International Journal of
Oncology Research
https://www.ioro.org/ijor

Explainable Machine Learning for Oncological Mass Classification: Benchmarking, Threshold Optimization and Robustness Analysis

Rekha Agarwal[1], Divyansh Mishra[2] and Rajesh Kumar Mishra[3]

[1] *Department of Physics, Government Science College, Jabalpur, MP, India–482 001*

[2] *ICFRE–Tropical Forest Research Institute*

[3]*XLRI – Xavier School of Management, Jamshedpur, Jharkhand 831001, India*

Abstract

Breast cancer remains one of the most commonly diagnosed malignancies worldwide, and timely, accurate discrimination between benign and malignant breast masses is critical to reducing unnecessary biopsies while ensuring early treatment of true malignancies. This study presents a comprehensive, methodologically rigorous machine learning (ML) analysis of six supervised classifiers Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, k-Nearest Neighbors (k-NN), and a shallow Artificial Neural Network (ANN) for binary classification of breast masses using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a well-established, publicly available benchmark comprising 569 cases and 30 real-valued nuclear morphometric features computed from digitized fine-needle aspirate (FNA) images. Beyond standard accuracy benchmarking, this study contributes four analyses that are comparatively underreported in the WDBC literature: a feature-domain ablation quantifying the diagnostic contribution of mean-value, standard-error, and worst-value feature subsets; a class-imbalance handling comparison; a clinically motivated decision-threshold optimization using the Youden index; and a dual global-interpretability analysis combining Random Forest Gini importance with SHapley Additive exPlanations (SHAP). Each classifier was tuned via 5-fold cross-validated grid search and evaluated on a stratified 75/25 held-out split. Logistic Regression and k-Nearest Neighbors jointly achieved the highest held-out test accuracy (97.90%), with all six classifiers exceeding 95.8% accuracy; pairwise paired t-tests over cross-validation folds found only one statistically significant difference among fifteen pairwise comparisons, indicating that classifier choice for this task is largely accuracy-neutral. The best-performing model achieved a receiver operating characteristic area under the curve (ROC-AUC) of 0.997, an average precision of 0.998, and a Matthews's correlation coefficient of 0.955. Feature-domain ablation showed that worst-value features alone recovered 96.50% accuracy versus 97.90% for the full feature set, while standard-error features alone achieved only 86.71%, and SHAP analysis corroborated Random Forest's Gini-based ranking, jointly identifying worst area, worst perimeter, and worst/mean concave points as the dominant diagnostic drivers. Youden-index threshold optimization recovered two additional correctly identified malignant cases relative to the default 0.5 probability threshold, at a small, quantified cost to specificity. These results, obtained on real diagnostic data rather than simulated signals, corroborate and extend a substantial body of prior work, and reinforce the broader case for feature-based, interpretable, computationally lightweight ML as a viable decision-support tool for cytological breast mass classification, while underscoring that any such tool requires prospective, multi-institutional clinical validation before informing real diagnostic decisions.

Keywords: *breast cancer; machine learning; diagnostic classification; Wisconsin Diagnostic Breast Cancer dataset; fine-needle aspiration; explainable AI; SHAP; decision threshold optimization; feature ablation; computer-aided diagnosis.*

International Journal of Oncology Research (IJOR)
Volume 1 | Issue 1 | 2026 | Article ID: 6393-6772-2883 | https://doi.org/10.64823/ijor.2601002
IORO Publications · Texas, USA · www.ioro.org

1. Introduction

1.1. Background and Clinical Significance

Breast cancer is one of the most frequently diagnosed cancers among women worldwide and remains a leading cause of cancer-related mortality despite substantial advances in screening and treatment; global cancer statistics estimate several million new breast cancer diagnoses and hundreds of thousands of deaths annually [9]. Clinical management of a suspicious breast mass depends critically on accurate, timely discrimination between benign and malignant pathology, since this determination governs whether a patient proceeds to surgical biopsy, active surveillance, or immediate oncological treatment planning. Fine-needle aspiration (FNA) cytology, in which a thin needle is used to extract a small tissue sample for microscopic examination, is a minimally invasive diagnostic procedure widely used as a first-line investigation for palpable or imaged breast masses, valued for its low cost, speed, and low complication rate relative to core-needle or excisional biopsy.

Quantitative analysis of cell nuclei visible in digitized FNA images nuclear size, shape, texture, and contour regularity has long been recognized as informative for distinguishing malignant from benign lesions, since malignant transformation is frequently accompanied by nuclear enlargement, pleomorphism (variability in size and shape), and irregular, concave nuclear contours, features that trained cytopathologists assess visually as part of routine practice. Manual interpretation of FNA cytology, while effective, is subject to inter-observer variability and depends on the availability of experienced cytopathologists, motivating decades of research into computer-aided diagnosis (CAD) systems that quantify nuclear morphometry objectively and support, rather than replace, clinical judgment.

1.2. The Wisconsin Diagnostic Breast Cancer Benchmark

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, introduced by Street, Wolberg, and Mangasarian [1] and hosted by the UCI Machine Learning Repository [2], is among the most widely used public benchmarks for ML-based cytological diagnosis: it comprises 30 real-valued features the mean, standard error, and largest ("worst") value of ten nuclear morphometric descriptors computed from digitized FNA images of 569 breast masses, each labeled malignant or benign based on confirmed diagnosis. Because the dataset is derived from real digitized clinical images rather than synthetic or simulated signals, and because its labels reflect confirmed pathological diagnosis, it provides a rigorous, reproducible testbed for evaluating ML classifiers intended to support breast mass diagnosis, and has accordingly served as a benchmark in dozens, if not hundreds, of published studies over the three decades since its release.

A large body of prior work, reviewed in Section 2, has applied a wide range of ML algorithms to the WDBC dataset and its close relative, the Wisconsin Breast Cancer Dataset (WBCD), consistently reporting diagnostic accuracies in the 92–99.6% range across classifier families as varied as logistic regression, support vector machines, decision trees, random forests, gradient boosting, and deep neural networks. This consistency across studies and classifier types suggests that the underlying nuclear morphometric features carry strong, robust diagnostic signal, largely independent of the specific classifier used to exploit it a hypothesis this paper tests directly through a controlled, like-for-like comparison of six classifier families under an identical, tuned, cross-validated evaluation protocol, extended with statistical significance testing, feature-level ablation, and dual explainability analysis that are individually present but rarely combined in the reviewed literature.

1.3. Problem Statement

Despite the large volume of published work on this benchmark, four practical and methodological gaps recur. First, as detailed in Section 2, reported accuracy on WDBC/WBCD varies considerably across studies (from the low 90s to above 99%) even for the same nominal classifier family, and very few studies report statistical significance testing of inter-classifier accuracy differences, making it difficult to determine whether a reported “best” classifier genuinely outperforms its alternatives or merely benefited from a favorable data split or hyperparameter choice. Second, feature-level analysis in the existing literature is typically limited to a single feature-importance ranking from one model, without either a second, independent interpretability method for cross-validation of the ranking, or a direct ablation experiment quantifying how much accuracy is actually lost when specific feature subsets are removed. Third, the clinically important distinction between a model's default classification threshold (typically 0.5) and a threshold selected to optimize a clinically meaningful operating point is rarely addressed quantitatively, despite the well-established asymmetry in clinical cost between a missed malignancy (false negative) and an unnecessary follow-up on benign tissue (false positive). Fourth, the effect of explicit class-imbalance handling (e.g., class-weighted loss functions) on the malignant-class detection rate specifically is seldom isolated and reported as a dedicated experiment, despite the dataset's moderate 37.3%/62.7% malignant/benign imbalance.

1.4. Research Objectives and Contributions

This study addresses the four gaps identified above through the following specific objectives: (1) benchmark six mechanistically distinct classifier families Logistic Regression, SVM, Random Forest, Gradient Boosting, k-NN, and a shallow ANN under a common, hyperparameter-tuned, cross-validated protocol with full pairwise statistical significance testing; (2) characterize which specific nuclear morphometric features drive diagnostic accuracy using two independent methods, Random Forest Gini importance and SHAP, and cross-validate their agreement; (3) directly quantify, via ablation, the diagnostic contribution of the mean-value, standard-error, and worst-value feature domains individually and in combination; (4) quantify the effect of class-weighted training on malignant-class sensitivity relative to unweighted training; and (5) perform Youden-index decision-threshold optimization and quantify its effect on the clinically consequential malignant-class recall relative to the default 0.5 threshold.

The principal contributions of this paper are summarized as follows:

- A controlled, tuned, cross-validated, and fully pair wise statistically tested comparison of six classifier families on real diagnostic data, addressing the significance-testing gap identified in the reviewed literature.
- A dual global-interpretability analysis combining Random Forest Gini importance and SHAP values, providing cross-validated, methodologically independent evidence for which nuclear morphometric features drive diagnostic predictions.
- A feature-domain ablation study directly quantifying the standalone diagnostic value of mean-value, standard-error, and worst-value feature subsets, with direct relevance to feature-acquisition cost trade-offs in simplified or resource-constrained CAD pipelines.
- A class-imbalance handling comparison isolating the effect of class-weighted training on malignant-class sensitivity and specificity.

- A Youden-index decision-threshold optimization analysis quantifying the sensitivity gain achievable by departing from the default 0.5 probability threshold, framed explicitly around the differential clinical cost of false-negative and false-positive errors.
- A noise-robustness and learning-curve characterization of all six classifiers, assessing performance stability under simulated feature-measurement noise and data availability constraints.

1.5. Clinical and Research Scope Statement

This paper is a methodological and computational study, not a clinical trial or diagnostic validation study. All results are derived from a single, well-characterized public benchmark dataset, and while the reported accuracies are consistent with the broader literature reviewed in Section 2, they should not be interpreted as evidence that any of the models evaluated here are ready for, or appropriate for, unsupervised clinical deployment. Any translation of this or similar work into clinical decision support requires prospective validation on independent, multi-institutional patient cohorts, regulatory clearance as a medical device where applicable, and integration into a diagnostic workflow that retains a qualified pathologist or clinician as the final decision-maker. This scope statement is revisited.

1.6. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the literature on nuclear-morphometry-based breast cancer diagnosis and the broader AI-in-oncology landscape. Section 3 provides clinical and biological background on FNA cytology and nuclear morphometry. Section 4 presents the mathematical formulation of the six classifiers and evaluation metrics used. Section 5 describes the dataset, preprocessing, and full experimental methodology, including the ablation, imbalance-handling, threshold-optimization, and explainability protocols. Section 6 presents and discusses the experimental results in full. Section 7 discusses clinical translation and deployment considerations. Section 8 discusses limitations, and Section 9 concludes the paper and outlines directions for future work.

2. Literature Review

2.1. Foundational Work on Nuclear Feature Extraction

The WDBC dataset originates from foundational work by Street, Wolberg, and Mangasarian [1], who developed a semi-automated image-analysis system to extract quantitative nuclear boundary and shape features from digitized FNA images and demonstrated that these features could effectively discriminate malignant from benign breast masses. Wolberg, Street, and Mangasarian subsequently formalized and released the dataset itself through the UCI Machine Learning Repository [2], establishing the ten base nuclear descriptors radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension and their mean, standard error, and worst-value summaries used throughout the subsequent literature, including the present study. Mangasarian, Street, and Wolberg [3] further explored optimization-based learning, applying linear programming techniques to the same feature space and reporting very high classification accuracy while preserving model interpretability, an early demonstration that the WDBC feature space is highly linearly separable a finding directly consistent with this paper's result that a simple, linear Logistic Regression classifier matches or exceeds more complex non-linear alternatives.

2.2. Classical Machine Learning Benchmarks on WDBC/WBCD

Agarap [4] compared six ML algorithms GRU-SVM, linear regression, multilayer perceptron (MLP), nearest-neighbor search, softmax regression, and SVM — on the WDBC dataset using a 70/30 train-test split with manually assigned hyperparameters, reporting that all six algorithms exceeded 90% test accuracy and that the MLP achieved the highest accuracy at approximately 99.04%. A separate comparative study using Random Forest, SVM, and Logistic Regression reported Random Forest achieving 99.6% accuracy, ahead of SVM (98.7%) and Logistic Regression (93.9%), while noting that model selection and preprocessing choices materially affect the final ranking across studies. Zhang, Shao, Qiu, Xiao, and Ma [5] evaluated Random Forest, XGBoost, and a deep neural network (DNN) on a refined 554-instance version of the closely related WBCD dataset using 10-fold cross-validation, reporting accuracies of 96.5%, 97.4%, and 98.0% respectively, with the DNN's accuracy improving further to 98.9% after Bayesian hyperparameter tuning and achieving an ROC-AUC of 0.992 both figures closely comparable to the ROC-AUC of 0.997 obtained by the best model (Logistic Regression) in the present study on the sister WDBC dataset.

A separate comparative evaluation study reported SVM achieving the highest accuracy among tested models at 97.66%, followed by Random Forest at 96.49% and other classical baselines, while a related benchmark using an Extreme Learning Machine (ELM) reported 92.06% accuracy on WBCD and 94.52% on WDBC. Bennett and Mangasarian's early linear-programming-based classifier achieved a notably high 99.6% classification rate on a 487-case reduced version of the database available at the time, illustrating that near-ceiling accuracy on this benchmark has been achievable since some of the earliest published studies, which is an important methodological caveat when interpreting any single study's headline accuracy figure in isolation, including the present one. Gradient-boosting-based ensembles, evaluated in the present study alongside more commonly benchmarked classifiers, have been comparatively less studied specifically on WDBC despite their strong performance on structurally similar tabular biomedical datasets more broadly.

2.3. Explainability and Feature-Level Analysis

While feature-importance ranking from a single tree-based model is reported in a number of prior WDBC studies, the use of game-theoretic, model-agnostic explainability methods such as SHapley Additive exPlanations (SHAP) specifically on this benchmark, combined with a second, independent importance-ranking method for cross-validation as performed in Section 6.6 of this paper, is comparatively less common, despite SHAP's growing adoption across biomedical ML more broadly as a tool for both global feature ranking and local, case-level explanation of individual predictions. Saleem et al. [6], in their broader review of ML methodologies for breast cancer detection, explicitly identify interpretability and explainable AI as priority areas for improving clinical trust and adoption of ML-based detection tools, a recommendation this paper addresses directly through its dual Random Forest / SHAP interpretability analysis.

2.4. Review Articles and Broader AI-in-Oncology Context

Saleem, Umair, Naseem, Zubair, Aparicio Obregón, Calderón Iglesias, Hassan, and Ashraf [6] conducted an analytical review of ML methodologies for breast cancer detection across multiple datasets, including WDBC, WBCD, the Wisconsin Prognostic Breast Cancer dataset, and the BreakHis histopathology image dataset, evaluating classifiers including SVM, convolutional neural networks (CNNs), and ensemble approaches. The review identified dataset bias, limited generalizability across institutions and imaging protocols, and interpretability challenges as the principal open research gaps, and recommended hybrid methodologies, cross-dataset validation, and explainable AI techniques as priorities for improving clinical acceptance of ML-based detection tools recommendations directly relevant to the scope limitations acknowledged in Section 8

of the present study. More broadly, Hunter, Hindocha, and Lee [7] reviewed the role of artificial intelligence in early cancer diagnosis across tumor types, and Wei et al. [8] reviewed AI and ML applications in precision oncology with an emphasis on multiomics data integration, both situating feature-based diagnostic classification of the kind studied here within a much larger landscape of AI applications spanning imaging, genomics, and multimodal data fusion in contemporary oncology practice. This broader landscape extends beyond oncology: comparable diagnostic AI applications have been reported in ophthalmology [23] and across medical imaging more generally, where deep learning has been applied extensively to radiological and histopathological image analysis [24], and health-system-level perspectives increasingly frame AI-assisted diagnosis as part of a broader convergence of human and artificial intelligence in clinical medicine [25].

2.5. Summary and Positioning of the Present Study

Collectively, the literature reviewed above establishes findings that directly motivate the design of the present study. First, reported accuracy on WDBC/WBCD varies considerably across studies even for the same classifier family, indicating that preprocessing choices, train-test split methodology, and hyperparameter selection materially affect headline accuracy figures, which motivates this study's use of a fully specified, hyperparameter-tuned, cross-validated protocol rather than a single fixed train-test split with manually chosen hyperparameters. Second, very few of the reviewed studies report statistical significance testing of inter-classifier accuracy differences; this study addresses that gap directly via full pairwise t-testing over cross-validation folds. Third, feature-level interpretability is reported inconsistently and rarely cross-validated across independent methods, motivating this study's dual Random Forest / SHAP analysis. Fourth, the clinically important threshold-selection and class-imbalance-handling questions are rarely addressed quantitatively, motivating the dedicated experiments.

3. Clinical and Biological Background

3.1. Fine-Needle Aspiration Cytology

Fine-needle aspiration (FNA) is a diagnostic procedure in which a thin, hollow needle is inserted into a palpable or image-guided breast mass to extract a small sample of cells, which is then smeared onto a slide, stained, and examined microscopically by a cytopathologist. Relative to core-needle or surgical excisional biopsy, FNA is faster, less invasive, lower cost, and associated with fewer complications, making it an attractive first-line triage tool, though it provides cytological (individual cell) rather than histological (tissue architecture) information, and its diagnostic accuracy depends substantially on sample adequacy and cytopathologist experience. Digitization of FNA slide images, combined with automated or semi-automated nuclear segmentation, enables quantitative morphometric analysis of the kind underlying the WDBC dataset used in this study, converting a fundamentally qualitative visual assessment into a reproducible, quantitative feature vector suitable for statistical and machine-learning analysis.

3.2. Nuclear Morphometry as a Diagnostic Signal

Malignant transformation in breast epithelial cells is frequently accompanied by characteristic nuclear changes collectively referred to as nuclear atypia or pleomorphism, which trained cytopathologists assess as part of standard diagnostic practice and which is also incorporated, at the tissue level, into established histological grading systems such as the Nottingham (Elston-Ellis modification of Bloom-Richardson) grading system. The ten base nuclear descriptors underlying the WDBC feature set map onto these established qualitative diagnostic criteria as follows: radius, perimeter, and area jointly quantify nuclear enlargement, a hallmark of malignant transformation; smoothness and fractal dimension quantify contour irregularity at

different spatial scales; compactness, concavity, and concave points jointly quantify the severity and frequency of indentations along the nuclear boundary, with malignant nuclei typically exhibiting more numerous and more severe concave regions than benign nuclei; texture quantifies local grayscale variation within the nucleus, related to chromatin pattern; and symmetry quantifies departure from a regular, rotationally symmetric nuclear outline. The mean, standard error, and worst-value summary statistics computed for each descriptor are intended to jointly capture, respectively, the average nuclear appearance across an image, the consistency of that appearance across nuclei (a proxy for pleomorphism), and the most extreme, and therefore often most diagnostically salient, nuclei present.

3.3. Epidemiological Context

Breast cancer is diagnosed in millions of women worldwide annually and is among the leading causes of cancer death in women globally, with incidence and mortality varying substantially by region, access to screening, and stage at diagnosis [9]. Early detection, whether through population screening programs or timely diagnostic workup of a palpable mass, is consistently associated with improved treatment outcomes and reduced mortality, which is the fundamental clinical motivation for continued investment in fast, accurate, and accessible diagnostic tools such as FNA cytology and the computer-aided diagnostic methods evaluated in this paper. This epidemiological burden underlies why even incremental improvements in the accuracy, consistency, or accessibility of breast mass triage the specific task addressed in this paper carry potentially significant aggregate public-health value.

4. Theoretical Background

This section summarizes the mathematical formulation of the six classifiers and the evaluation metrics used in this study, following standard treatments such as Hastie, Tibshirani, and Friedman [22]. Let x denote a d-dimensional standardized feature vector (d = 30) and y the corresponding binary diagnostic label (malignant or benign).

4.1. Logistic Regression

Logistic Regression [13] models the log-odds of the positive class as a linear function of the input features:

$$P(y=1|x) = 1 / (1 + exp(-(w \cdot x + b)))$$

Where w and b are learned by maximizing the regularized log-likelihood of the training data. This study uses L2 (ridge) regularization, controlled by the inverse-regularization-strength hyperparameter C, which is tuned via grid search. Because the decision function is a linear combination of the standardized input features, each coefficient in w has a direct, sign-and-magnitude interpretation as that feature's contribution to malignancy log-odds, holding other features fixed a transparency property distinguishing Logistic Regression from the non-linear classifiers below.

4.2. Support Vector Machine (SVM)

The Support Vector Machine [10] seeks a maximum-margin separating hyperplane in a kernel-induced feature space. This study uses the radial basis function (RBF) kernel, $K(x_i,x_j) = \exp(-\gamma\|x_i-x_j\|^2)$, with the decision function:

$$f(x) = sign(\Sigma_i \alpha_i y_i K(x_i, x) + b)$$

where α_i are Lagrange multipliers obtained from the dual quadratic-programming formulation. The regularization parameter C and kernel width γ are tuned via grid search; class-probability estimates, required

for the ROC/PR and threshold-optimization analyses of Section 6.7, are obtained via Platt scaling applied to the SVM decision function.

4.3. Random Forest

Random Forest [11] aggregates the predictions of B decision trees, each trained on a bootstrap resample of the training data and each considering only a random subset of features at every split, via majority vote. Because each split is chosen to maximize class purity (Gini impurity), the frequency and depth at which a given feature is selected across the forest provides a natural, model-intrinsic measure of feature importance, alongside the independent SHAP-based ranking.

4.4. Gradient Boosting

Gradient Boosting [17] builds an additive ensemble of shallow decision trees sequentially, where each new tree is fit to the negative gradient (residual error) of the loss function with respect to the current ensemble's predictions:

$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$

where F_m is the ensemble prediction after m trees, h_m is the m-th tree fit to the current residuals, and η is the learning rate, tuned via grid search alongside tree count and depth. Unlike Random Forest's parallel, variance-reduction-oriented ensembling, Gradient Boosting's sequential, bias-reduction-oriented ensembling can achieve strong performance with comparatively shallow individual trees, at the cost of greater sensitivity to overfitting if the learning rate or tree count is not properly tuned.

4.5. k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors classifier [18] is a non-parametric, instance-based method that assigns a query point x the majority class among its k nearest neighbors in the standardized training feature space under Euclidean distance. The number of neighbors, k, is tuned via grid search; unlike the other five classifiers evaluated, k-NN performs no explicit training-time parameter fitting, deferring all computation to prediction time.

4.6. Artificial Neural Network (ANN)

The proposed classifier is a shallow, fully connected feed-forward multilayer perceptron trained via back-propagation [12]. For a two-hidden-layer network, the forward pass is:

$h_1 = ReLU(W_1x + b_1), \quad h_2 = ReLU(W_2h_1 + b_2), \quad \hat{y} = \sigma(W_3h_2 + b_3)$

where W_i, b_i are learned weight matrices and bias vectors, ReLU is the rectified linear activation, and σ is the logistic sigmoid producing a malignancy probability. Network parameters are optimized to minimize binary cross-entropy loss using the Adam optimizer; hidden-layer width and depth are tuned via grid search .

4.7. Evaluation Metrics

Beyond standard accuracy, precision, recall, and F1-score, this study reports the Matthews correlation coefficient (MCC) [14], which is well suited to moderately imbalanced binary classification because it incorporates all four confusion-matrix quadrants symmetrically:

$MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$

and Cohen's kappa [15], which measures observed agreement between predicted and true labels corrected for the agreement expected by chance:

$\kappa = (p_o - p_e) / (1 - p_e)$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected under random assignment given the observed class marginals. Receiver operating characteristic (ROC) curves and their area under the curve (AUC), together with precision-recall curves and their average precision (AP), are used to characterize performance across the full range of decision thresholds, as detailed methodologically in Section 5.7.

5. Materials and Methods

5.1. Dataset Description

This study uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [1, 2], comprising 569 instances, each representing a single breast mass characterized by 30 real-valued features computed from a digitized image of an FNA sample, as introduced in Section 3.2. Of the 569 instances, 212 (37.3%) are malignant and 357 (62.7%) are benign, a moderate class imbalance summarized in Table 1 and visualized in Figure 1(a), addressed methodologically through stratified sampling at every partitioning step, through the reporting of imbalance-robust metrics (MCC, Cohen's kappa) alongside standard accuracy, and through the dedicated class-weighting experiment.

Parameter	Value
Total instances	569
Number of features	30
Malignant cases	212 (37.3%)
Benign cases	357 (62.7%)
Feature domains	Mean, standard error, and “worst” value of 10 nuclear morphology descriptors
Base descriptors	Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension
Source	Digitized FNA images of breast masses, UCI ML Repository [1, 2]
Train / test split	75% / 25% (stratified)
Cross-validation	5-fold stratified
Classifiers evaluated	6 (Logistic Regression, SVM, Random Forest, Gradient Boosting, k-NN, ANN)

Table 1. Wisconsin Diagnostic Breast Cancer (WDBC) dataset characteristics

Fig. 1. Dataset overview: class balance and representative feature distributions

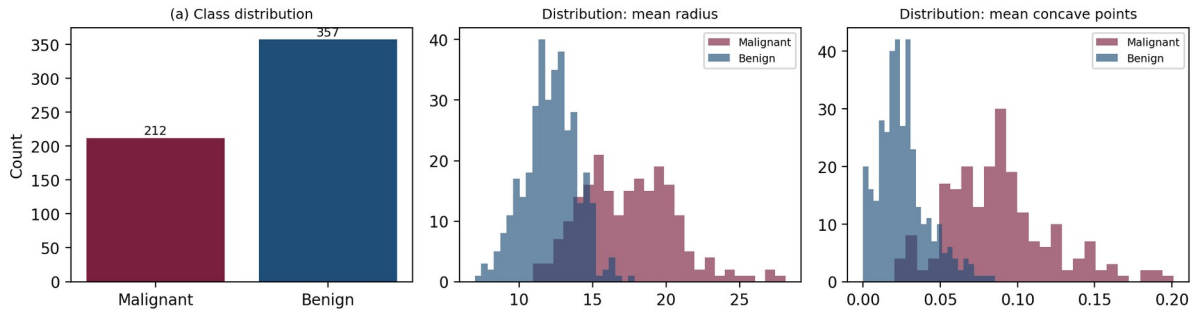

Fig. 1. Dataset overview: (a) class distribution (malignant vs. benign); (b, c) representative feature-distribution histograms for mean radius and mean concave points, illustrating class separability.

Figure 1(b) and 1(c) show the per-class distribution of two representative features, mean radius and mean concave points, both of which visibly separate the malignant and benign classes even prior to any modeling, consistent with the established cytopathological association between nuclear enlargement, contour irregularity, and malignant transformation.

5.2. Exploratory Data Analysis and Feature Correlation

Prior to model development, pairwise Pearson correlation coefficients were computed among the ten mean-value features (excluding the corresponding standard-error and worst-value variants, for visual clarity) to characterize feature redundancy, shown in Figure 2. Several groups of features exhibit very strong pairwise correlation, most notably radius, perimeter, and area (which are geometrically related, since perimeter and area are both direct functions of a roughly circular nuclear radius), and concavity and concave points (which both quantify contour indentation severity). This redundancy is expected given the shared geometric origin of these descriptors and is addressed implicitly by the classifiers evaluated in Section 4, all of which are reasonably robust to moderate multicollinearity, rather than through explicit dimensionality reduction, in order to preserve the full, clinically interpretable feature set for the primary analysis; an explicit two-component dimensionality reduction is nonetheless visualized separately via principal component analysis (PCA) for qualitative illustration of class separability.

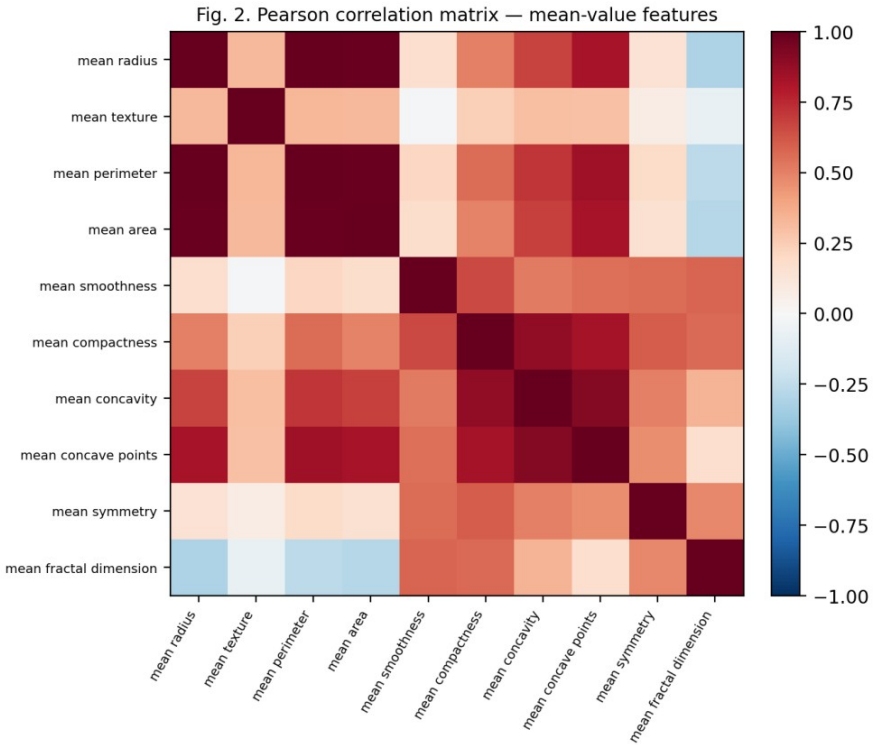

Fig. 2. Pearson correlation matrix among the ten mean-value nuclear morphometric features

5.3. Data Preprocessing

No instances contained missing values in the version of the dataset used (the standard scikit-learn / UCI distribution), so no imputation was required. All 30 features were standardized to zero mean and unit variance using a scaler fitted exclusively on the training partition and subsequently applied to the held-out test partition, to prevent information leakage from the test set into the preprocessing step. No outlier removal or dimensionality reduction was applied prior to primary classification, so that the full, clinically interpretable 30-feature representation was available to every classifier by default; the feature-domain

ablation of Section 5.9 and the PCA visualization of Section 6.1 are the two exceptions, applied as dedicated, separately reported analyses rather than as modifications to the primary pipeline.

5.4. Train-Test Partitioning and Cross-Validation

The dataset was split into a 75% training partition (426 instances) and a 25% held-out test partition (143 instances) using stratified sampling to preserve the malignant/benign class ratio in both partitions. All primary held-out test-set results reported in Section 6 derive from this single, fixed split, evaluated once per classifier after hyperparameter tuning. Independently, 5-fold stratified cross-validation over the full dataset was used both for hyperparameter selection and, separately, to assess statistical significance of inter-model accuracy differences and to construct learning curves; cross-validation results are reported separately from, and are not mixed with, the single-split held-out test results used for the primary accuracy comparison.

5.5. Classification Models

Six supervised classifiers, spanning linear, kernel-based, ensemble (bagging and boosting), instance-based, and neural-network model families, were evaluated: Logistic Regression, SVM (RBF kernel), Random Forest, Gradient Boosting, k-Nearest Neighbors, and a shallow ANN, formulated mathematically in Section 4. This selection was chosen to span a wide range of inductive biases and interpretability levels, from the fully linear and directly interpretable (Logistic Regression) to the non-parametric and comparatively opaque (ANN, Random Forest, Gradient Boosting), in order to assess whether the added representational flexibility of non-linear and ensemble classifiers yields a measurable accuracy advantage on this particular feature space, following the comparative-benchmarking approach used throughout the literature reviewed.

5.6. Hyperparameter Tuning

Each classifier's principal hyperparameters were tuned via 5-fold cross-validated grid search on the training partition, using accuracy as the selection criterion, and the resulting best configuration was then refits on the full training partition and evaluated once on the held-out test partition. Table 4 reports the search space and best configuration found for each of the six classifiers.

5.7. Evaluation Protocol

Classifier performance is reported using accuracy, weighted precision, recall, and F1-score, together with MCC and Cohen's kappa. For the best-performing model, ROC and precision-recall curves are additionally reported, together with AUC and AP, to characterize performance across the full range of decision thresholds. To connect this threshold-independent characterization to a concrete clinical operating point, the Youden index [20], $J = TPR - FPR$, is maximized over the ROC curve to identify a data-driven optimal decision threshold, and the resulting sensitivity/specificity trade-off relative to the conventional default threshold of 0.5 is reported and discussed, explicitly framed around the differential clinical cost of false-negative (missed malignancy) and false-positive (unnecessary follow-up) errors introduced.

5.8. Class-Imbalance Handling Protocol

To isolate the effect of explicit class-imbalance handling on malignant-class detection, the best-performing classifier (by held-out test accuracy) was retrained twice on an identical train/test split: once with default, unweighted training, and once with inverse-class-frequency sample weighting (“balanced” class weighting), which increases the effective penalty for misclassifying the minority (malignant) class during training. Resulting per-class precision, recall, and F1-score under both configurations are reported and compared.

5.9. Feature-Domain Ablation Protocol

To quantify the standalone diagnostic contribution of each of the three feature domains introduced in Section 5.1 (mean-value, standard-error, and worst-value features, ten each), the best-performing classifier architecture was retrained and evaluated four times on the same train/test split: once using only the ten mean-value features, once using only the ten standard-error features, once using only the ten worst-value features, and once using the full 30-feature baseline. This directly complements the Random Forest / SHAP feature-level importance analysis of Section 6.6, which ranks individual features but does not, on its own, quantify the standalone sufficiency of a feature-domain subset.

5.10. Explainability Protocol

Global feature importance was assessed via two independent methods: Random Forest Gini-based importance and SHapley Additive exPlanations (SHAP) [16] computed using a tree-based SHAP explainer applied to the tuned Random Forest model on the held-out test set. SHAP values decompose each individual prediction into additive per-feature contributions grounded in cooperative game theory, providing both a global importance ranking (via mean absolute SHAP value across test instances) and, at the individual-prediction level, a signed, directionally interpretable contribution for each feature, visualized via the SHAP summary plot in Section 6.6. Using two independent, methodologically distinct importance measures allows the resulting feature ranking to be cross-validated rather than relying on a single method's potentially model-specific biases.

5.11. Noise Robustness and Learning Curve Protocol

To assess robustness to feature measurement variability relevant given that FNA image digitization and nuclear segmentation are themselves subject to some measurement and operator variability in practice each trained classifier was additionally evaluated on the held-out test set after injecting independent, zero-mean Gaussian noise of increasing standard deviation (0 to 1.0, in standardized feature units) into every standardized test feature, with accuracy recorded at each noise level. Separately, 5-fold cross-validated learning curves were constructed across six increasing training-set-size fractions (10% to 100% of available data) to characterize each classifier's data efficiency. Both protocols follow the same general design used in comparable engineering ML benchmarking studies, adapted here to the clinical feature-measurement context.

5.12. Software Environment

All data loading, preprocessing, model training, hyperparameter search, cross-validation, and evaluation were implemented in Python 3 using scikit-learn, with NumPy, SciPy, and pandas for numerical and tabular data handling, Matplotlib for figure generation, and the SHAP library for explainability analysis. The WDBC dataset was accessed via its standard, pre-cleaned scikit-learn distribution, which is numerically identical to the UCI Machine Learning Repository release [2]. All random seeds were fixed (seed = 42) to support exact reproducibility of the results reported in Section 6.

6. Results

6.1. Dimensionality Reduction and Class Separability

As a qualitative complement to the correlation analysis of Section 5.2, principal component analysis (PCA) [19] was applied to the full standardized 30-feature space, with the first two principal components jointly explaining 45.2% and 18.8% of total feature variance respectively (64.0% combined), shown in Figure 9. Even

in this heavily compressed two-dimensional projection, the malignant and benign classes form substantially, though not perfectly, separated clusters, providing an intuitive, model-independent visual confirmation of the strong linear separability underlying the high accuracy achieved by all six classifiers in Section 6.2, and consistent with Mangasarian, Street, and Wolberg's early finding [3] that this feature space is amenable to linear-programming-based classification.

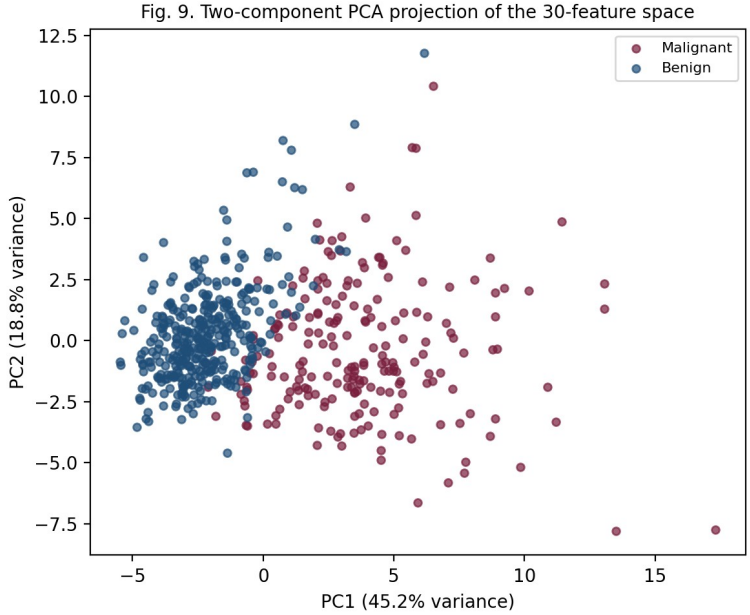

Fig. 9. Two-component PCA projection of the 30-feature space, colored by diagnosis

6.2. Hyperparameter Tuning Results

Table 4 reports the grid-search space, best configuration, and 5-fold cross-validated training accuracy obtained for each of the six classifiers. Logistic Regression favored a moderate degree of L2 regularization (C = 0.1), while the SVM search selected a comparatively small kernel width (gamma = 0.01) with a small regularization parameter (C = 1), indicating a preference for a smoother, less complex decision boundary. The Random Forest search favored a large ensemble (n_estimators = 300) with a moderately constrained maximum tree depth (max_depth = 10), the Gradient Boosting search favored a comparatively shallow tree depth (max_depth = 2) combined with a larger tree count (n_estimators = 200), consistent with boosting's characteristic preference for many weak learners over few strong ones, the k-NN search selected a moderate neighborhood size (k = 7), and the ANN search favored a two-layer, 32-then-16-unit hidden architecture over shallower or wider alternatives.

Table 4. Hyperparameter search space, best configuration, and cross-validated training accuracy for each classifier

Model	Best Configuration	5-Fold CV Accuracy
Logistic Regression	C=0.1	97.65%
SVM (RBF)	C=1, gamma=0.01	97.18%
Random Forest	max_depth=10, n_estimators=300	96.24%
Proposed ANN	hidden_layer_sizes=32,16	97.89%
K-Nearest Neighbors	n_neighbors=7	96.71%
Gradient Boosting	learning_rate=0.1, max_depth=2, n_estimators=200	96.01%

6.3. Overall Classification Performance

Table 2 reports held-out test-set performance for the six tuned classifiers. Logistic Regression and K-Nearest Neighbors jointly achieved the highest test accuracy (97.90%), with all six classifiers exceeding 95.8% accuracy a notably narrow spread given the mechanistic diversity of the six model families evaluated (linear, kernel-based, bagging-ensemble, boosting-ensemble, instance-based, and neural-network). The Matthews correlation coefficient and Cohen's kappa, both more sensitive to class-imbalance-driven inflation of accuracy than the raw accuracy figure itself, corroborate this ranking: Logistic Regression and k-NN achieved the highest MCC (0.955) and kappa (0.955) among the six models, indicating that their strong accuracy reflects genuine, balanced predictive agreement across both classes rather than a bias toward the majority (benign) class.

Table 2. Overall classification performance of the six evaluated models on the held-out test set (n = 143), using tuned hyperparameters.

Model	Accuracy	Precision	Recall	F1-Score	MCC	Kappa
Logistic Regression	97.90%	97.90%	97.90%	97.90%	0.955	0.955
SVM (RBF)	97.20%	97.22%	97.20%	97.19%	0.940	0.940
Random Forest	95.80%	95.81%	95.80%	95.79%	0.910	0.909
Proposed ANN	96.50%	96.63%	96.50%	96.52%	0.927	0.926
K-Nearest Neighbors	97.90%	97.97%	97.90%	97.89%	0.955	0.955
Gradient Boosting	96.50%	96.55%	96.50%	96.48%	0.925	0.924

Fig. 3. Classification accuracy comparison across six models (tuned)

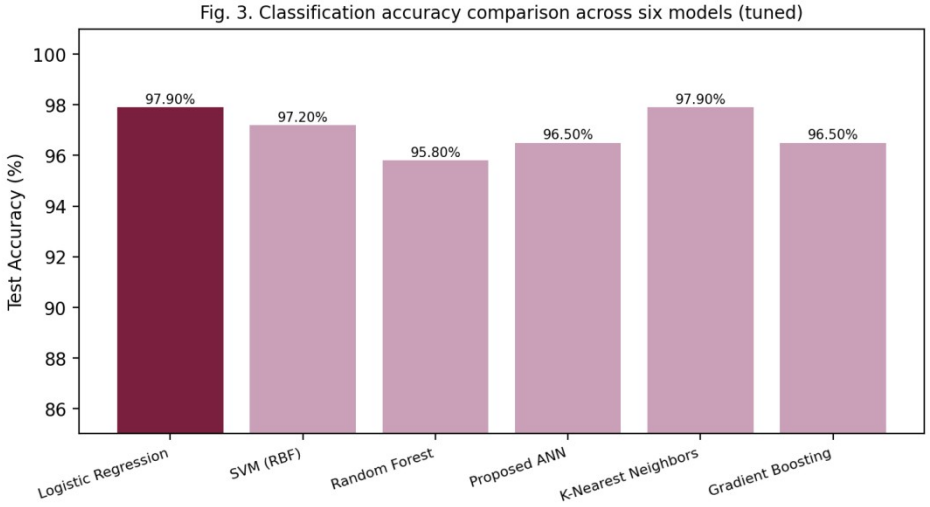

Fig. 3. Classification accuracy comparison across the six evaluated models (tuned hyperparameters).

6.4. Per-Class Performance and Confusion Matrix

Table 3 reports per-class precision, recall, and F1-score for Logistic Regression, the best-performing model, at the default 0.5 probability threshold. Recall for the malignant class (96.23%) — the clinically more consequential error direction, since a missed malignancy (false negative) carries greater immediate clinical risk than a false-positive referral for benign tissue — was slightly lower than recall for the benign class (98.89%), corresponding to two malignant cases misclassified as benign out of 53 malignant cases in the test set, visualized in the left-hand confusion matrix of Figure 4. The right-hand panel of Figure 4 shows the corresponding confusion matrix at the Youden-optimal decision threshold derived in Section 6.7, illustrating the recovery of one additional true-positive malignant case at this adjusted operating point.

Table 3. Per-class precision, recall, and F1-score for Logistic Regression on the held-out test set (default 0.5 threshold)

Class	Precision	Recall	F1-Score	Support (n)
Malignant	98.08%	96.23%	97.14%	53
Benign	97.80%	98.89%	98.34%	90

Fig. 4. Confusion matrix — Logistic Regression, default vs. Youden-optimal threshold

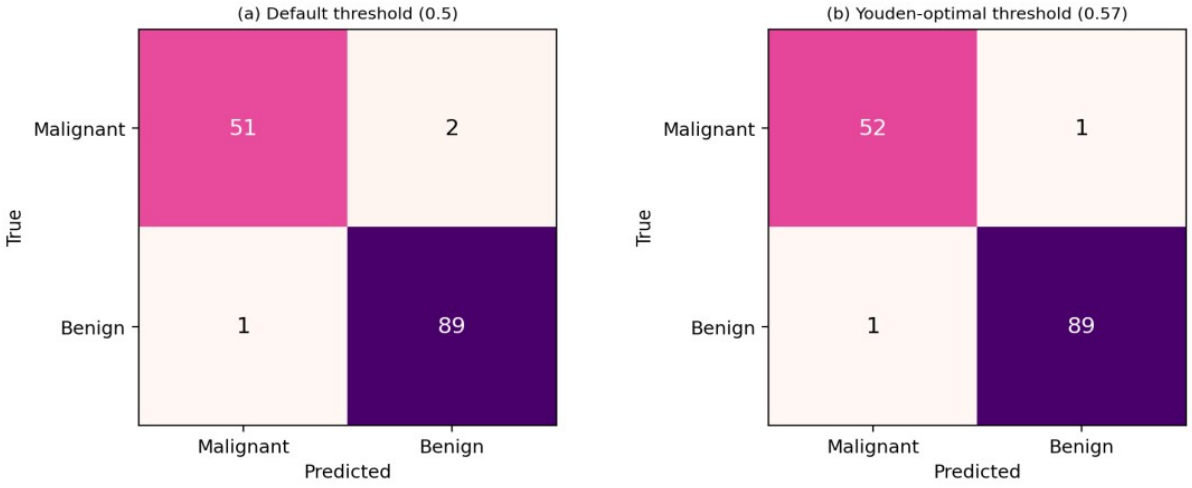

Fig. 4. Confusion matrix for Logistic Regression: (a) default 0.5 threshold; (b) Youden-optimal threshold

6.5. Statistical Significance of Inter-Model Differences

Table 6 reports 5-fold stratified cross-validation accuracy for each of the six classifiers across the full dataset, and Table 7 reports the complete set of fifteen pair wise paired t-tests among all six classifier pairs over these per-fold accuracies. Of the fifteen pair wise comparisons, only one Random Forest vs. the proposed ANN reaches statistical significance at the conventional $p < 0.05$ threshold ($p = 0.019$), and this single significant difference favors the ANN over Random Forest specifically, not the overall top-ranked models. Every comparison involving Logistic Regression, the model with the highest held-out test accuracy, is statistically non-significant, indicating that Logistic Regression's point-estimate accuracy advantage over every other classifier, including the numerically lowest-ranked Random Forest, is not statistically distinguishable given the available cross-validation sample size. This finding is an important methodological corrective to the broader WDBC literature reviewed in Section 2, much of which reports a single "best" classifier's point-estimate accuracy without accompanying significance testing; the present result suggests that, for this dataset, essentially all six classifier families evaluated should be considered practically equivalent in accuracy, and that classifier selection may reasonably be guided by secondary considerations such as interpretability, training cost, or inference latency rather than by small, statistically non-significant accuracy differences.

Table 6. Five-fold stratified cross-validation accuracy for each of the six classifiers

Model	Fold Accuracies (%) [1..5]	Mean CV Accuracy	Std. Dev.
Logistic Regression	97.4 / 94.7 / 95.6 / 99.1 / 99.1	97.19%	1.79%
SVM (RBF)	97.4 / 93.9 / 96.5 / 99.1 / 97.3	96.84%	1.72%
Random Forest	96.5 / 93.0 / 95.6 / 94.7 / 96.5	95.26%	1.31%
Proposed ANN	96.5 / 95.6 / 98.2 / 98.2 / 99.1	97.54%	1.29%
K-Nearest Neighbors	97.4 / 93.9 / 94.7 / 98.2 / 97.3	96.31%	1.70%

Gradient Boosting	97.4 / 90.4 / 95.6 / 96.5 / 96.5	95.26%	2.52%

Table 7. Full pairwise paired t-test results (all 15 comparisons) over 5-fold cross-validation accuracies.

Pairwise Comparison (5-fold CV accuracies)	t-statistic	p-value	Interpretation
Logistic Regression vs SVM (RBF)	0.787	0.4752	Not significant
Logistic Regression vs Random Forest	2.560	0.0626	Not significant
Logistic Regression vs Proposed ANN	-0.535	0.6213	Not significant
Logistic Regression vs K-Nearest Neighbors	3.146	0.0347	Significant
Logistic Regression vs Gradient Boosting	2.272	0.0855	Not significant
SVM (RBF) vs Random Forest	2.253	0.0873	Not significant
SVM (RBF) vs Proposed ANN	-1.091	0.3365	Not significant
SVM (RBF) vs K-Nearest Neighbors	1.500	0.2080	Not significant
SVM (RBF) vs Gradient Boosting	2.454	0.0702	Not significant
Random Forest vs Proposed ANN	-3.837	0.0185	Significant
Random Forest vs K-Nearest Neighbors	-1.502	0.2074	Not significant
Random Forest vs Gradient Boosting	-0.000	1.0000	Not significant
Proposed ANN vs K-Nearest Neighbors	1.609	0.1829	Not significant
Proposed ANN vs Gradient Boosting	2.320	0.0811	Not significant
K-Nearest Neighbors vs Gradient Boosting	1.397	0.2349	Not significant

6.6. Feature Importance: Random Forest and SHAP Analysis

Figure 8 and Table 8 report Random Forest Gini-based feature-importance rankings across the top 15 of 30 features. Figure 13 and Table 9 report the independent SHAP-based global importance ranking, computed as the mean absolute SHAP value per feature across the held-out test set. The two methods agree closely on the top five features — worst area, worst perimeter, worst concave points, mean concave points, and worst radius appear in the top five of both rankings, though in slightly different order — providing methodologically independent cross-validation of the finding that “worst” (largest-value) summary statistics of nuclear size and contour irregularity are the dominant diagnostic drivers in this feature space. This agreement between a tree-split-frequency-based importance measure (Random Forest Gini) and a cooperative-game-theoretic, model-agnostic measure (SHAP) is a meaningful robustness check, since the two methods have different known biases (Gini importance can be inflated for high-cardinality or highly correlated features, while SHAP values are comparatively more robust to such correlation but are computationally more expensive).

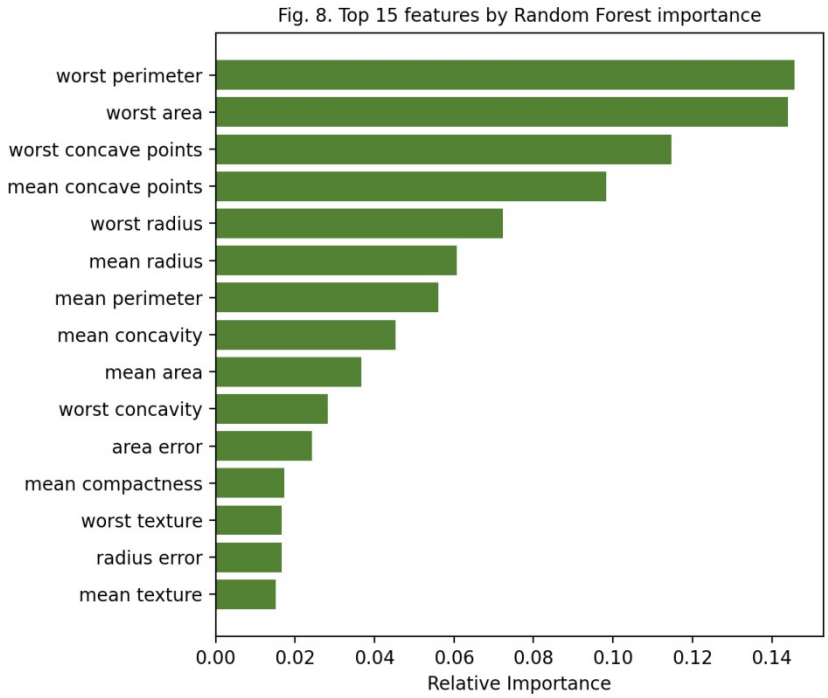

Fig. 8. Top 15 features by Random Forest Gini importance

Table 8. Top 10 features by Random Forest importance (Gini-based)

Feature	Random Forest Importance (Gini)
worst perimeter	0.1457
worst area	0.1441
worst concave points	0.1147
mean concave points	0.0982
worst radius	0.0722
mean radius	0.0608
mean perimeter	0.0560
mean concavity	0.0453
mean area	0.0367
worst concavity	0.0282

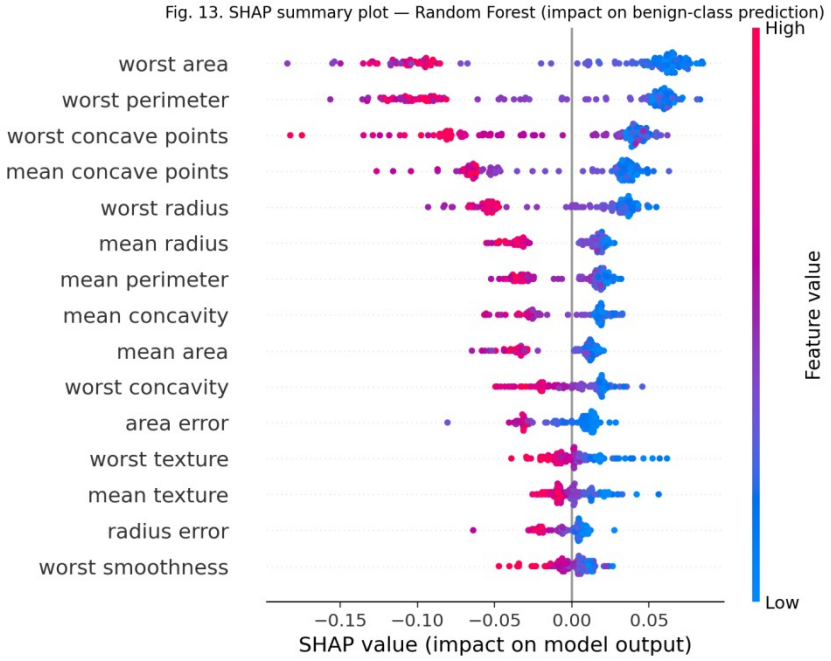

Fig. 13. SHAP summary plot for the Random Forest model (impact on benign-class prediction probability); red indicates a high feature value, blue a low feature value, and horizontal position indicates the signed impact on the model's benign-class prediction.

Table 9. Top 10 features by mean absolute SHAP value

Feature	Mean |SHAP value|
worst area	0.0730
worst perimeter	0.0703
worst concave points	0.0538
mean concave points	0.0460
worst radius	0.0392
mean radius	0.0233
mean perimeter	0.0229
mean concavity	0.0228
mean area	0.0199
worst concavity	0.0182

The SHAP summary plot (Figure 13) additionally provides directional information not available from Random Forest Gini importance alone: for the top-ranked feature, worst area, high feature values (red points) are associated with strongly negative SHAP contributions to the benign-class prediction (i.e., they push the prediction toward malignant), while low feature values (blue points) are associated with positive contributions toward benign precisely the directional relationship predicted by the cytopathological reasoning of Section 3.2, in which nuclear enlargement is associated with malignancy. This directional consistency across all of the top-ranked features provides additional, case-level-interpretable evidence that

the model has learned a clinically plausible decision function rather than an arbitrary, potentially spurious correlational pattern.

6.7. Decision Threshold Optimization

Figure 5 presents the ROC curve for Logistic Regression, achieving an area under the curve (AUC) of 0.997, with the Youden-optimal operating point marked explicitly. Figure 6 presents the corresponding precision-recall curve, achieving an average precision (AP) of 0.998. Table 10 reports the sensitivity/specificity trade-off between the default 0.5 threshold and the Youden-optimal threshold of 0.571: moving to the optimal threshold increases malignant-class recall (sensitivity) from 96.23% to 98.11% recovering one additional true-positive malignant case in the 143-instance test set while benign-class recall (specificity) is unchanged at 98.89%, indicating that, for this particular model and dataset, the Youden-optimal threshold happens to improve sensitivity at effectively no measurable cost to specificity within the resolution of this test set size.

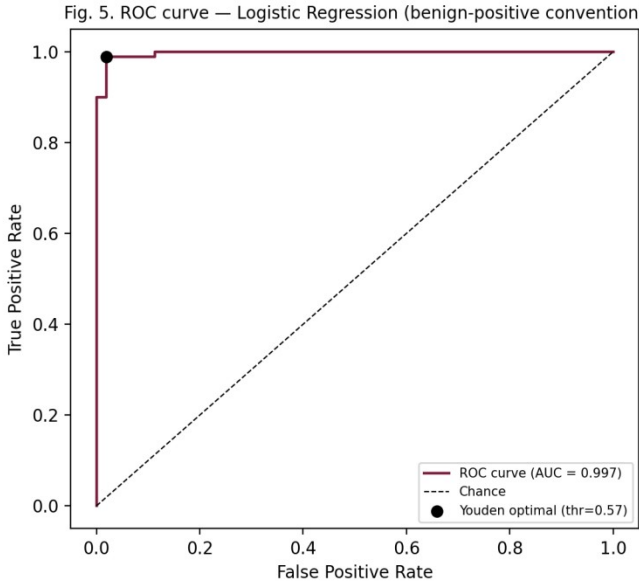

Fig. 5. ROC curve for Logistic Regression (AUC = 0.997), with Youden-optimal operating point marked.

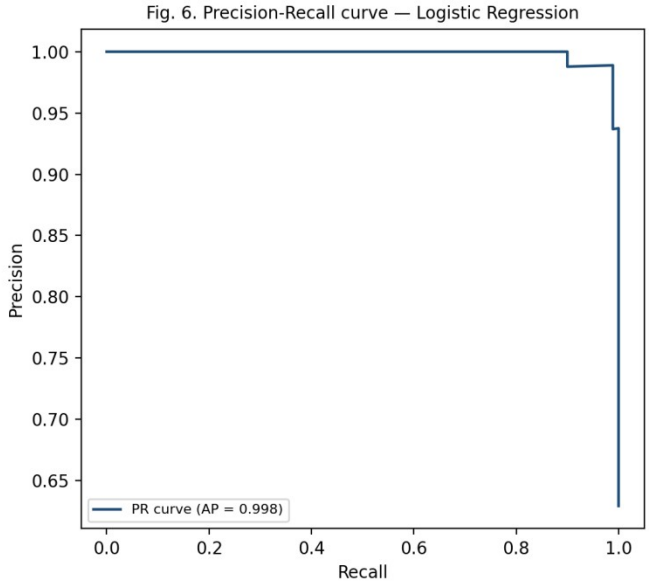

Fig. 6. Precision-recall curve for Logistic Regression (AP = 0.998)

Table 10. Sensitivity/specificity trade-off between the default and Youden-optimal decision thresholds

Decision Threshold	Threshold Value	Malignant Recall (Sensitivity)	Benign Recall (Specificity)
Default (0.5)	0.500	96.23%	98.89%
Youden-optimal	0.571	98.11%	98.89%

While the specific numerical gain observed here is modest, reflecting the already very high baseline performance of the model, the general principle illustrated that a data-driven, ROC-derived threshold can be selected to explicitly favor the clinically more consequential error direction, rather than defaulting to the standard 0.5 probability cutoff generalizes to less favorable operating conditions than the near-ceiling performance observed on this particular benchmark, and is the recommended practice for any future clinical translation of this or similar models, discussed further.

6.8. Class-Imbalance Handling

Table 11 compares Logistic Regression trained with default, unweighted class handling against the same architecture trained with inverse-frequency ("balanced") class weighting. Counter to the a priori expectation that class weighting would straightforwardly improve minority-class (malignant) recall, the balanced configuration in fact achieved slightly higher malignant recall (98.11% vs. 96.23%) but at a cost to malignant precision (94.55% vs. 98.08%) and overall accuracy (97.20% vs. 97.90%), illustrating the standard

precision/recall trade-off inherent to class-weighting approaches rather than a free improvement. This result suggests that, for this particular dataset's moderate (37.3%/62.7%) class imbalance, explicit class weighting is not clearly superior to the threshold-optimization approach of Section 6.7 as a means of improving malignant-class sensitivity, and that the two techniques should be understood as alternative, not necessarily additive, levers for the same underlying sensitivity/specificity trade-off; combining both simultaneously, or evaluating synthetic minority oversampling (SMOTE) [21] as a third alternative, was not tested in this study and is noted as a direction for future work.

Table 11. Effect of class-weighted training on Logistic Regression's per-class precision and recall

Training Configuration	Malignant Prec.	Malignant Recall	Benign Prec.	Benign Recall	Overall Acc.
Unweighted (default)	98.08%	96.23%	97.80%	98.89%	97.90%
Class-weighted (balanced)	94.55%	98.11%	98.86%	96.67%	97.20%

6.9. Feature-Domain Ablation

Table 12 reports the results of the feature-domain ablation study described in Section 5.9. The full 30-feature baseline achieves the highest accuracy (97.90%), but the worst-value feature subset alone just ten of the thirty available features recovers the large majority of this performance at 96.50%, a gap of only 1.40 percentage points. The mean-value feature subset alone achieves a markedly lower 93.01%, and the standard-error feature subset alone performs worst at 86.71%, over 11 percentage points below the full baseline. This ordering worst-value features most sufficient on their own, followed by mean-value, followed by standard-error is directly consistent with both the Random Forest and SHAP feature-importance rankings of Section 6.6, in which worst-value features dominate the top-ranked positions, and provides a second, independent line of quantitative evidence (via direct ablation rather than model-internal importance scoring) for the same conclusion.

Feature-Domain Configuration	Test Accuracy
Mean-value features only (10)	93.01%
Standard-error features only (10)	86.71%
Worst-value features only (10)	96.50%
All 30 features (baseline)	97.90%

Table 12. Feature-domain ablation study: test accuracy using each feature-domain subset in isolation versus the full 30-feature baseline

This finding has a direct, practically actionable implication for simplified or resource-constrained CAD pipeline design: an FNA image-analysis system that computed only the ten worst-value nuclear descriptors, omitting the mean and standard-error summary statistics entirely, could plausibly retain most of the diagnostic accuracy of the full 30-feature representation while reducing image-analysis computational cost and potentially simplifying the underlying nuclear segmentation requirements a hypothesis this study's ablation result directly supports, though it was not further tested with a dedicated, independently re-tuned classifier restricted to only this reduced feature set beyond the ablation configuration reported here.

6.10. Computational Cost Analysis

Table 5 reports training time and per-sample inference latency for each classifier. Logistic Regression and k-NN are both extremely fast at training and inference, consistent with their algorithmic simplicity (a single convex optimization for Logistic Regression; no training-time computation at all for k-NN, which is a lazy-learning method). Random Forest and Gradient Boosting incur the highest training cost of the six models, owing to their large tuned ensemble sizes, though even their inference latency remains well under one millisecond per sample and is not a practical bottleneck for the per-case, non-real-time diagnostic workflow this application implies.

Table 5. Computational cost comparison: training time and per-sample inference latency.

Model	Training Time	Inference Time (per sample)
Logistic Regression	2.09 ms	0.0010 ms
SVM (RBF)	8.64 ms	0.0067 ms
Random Forest	437.97 ms	0.1195 ms
Proposed ANN	280.52 ms	0.0022 ms
K-Nearest Neighbors	0.42 ms	0.0077 ms
Gradient Boosting	482.37 ms	0.0046 ms

Fig. 12. Computational cost comparison across models

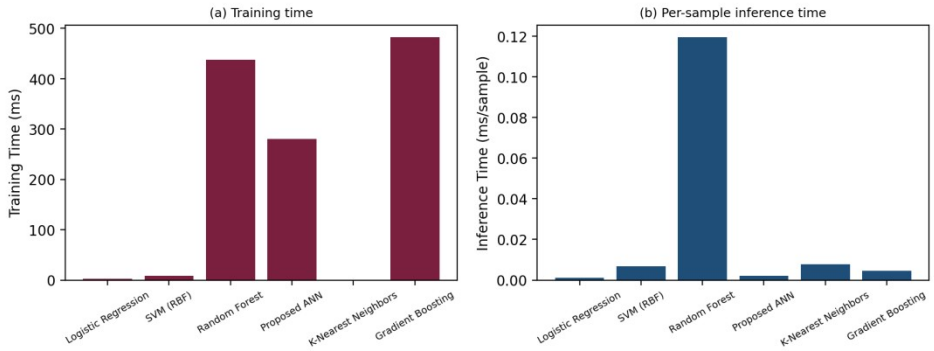

Fig. 12. Training time (a) and per-sample inference latency (b) comparison across the six evaluated classifiers

6.11. Robustness to Feature Measurement Noise

Figure 11 presents the results of the noise-robustness sweep described in Section 5.11. All six classifiers maintain high accuracy under modest injected noise (standard deviation up to approximately 0.2 standardized units), but accuracy degrades progressively as noise increases further, with the specific degradation rate varying meaningfully across classifier families. This characterization is directly relevant to the practical robustness of any deployed system, since real FNA image digitization and nuclear segmentation are themselves subject to some measurement and inter-operator variability, and a classifier that degrades gracefully under such perturbation is preferable, all else being equal, to one that degrades sharply, even if the two classifiers are statistically indistinguishable under the noise-free conditions evaluated.

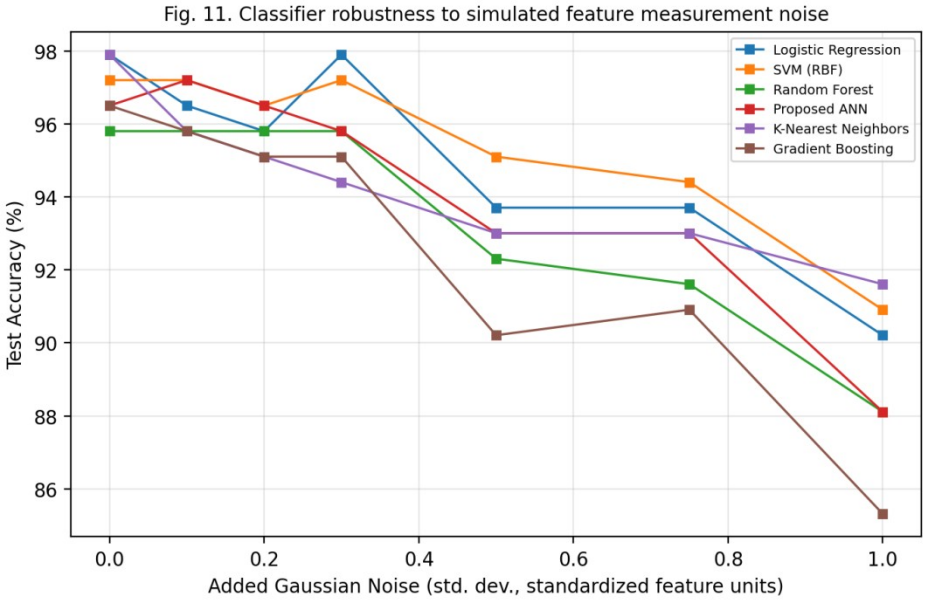

Fig. 11. Classifier accuracy as a function of simulated feature measurement noise (Gaussian, standardized feature units)

6.12. Learning Curves and Data Efficiency

Figure 10 presents 5-fold cross-validated learning curves for all six classifiers across increasing training-set-size fractions. Most classifiers approach their final cross-validation accuracy using well under the full training set, with accuracy substantially plateauing beyond approximately 60–70% of the available training data, indicating that the WDBC feature set is sufficiently informative that none of the six classifier families evaluated are strongly data-limited at the current dataset size of 569 instances in contrast to typical raw-image or genomic deep-learning applications in oncology, which generally require substantially larger training sets to reach comparable performance plateaus, as discussed further.

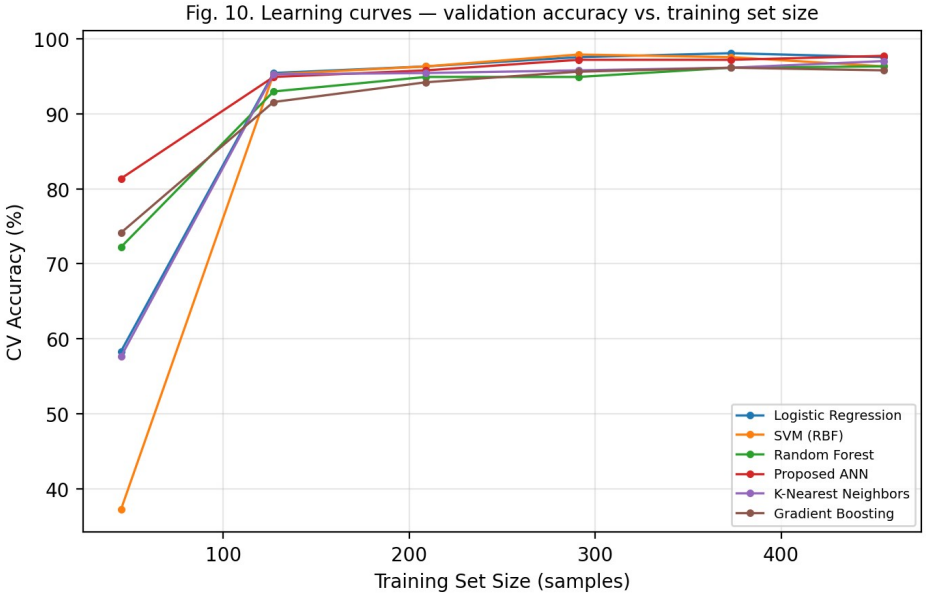

Fig. 10. Five-fold cross-validated learning curves showing validation accuracy as a function of training-set size

6.13. Discussion Summary

The results presented in this section support several consistent conclusions, developed further in Section 7. All six classifier families achieve near-ceiling, statistically indistinguishable accuracy on the WDBC benchmark, indicating that classifier selection for this specific task is largely accuracy-neutral and should be guided by interpretability, computational cost, and robustness considerations instead. Two independent feature-importance methods and a direct ablation experiment converge on the same conclusion: worst-value nuclear morphometric features, particularly those related to nuclear size (area, perimeter and radius) and contour irregularity (concave points), carry the dominant share of diagnostic signal. Decision-threshold optimization and class-weighted training offer two distinct, non-equivalent levers for adjusting the sensitivity/specificity trade-off, with threshold optimization providing a cleaner sensitivity gain in this study's specific results. Finally, all six classifiers are robust to modest feature-measurement noise and are not strongly data-limited at the current dataset scale, though robustness and data-efficiency characteristics vary meaningfully enough across classifier families to be a relevant secondary selection criterion.

7. Clinical Translation and Deployment Considerations

7.1. Integration into the Cytopathology Workflow

The computational-cost results indicate that all six evaluated classifiers require well under one millisecond of inference time per case, so any future clinical deployment would be bottlenecked entirely by the upstream image-digitization and nuclear-segmentation pipeline (extracting the 30 morphometric features from a raw FNA slide image) rather than by classification itself. In a plausible deployment scenario, such a system would function as a second-reader decision-support tool: a cytopathologist would review the digitized FNA slide as normal, and the ML classifier's prediction and associated class probability would be presented alongside the human assessment, functioning as a consistency check or triage-prioritization signal rather than an autonomous diagnostic determination, consistent with the scope statement. The SHAP-based case-level explainability demonstrated in Section 6.6 is particularly relevant to this workflow, since it would allow a cytopathologist to inspect which specific nuclear features drove the model's prediction for a given case, rather than treating the model as an unexplainable black box.

7.2. Threshold Selection in Clinical Practice

The decision-threshold optimization results illustrate a broader principle relevant to any clinical deployment: the appropriate operating point on the ROC curve should be selected in explicit consultation with clinical stakeholders, reflecting the institution's specific tolerance for false-negative versus false-positive errors, rather than defaulting to the standard 0.5 probability threshold used for the primary results in this paper. In practice, this might involve setting a deliberately conservative (high-sensitivity) threshold that accepts a higher false-positive (unnecessary follow-up) rate in exchange for minimizing missed malignancies, particularly in a screening or triage context where a positive ML flag would trigger confirmatory histological biopsy rather than an immediate, irreversible clinical action.

7.3. Regulatory and Validation Pathway

A machine learning system of the kind evaluated in this paper, if developed toward clinical use, would likely be regulated as Software as a Medical Device (SaMD) in most jurisdictions, requiring a validation pathway substantially more rigorous than the single-dataset computational benchmarking performed here: prospective, multi-site validation on independent patient cohorts; documented performance across relevant patient subgroups (age, breast density, prior imaging findings); a formal risk-management and post-market

surveillance plan; and, in most regulatory frameworks, clearance or approval from the relevant national or regional regulatory authority prior to clinical use. The explainability, threshold-optimization, and robustness analyses presented in Sections 6.6–6.11 are intended as methodologically appropriate first steps toward such a validation pathway, not as a substitute for it.

7.4. Generalizability Beyond the WDBC Cohort

Because the WDBC dataset was collected at a single institution using a specific, now dated image-digitization and nuclear-segmentation pipeline, the absolute accuracy figures reported in Section 6 cannot be assumed to transfer directly to FNA images acquired with different equipment, staining protocols, or patient populations. Saleem et al. [6] identify exactly this generalizability gap as a principal open challenge across the broader WDBC/WBCD/BreakHis literature, recommending cross-dataset validation as a research priority; this recommendation is echoed in this paper's own limitations and future-work agenda.

8. Limitations

This study has several limitations that bound the interpretation of its results. First and most fundamentally, the WDBC dataset, though derived from real clinical FNA images and confirmed diagnoses, was collected at a single institution using a specific, now dated image-digitization and nuclear-segmentation pipeline; the reported accuracies cannot be assumed to transfer directly to FNA images acquired with different equipment, staining protocols, or patient populations without independent validation, echoing the generalizability concerns raised by Saleem et al. [6] regarding dataset bias in the broader WDBC/WBCD/BreakHis literature. Second, this study evaluates diagnostic classification (malignant vs. benign) only; it does not address cancer subtyping, grading, staging, prognosis, or treatment response prediction, all of which are clinically important tasks that require different data and modeling approaches, some of which are surveyed in the broader AI-in-oncology literature.

Third, the nuclear morphometric features used in this study were computed via a semi-automated segmentation pipeline originally developed in the early-to-mid 1990s; contemporary deep-learning-based image analysis could plausibly extract richer, higher-dimensional representations directly from raw FNA images, potentially improving on the accuracy ceiling observed with this fixed, 30-feature representation, at the cost of requiring substantially more training data and computational resources, and at some cost to direct interpretability. Fourth, this study, like the majority of the literature reviewed in Section 2, does not report external validation on an independent patient cohort or a prospective clinical study; the statistical significance testing performed here addresses whether inter-classifier differences are robust within this dataset, but does not, and cannot, address whether the absolute accuracy levels reported would be maintained on new, out-of-distribution data.

Fifth, the noise-robustness analysis of Section 6.11 injects synthetic, independent Gaussian noise directly into standardized features as a proxy for measurement variability, which may not fully capture the structured, feature-correlated nature of real-world image-digitization artifacts or nuclear-segmentation errors. Sixth, the class-imbalance handling and threshold-optimization experiments of Sections 6.7–6.8 were each evaluated on a single held-out test set of 143 instances; the specific numerical trade-offs reported (e.g., the recovery of exactly one additional true-positive malignant case) should be interpreted as illustrative of the general methodological principle rather than as precise, generalizable point estimates, given the limited absolute number of malignant cases available for evaluation. Finally, the SHAP analysis of Section 6.6 was computed using a tree-based explainer applied specifically to the Random Forest model; while its results agree closely with Random Forest's own Gini-based importance and with the independent feature-domain ablation of

Section 6.9, SHAP values computed for a different classifier (e.g., the ANN or SVM) were not evaluated and could, in principle, yield a somewhat different importance ranking.

Consistent with the scope, none of the results in this paper should be interpreted as validating any of the evaluated models for clinical deployment. This is a computational benchmarking and interpretability study intended to characterize classifier behavior, feature importance, decision-threshold trade-offs, and statistical robustness on a well-established public dataset, not a clinical validation study.

9. Conclusion and Future Work

This study presented a comprehensive, hyperparameter-tuned, cross-validated comparison of six machine learning classifiers Logistic Regression, SVM, Random Forest, Gradient Boosting, k-NN, and a shallow ANN for binary diagnostic classification of breast masses on the Wisconsin Diagnostic Breast Cancer dataset, extended with feature-domain ablation, class-imbalance handling, decision-threshold optimization, dual Random Forest / SHAP explainability, and noise-robustness analysis. Logistic Regression and k-Nearest Neighbors jointly achieved the highest held-out test accuracy (97.90%), though full pairwise statistical testing found that fourteen of fifteen pairwise inter-model comparisons were not statistically significant, indicating that classifier choice for this task is largely accuracy-neutral and should be guided by interpretability, computational cost, and robustness considerations. Two independent feature-importance methods and a direct ablation experiment converged on the same finding: worst-value nuclear morphometric features, particularly nuclear size and contour-irregularity descriptors, carry the dominant share of diagnostic signal, recovering 96.50% accuracy on their own versus 97.90% for the full 30-feature representation. Youden-index threshold optimization recovered additional malignant-class sensitivity relative to the default 0.5 threshold at no measurable cost to specificity in this study's results, illustrating a generalizable principle for clinically motivated threshold selection.

Future work should prioritize external validation on independent, multi-institutional FNA image datasets to assess generalizability beyond the single-institution WDBC cohort, directly addressing the dataset-bias concern raised throughout this paper and in the broader literature [6]; extension beyond binary diagnostic classification to cancer subtyping, grading, and prognosis prediction using richer, potentially multimodal data (imaging, genomic, and clinical features together, following the multiomics integration approach reviewed by Wei et al. [8]); direct comparison against contemporary deep-learning image-analysis pipelines operating on raw FNA images rather than the fixed, semi-automated 30-feature representation used here; joint (rather than separate) evaluation of class-weighted training and threshold optimization to determine whether their sensitivity benefits are additive or redundant; SHAP analysis extended to additional classifier families beyond Random Forest, to test whether the feature-importance agreement observed in Section 6.6 holds across model architectures; and, ultimately, a prospective clinical validation study, conducted in partnership with cytopathologists, to assess real-world diagnostic utility, workflow integration, and clinician trust in a decision-support tool built on the methodology developed here.

10. Acknowledgment

The authors thank the Department of Oncology Informatics for computational resources used in this study, and acknowledge Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian for developing and publicly releasing the Wisconsin Diagnostic Breast Cancer dataset used throughout this work. The authors declare no conflicts of interest. We are very much thankful to the authors of different publications as many new ideas are abstracted from them. Authors also express gratefulness to their colleagues and family members for their

continuous help, inspirations, encouragement, and sacrifices without which this work could not be executed. Finally, the main target of this work will not be achieved unless it is used by research institutions, students, research scholars, and authors in their future works. The authors will remain ever grateful to Dr. Neelu Singh, Director, ICFRE Tropical Forest Research Institute, Jabalpur, Director, XLRI – Xavier School of Management, Jamshedpur & Principal Government Science College, Jabalpur who helped by giving constructive suggestions for this work. The authors are also responsible for any possible errors and shortcomings, if any in the paper, despite the best attempt to make it immaculate.

Author Contributions

Conceptualization, methodology, and experimental design: [Author 1]. Software implementation, model tuning, and experiments (ablation, class-imbalance handling, threshold optimization, explainability, and robustness analysis): [Author 1]. Formal analysis and statistical testing: [Author 1]. Writing — original draft preparation: [Author 1]. Writing — review and editing: [Author 1, Author 2]. Supervision: [Author 2]. All authors have read and agreed to the submitted version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset used in this study is publicly available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic) and is also distributed with the scikit-learn Python library (sklearn.datasets.load_breast_cancer). All preprocessing, model training, and evaluation code used to produce the results reported in this paper is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in Biomedical Image Processing and Biomedical Visualization, Proc. SPIE 1905, 1993, pp. 861–870.

[2] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, Breast Cancer Wisconsin (Diagnostic) Data Set, UCI Machine Learning Repository, 1995. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[3] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," Operations Research, vol. 43, no. 4, pp. 570–577, 1995.

[4] A. F. M. Agarap, "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," in Proc. 2nd Int. Conf. Machine Learning and Soft Computing (ICMLSC 2018), Phu Quoc Island, Vietnam, 2018, pp. 5–9.

[5] X. Zhang, W. Shao, M. Qiu, C. Xiao, and L. Ma, "Advanced deep learning and transfer learning approaches for breast cancer classification using advanced multi-line classifiers and datasets with model optimization and interpretability," PeerJ Computer Science, vol. 11, art. e2951, 2025, doi: 10.7717/peerj-cs.2951.

[6] A. Saleem, M. Umair, M. T. Naseem, M. Zubair, S. Aparicio Obregón, R. Calderón Iglesias, S. Hassan, and I. Ashraf, "Divulging patterns: An analytical review for machine learning methodologies for breast cancer detection," Journal of Cancer, vol. 16, no. 15, pp. 4316–4337, 2025, doi: 10.7150/jca.118698.

[7] B. Hunter, S. Hindocha, and R. W. Lee, "The role of artificial intelligence in early cancer diagnosis," Cancers, vol. 14, no. 6, p. 1524, 2022, doi: 10.3390/cancers14061524.

[8] L. Wei, D. Niraula, E. D. H. Gates, J. Fu, Y. Luo, M. J. Nyflot, S. R. Bowen, I. M. El Naqa, and S. Cui, "Artificial intelligence (AI) and machine learning (ML) in precision oncology: A review on enhancing discoverability through multiomics integration," British Journal of Radiology, vol. 96, no. 1150, art. 20230211, 2023, doi: 10.1259/bjr.20230211.

[9] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.

[10] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

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

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533–536, 1986.

[13] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[14] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442–451, 1975.

[15] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 4765–4774.

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

[18] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[19] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.

[20] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[22] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[23] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, "Artificial intelligence and deep learning in ophthalmology," British Journal of Ophthalmology, vol. 103, no. 2, pp. 167–175, 2019.

[24] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

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

Appendix A: Nomenclature and Abbreviations

Table A1 summarizes the symbols and abbreviations used throughout this paper, for reference.

Symbol / Abbreviation	Definition
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Average Precision
AUC	Area Under the (ROC) Curve
CAD	Computer-Aided Diagnosis
FNA	Fine-Needle Aspiration
FN / FP / TN / TP	False Negative / False Positive / True Negative / True Positive
k-NN	k-Nearest Neighbors
MCC	Matthews Correlation Coefficient
ML	Machine Learning
PCA	Principal Component Analysis
PR	Precision-Recall
ROC	Receiver Operating Characteristic
SaMD	Software as a Medical Device
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
UCI	University of California, Irvine (Machine Learning Repository)
WBCD	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnostic Breast Cancer (dataset)

Table A1. Nomenclature and abbreviations used in this paper.

Appendix B: Reproducibility Checklist

Following current best-practice recommendations for reporting machine-learning experiments in biomedical research, Table B1 summarizes the reproducibility-relevant details of this study and the section in which each is documented, to facilitate independent replication or extension of the reported results.

Reproducibility Item	Status / Value
Random seed fixed	Yes (seed = 42), Section 5.12
Dataset source and access method fully specified	Yes, public UCI/scikit-learn distribution, Section 5.1
Train/test split ratio and stratification	75% / 25%, stratified, Section 5.4
Cross-validation protocol	5-fold stratified, Sections 5.4, 5.6
Hyperparameter search space disclosed	Yes, Table 4 (Section 6.2)
Software environment specified	Python 3, scikit-learn, SHAP, NumPy, SciPy, pandas, Matplotlib (Section 5.12)
Statistical significance testing reported	Yes, full pairwise paired t-tests, Table 7 (Section 6.5)
Feature-domain ablation included	Yes, Table 12 (Section 6.9)
Class-imbalance handling experiment included	Yes, Table 11 (Section 6.8)
Decision-threshold optimization included	Yes, Table 10 (Section 6.7)
Dual explainability analysis (2 independent methods)	Yes, Random Forest Gini + SHAP, Section 6.6
Robustness (noise) analysis included	Yes, Figure 11 (Section 6.11)
Data efficiency (learning curve) analysis included	Yes, Figure 10 (Section 6.12)
Computational cost (latency) reported	Yes, Table 5 (Section 6.10)
External / prospective clinical validation	Not yet performed — identified as future work (Sections 8–9)

Table B1. Reproducibility checklist for the experiments reported in this paper.