error and worst perimeter feature values. These features can be said to represent nuclear size, non-convexity of the nuclear boundary, variation and texture within the nucleus-the same broad morphometric categories from which the WDBC diagnostic representation is derived [2], [22]. Since permutation only occurred on held-out outer-test folds, it is important to understand that estimations were separated from the training data. Also, the use of correlated predictors may lead to one feature replacing the other and the ranks given show where the most reliance occurs for prediction rather than indicating a direction of contribution, an individual patient's reason for prediction or the reason behind causal actions [15].

5.3 Potential Decision Utility

The decision curves demonstrated that all three models' predicted net benefits were superior to the default strategy over the majority of thresholds analyzed, with the two boosted models outperforming from 0.01-0.60. DCA accounts for the relative costs of true and false positive and false negative events thereby informing the results of the discrimination and calibration analysis [18]. The curves, however, are based on the pooled external out-of-fold predictions using an historical reference data set. They do not serve as definitive proof of bed-side readiness and reduced biopses.

5.4 Study Limitations

- The study used one small, historical, highly curated benchmark dataset. Repeated cross-validation improves internal precision but does not replace external, temporal, geographic, or prospective validation.
- The dataset contains derived cytology morphometry but lacks patient demographics, comorbidities, acquisition-site information, and other clinical variables needed to assess subgroup transportability and fairness.
- The KNN imputer was included in the leakage-controlled pipeline but was not empirically exercised because the dataset contained no missing values; its real-world performance therefore remains untested.
- Youden thresholds optimize sensitivity and specificity symmetrically within the training data and may not correspond to actual clinical costs or workflow-specific intervention thresholds.
- Permutation importance is global, model dependent, sensitive to correlated predictors, and non-causal. It does not provide local patient explanations or feature-effect direction.
- Calibration and decision utility were estimated internally. Prospective impact studies are required before claims about diagnostic workflow improvement, biopsy reduction, or clinical safety can be made.

6. CONCLUSION

A consistently used, leakage-controlled validation scheme found HistGradientBoosting to be the best classifier overall in the WDBC cytology benchmark. The model exhibited high pooled discrimination, balanced sensitivity and specificity, good raw probability accuracy, stable reliance on morphometric features, and positive internal net benefit. The research outlines an

evaluation pathway that can be repeated and includes uncertainty, calibration, global explainability, and decision analysis. Still, external validation and future assessment of its prospective clinical impact are necessary before it can be used clinically.

FUNDING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. **KFU264421**]

REFERENCES

[1] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," New England Journal of Medicine, vol. 363, no. 4, pp. 301–304, 2010.

[2] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from fine-needle aspirates," Archives of Surgery, vol. 130, no. 5, pp. 511–516, 1995.

[3] R. E. Steiner and L. J. King, "Cognitive errors and fatigue in diagnostic radiology," Academic Radiology, vol. 25, no. 7, pp. 885–893, 2018.

[4] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.

[5] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

[6] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023.

[7] D. B. Kaufman and S. Rosset, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, pp. 15:1–15:37, 2012.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[9] A. P. Dawid, "The well-calibrated Bayesian," Journal of the American Statistical Association, vol. 77, no. 379, pp. 605–610, 1982.

[10] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. New York, NY, USA: Springer, 2019.

[11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 3146–3154.

[12] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 507–520.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[14] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.

[15] A. Altmann et al., "Permutation importance: a corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340–1347, 2010.

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[18] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," Medical Decision Making, vol. 26, no. 6, pp. 565–574, 2006.

[19] O. Troyanskaya et al., "Missing value estimation algorithms for microarray data," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[22] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, doi: 10.24432/C5DW2B.