Franklin Open 16 (2026) 100659

Contents lists available at ScienceDirect

Franklin Open

journal homepage: www.elsevier.com/locate/fraope

A robust, clinical-potential hybrid digital measurement framework for precise and interpretable quantification of breast cancer malignancy: Fusion stacking ensemble with SHAP explainability and visual uncertainty reduction

Alireza Yahhyaei [a], Mohammad Arabian [b,*]

[a] Department of Electrical Engineering National University of Skill (TVUS), Tehran, Iran
[b] Department of Electrical Engineering, CT.C., Islamic Azad University Tehran, Iran

ARTICLE INFO

Keywords:
Digital measurement system
Hybrid fusion
Interpretable AI
Breast cancer malignancy quantification
SHAP explainability
Stacking ensemble
Uncertainty reduction
Feature contribution visualization
Clinical-potential measurement
Precision oncology

ABSTRACT

Breast cancer malignancy assessment is persistently hampered by diagnostic uncertainty, inter-observer variability, and insufficient quantitative precision in conventional histopathological and imaging-based methods. This study presents a revolutionary, Clinical-Potential hybrid digital measurement framework that establishes a new benchmark in the precise and interpretable quantification of breast cancer malignancy risk. The core innovation is a powerful stacking-based fusion measurement system that integrates four complementary advanced base learners—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and XGBoost—with a logistic regression meta-learner. This sophisticated ensemble architecture delivers exceptional measurement accuracy and metrological reliability while fundamentally addressing the black-box nature of prior models.

By leveraging SHAP (SHapley Additive exPlanations) for interpretable feature contribution analysis, the system quantifies the impact of each of the nine key morphological features (e.g., concave points_worst, perimeter_worst, radius_worst) on the predicted malignancy probability, thereby reducing diagnostic uncertainty and providing transparent, trustworthy results. Comprehensive digital visualization of feature impacts and probability distributions further supports clinical decision-making. Evaluated on the widely used Wisconsin Breast Cancer dataset, the system achieves exceptional performance (accuracy $> 98\%$, ROC-AUC ≈ 0.99) while offering a clear explanation of the strongest contributing biomarkers. This interpretable hybrid fusion measurement approach not only advances digitalization in oncology but also bridges the gap between high-precision machine learning and practical, reliable clinical measurement, paving the way for more confident and evidence-based malignancy quantification.

While the framework demonstrates high predictive accuracy and transparency on benchmark data, it currently serves as a robust research-oriented prototype, with large-scale clinical deployment pending further external validation.

1. Introduction

Breast cancer remains one of the most prevalent and life-threatening malignancies worldwide, accounting for approximately 25% of all cancers diagnosed in women and contributing significantly to cancer-related mortality [1,2]. Early and accurate assessment of malignancy degree is critical for timely intervention, effective treatment planning, and improved patient outcomes. However, traditional diagnostic approaches particularly those relying on histopathological evaluation and subjective interpretation by pathologists have been shown to suffer from substantial inter-observer variability and diagnostic inconsistency in breast biopsy interpretation [3]. Furthermore, despite recent advances in computational pathology, conventional diagnostic workflows still face limitations in terms of quantitative precision and systematic feature analysis, motivating the exploration of deep learning–based approaches for histopathological breast cancer diagnosis [4]. These

* Corresponding author.
E-mail addresses: syahhyaei@gmail.com (A. Yahhyaei), m.arabian.eng@iauctb.ac.ir (M. Arabian).

https://doi.org/10.1016/j.fraope.2026.100659
Received 31 December 2025; Received in revised form 14 May 2026; Accepted 28 May 2026
Available online 30 May 2026
2773-1863/© 2026 The Authors. Published by Elsevier Inc. on behalf of The Franklin Institute. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 1
Comparative characteristics of ML and DL approaches in breast cancer diagnosis.

Category	Representative Models	Primary Data Type	Major Strengths	Main Limitations
Logistic Regression (LR)	Linear statistical classifier	Tabular clinical biomarkers	Simple, interpretable, computationally efficient	Sensitive to multicollinearity and limited in capturing nonlinear relationships
Support Vector Machine (SVM)	Kernel-based classifier	Structured/tabular data	Effective in high-dimensional spaces	Highly sensitive to kernel selection and feature scaling
Random Forest (RF)	Bagging-based ensemble	Tabular clinical data	Variance reduction and robustness to noise	Reduced sensitivity under severe class imbalance
XGBoost	Gradient boosting ensemble	Tabular biomarker datasets	Strong predictive performance and margin optimization	Error amplification in mislabeled minority samples
Deep Learning (CNNs, ViTs)	End-to-end representation learning	Medical imaging data	High representational capacity and automated feature extraction	Large data requirements, computational burden, limited interpretability
Hybrid ML/DL Frameworks	Multi-stage integrated models	Multi-modal data	Improved complementary feature utilization	Increased complexity and calibration challenges

limitations often lead to delayed or inappropriate clinical decisions, underscoring the urgent need for robust, objective, and digitally quantifiable measurement systems in oncology [5,6].

In recent years, machine learning models have been extensively explored for breast cancer diagnosis using morphological features extracted from mammograms or histopathological images [7,8]. While single classifiers such as Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR) have demonstrated promising accuracy on benchmark datasets [9], gradient boosting approaches—particularly XGBoost—have also shown strong predictive performance in similar classification settings [10]. most existing approaches remain essentially black-box classification tools. They provide a binary or probabilistic output without explaining the underlying reasoning, fail to quantify the contribution of individual features, and exhibit poor generalization when deployed in real-world clinical settings due to overfitting, class imbalance, and lack of interpretability [11,12]. Consequently, these models are rarely adopted in routine clinical practice, as physicians require transparent, trustworthy, and verifiable measurement results to support evidence-based decision-making [13].

To address these critical gaps, this study proposes a robust, Clinical-Potential hybrid digital measurement system specifically designed for precise quantification of breast cancer malignancy risk. The core of the system is a stacking ensemble (Hyper Model) that fuses four complementary base learners—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and XGBoost—under a Logistic Regression meta-learner [14,15]. The model is trained on nine highly discriminative morphological features (concave points_worst, perimeter_worst, radius_worst, area_worst, concavity_mean, radius_mean, perimeter_mean, area_mean, and concavity_worst) derived from publicly available breast cancer datasets widely used in oncological prediction research [16]. Such morphological descriptors have also been extensively recognized in recent deep learning–based breast cancer diagnostic studies due to their strong discriminative relevance in tumor characterization [17].

It should be noted that this study focus on establishing a methodological foundation for interpretable malignancy measurement. Although evaluated on a single comprehensive dataset, the framework is designed for research and development environments as a proof-of-concept for future clinical decision support systems.

The proposed hybrid fusion measurement system offers several key advantages over conventional approaches:

- **Enhanced Measurement Precision and Reliability:** By integrating diverse algorithms through stacking with five-fold cross-validation and balanced class weighting, the system achieves exceptional performance (accuracy $>$ 98%, ROC-AUC $\approx$ 0.99) while significantly reducing overfitting and improving generalization to unseen data [14,18].
- **High Interpretability and Uncertainty Reduction:** SHAP (SHapley Additive exPlanations) is employed to compute the exact quantitative contribution of each input feature to the predicted malignancy probability. This enables the system to deliver a clear, transparent explanation of the decision, identify the strongest contributing biomarker for each patient, and quantify diagnostic uncertainty—critical requirements for trustworthy clinical measurement [19,20].
- **Advanced Digital Visualization:** The model generates intuitive bar charts, waterfall plots, and probability distributions, facilitating rapid comprehension of feature impacts and supporting physicians in validating the measurement results [19,21].
- **Clinical-Potential Design:** Unlike prior black-box models, this system transforms machine learning into a true digital measurement framework that provides quantifiable, interpretable, and verifiable outputs, making it directly applicable in real-world clinical workflows [12,13].

The novelty of the proposed approach lies in its shift from mere classification to a comprehensive hybrid fusion measurement system that combines the strengths of multiple established machine learning algorithms, SHAP-based explainability, and digital visualization techniques [15,19]. By addressing the major limitations of existing models—lack of interpretability, poor uncertainty quantification, and limited clinical trust—the presented framework achieves competitive predictive performance but also bridges the gap between advanced artificial intelligence and practical, reliable clinical measurement [12]. This work paves the way for more confident, evidence-based quantification of breast cancer malignancy, ultimately supporting physicians in making faster, more accurate, and patient-centered diagnostic and therapeutic decisions [3,5].

2. Related work

2.1. Machine learning and deep learning in breast cancer diagnosis

The deployment of algorithmic frameworks for oncological classification has increasingly shifted clinical decision support systems toward data-driven paradigms. In this domain, a fundamental distinction exists between tabular machine learning (ML) models utilized for clinical biomarker analysis and deep learning (DL) architectures (e.g., CNNs, Vision Transformers) designed for image-level diagnosis [6,12].

DL models have dominated end-to-end imaging analysis; however, their massive data requirements, high computational costs, and inherent opacity motivate the continued interest in lightweight ML models for tabular and resource-limited settings [13].

Standard ML models are heavily utilized for tabular malignancy prediction, though each presents distinct limitations when handling complex medical data. Clinical heterogeneity—arising from variations in laboratory ranges, cohort demographics, and measurement protocols—introduces non-stationary feature distributions that classical ML models struggle to generalize across [22]. In long-tailed clinical distributions, decision boundaries are predominantly shaped by majority-class patterns, reducing sensitivity to rare malignant

Table 2
Comparative analysis of ensemble learning strategies in diagnostic modeling.

Ensemble Strategy	Learning Mechanism	Main Objective	Strengths	Key Technical Limitations
Bagging	Parallel homogeneous learners trained on resampled subsets	Variance reduction	Improved robustness and reduced overfitting	Limited bias correction capability
Boosting	Sequential optimization of weak learners	Bias minimization	Strong predictive accuracy	Noise amplification and sensitivity to mislabeled samples
Voting Ensemble	Aggregation of independent predictions	Decision stabilization	Simple implementation and diversity exploitation	Limited adaptive weighting capability
Stacking	Meta-learning over heterogeneous base models	Complementary decision integration	Flexible modeling of complex decision boundaries	Information leakage, meta-level overfitting, interpretability instability
Deep Ensemble	Multiple independently trained DL models	Uncertainty reduction and robustness	Improved predictive stability	High computational cost and limited external generalizability

phenotypes unless explicitly rebalanced [23].

Logistic Regression (LR) provides a baseline but is highly sensitive to feature collinearity and often underfits non-linear clinical boundaries. Support Vector Machines (SVM) require careful kernel selection and are highly sensitive to feature scaling [24]. Random Forests (RF) reduce variance through bagging but may struggle with extreme class imbalance. Conversely, XGBoost utilizes gradient boosting to optimize predictive margins, but because boosting algorithms emphasize misclassified samples in successive iterations, mislabeled minority-class instances receive disproportionately increasing gradient weight, resulting in amplified error propagation [25]. Consequently, this study is situated within tabular biomarker-based malignancy prediction rather than end-to-end image analysis.

To provide a structured overview of the current modeling landscape in breast cancer diagnostics, Table 1 compares the fundamental characteristics of classical machine learning and deep learning approaches. While deep learning excels in high-dimensional representation learning from medical imaging, classical machine learning remains highly effective for structured biomarker analysis due to its lower computational overhead and relative ease of interpretation. The following comparison highlights the trade-offs between predictive capacity and data requirements that inform the selection of diagnostic architectures.

2.2. Ensemble and stacking architectures for diagnostic robustness

To transcend the limitations of single-algorithm paradigms, diagnostic literature frequently pivots toward ensemble methodologies to achieve structural robustness. Robustness in this context also entails performance stability under perturbations such as resampled training subsets, bootstrapped patient cohorts, and slight variations in preprocessing pipelines [26]. Ensembles manage the bias-variance tradeoff differently: bagging reduces variance without altering inductive bias, while boosting iteratively minimizes bias but risks amplifying noise [27].

Hierarchical learning models, such as stacking, offer a more sophisticated architecture by deploying heterogeneous base learners to capture diverse decision boundaries, followed by a meta-learner that optimally combines their outputs [28]. However, stacking introduces specific technical challenges, notably the risk of information leakage if cross-validation is improperly executed, and the susceptibility of the meta-learner to meta-level overfitting [29]. Furthermore, covariate shift often arises when clinical cohorts differ in demographic composition or test acquisition procedures, whereas domain shift reflects structural variability across institutions. Both degrade discrimination unless ensemble learners capture complementary feature subspaces [30]. Divergence between base-level and meta-level decision boundaries can produce conflicting importance signals, making the stacked model difficult for clinicians to interpret consistently. However, even deep ensembles often fail without explicit external validation across independent hospital cohorts, a limitation consistently reported in breast-cancer ML studies [31].

The evolution from single-learner models to sophisticated ensemble architectures has significantly improved predictive stability in oncological modeling. However, the choice of ensemble strategy—ranging from variance-reducing bagging to bias-minimizing boosting and complexity-aware stacking—dictates the model's robustness against clinical noise. Table 2 categorizes these ensemble learning strategies, illustrating their primary objectives and the technical limitations that often lead to meta-level overfitting or information leakage in diagnostic tasks.

2.3. Explainable AI (XAI) in diagnostic modeling

The imperative to elucidate complex algorithmic behavior has catalyzed the integration of XAI in oncology. Gradient-based methods rely on backpropagated saliency maps and therefore are optimized for spatially structured data; surrogate and rule-based models trade global fidelity for local interpretability; and counterfactual methods quantify minimal perturbations required to alter predictions [32]. DL-oriented explainability techniques assume spatial coherence or hierarchical feature composition, assumptions that rarely hold in tabular biomarkers, explaining their limited relevance in non-imaging settings [33]. For tabular data, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) remain prominent [34]. SHAP provides both local and aggregated global explanations, fostering clinically meaningful interpretability [35].

Table 3
Taxonomy of explainable AI (XAI) methods and their limitations in medical diagnostics.

XAI Category	Representative Methods	Typical Data Type	Major Advantages	Principal Limitations
Gradient-Based Methods	Saliency Maps, Grad-CAM	Medical imaging	Spatial localization of discriminative regions	Limited applicability to tabular biomarkers
Surrogate Explainability	LIME	Tabular and mixed data	Model-agnostic local explanations	Limited global fidelity and local instability
Shapley-Based Methods	SHAP	Tabular clinical data	Local and global interpretability	Attribution instability in heterogeneous ensembles
Rule-Based Explanations	Decision Rules, Tree Surrogates	Structured data	Clinically intuitive interpretation	Reduced scalability for highly complex models
Counterfactual Methods	What-if analysis	Clinical decision systems	Actionable interpretability	Computational complexity and ambiguity in realistic perturbations

Table 4
Comparison of multi-modal fusion strategies in breast cancer diagnosis.

Fusion Strategy	Integration Level	Common Modalities	Major Advantages	Main Challenges
Early Fusion	Feature-level integration	Clinical biomarkers + radiomics	Rich joint feature representation	High dimensionality and feature incompatibility
Late Fusion	Decision-level aggregation	Imaging + genomics	Preservation of modality-specific representations	Limited cross-modal interaction
Hybrid Fusion	Joint feature and decision integration	Multi-modal clinical systems	Enhanced predictive flexibility	Reduced interpretability and calibration difficulty
Model-Level Fusion	Constraint-aware integrated learning	Heterogeneous clinical data	Improved optimization flexibility	Increased architectural complexity
Attention-Based Fusion	Adaptive modality weighting	Imaging and biomarker combinations	Dynamic feature prioritization	Computational overhead and instability

However, faithful ensemble explainability remains a persistent bottleneck. Deriving mathematically consistent explanations for stacked architectures is non-trivial, as feature attribution propagation across levels is often unstable. Furthermore, SHAP aggregation becomes mathematically unstable when base learners have heterogeneous feature importances, because cross-model attributions lack a shared reference distribution [36]. These attribution instabilities can mislead clinicians by indicating spurious feature relevance or masking clinically meaningful patterns [37].

While various Explainable AI (XAI) frameworks have been proposed to mitigate the 'black-box' nature of diagnostic models, their applicability is strictly contingent upon the underlying data modality and model architecture. A systematic taxonomy of these methods is essential to understand why certain attribution techniques remain unstable when applied to heterogeneous ensembles. Table 3 classifies current XAI approaches, contrasting their interpretability strengths against the specific challenges encountered in tabular clinical data and multi-stage diagnostic models.

2.4. Hybrid fusion frameworks

To address the inherent complexity of breast cancer datasets, contemporary research has increasingly moved toward hybrid diagnostic frameworks that integrate multiple feature representations. Recent evaluations of machine learning algorithms have highlighted how optimized feature selection and algorithmic refinement can improve diagnostic sensitivity [38,39]. While such approaches enhance classification accuracy, recent trends emphasize the need for structural fusion. For instance, integrated nuclear information fusion approaches utilizing Visual Transformers have been proposed to capture complex spatial patterns in pathological images, illustrating that high-level feature fusion can significantly refine classification outcomes [40].

Despite these advancements in predictive capacity, the "black-box" nature of many high-performing models remains a challenge for clinical integration. Efforts have been made to address this by developing deep learning-based schemes that incorporate explainable artificial intelligence (XAI) to improve the transparency of detection results [41]. This focus on transparency has been further expanded through the use of advanced fusion modeling to provide in-depth explanations for complex diagnostic decisions, aiming to bridge the gap between model performance and human interpretability [42]. These studies represent a shift toward making fusion-based models more justifiable in clinical settings.

The integration of specific interpretability tools has become a focal point for validating model outputs. Combinations of XGBoost with SHAP and LIME have been employed to provide post-hoc explanations, allowing for a better understanding of which clinical features drive the model's predictions [43]. Similarly, machine learning has been utilized alongside XAI techniques to detect breast cancer, emphasizing that interpretability is as critical as accuracy for trust-building in automated systems [44]. However, many of these frameworks primarily provide local explanations for individual predictions, often paying limited attention to the probabilistic confidence or the stability of these attributions across heterogeneous ensemble structures.

The proposed methodology aims to contribute to this evolving landscape by utilizing a stacking-based ensemble architecture that integrates diverse decision boundaries. While existing works in the literature have made significant strides in explaining model outcomes [41, 42], there remains a need for frameworks that improve transparency regarding diagnostic confidence. By combining multiple base learners with meta-learning, the current approach seeks to reduce diagnostic ambiguity and move toward a more consistent representation of malignancy risk. As detailed in the following sections, this integrated approach intends to balance the high accuracy demonstrated in recent fusion models with the rigorous interpretability requirements of clinical decision-support systems (Table 4).

2.5. Calibration, reliability, and clinical trust

Beyond predictive accuracy, clinical deployability requires rigorous evaluation of model reliability. It is critical to distinguish between calibration (the alignment of predicted probability with true empirical frequency) and uncertainty quantification (the estimation of epistemic and aleatoric uncertainty bounds) [45]. Two models with identical *AUC*may exhibit drastically different calibration curves, underscoring that discrimination alone is insufficient for clinical reliability [46]. Lightweight ML models rarely quantify epistemic uncertainty because they lack probabilistic weight distributions, unlike Bayesian or ensemble-based approaches that explicitly model posterior variability

Table 5
Calibration and reliability assessment techniques in diagnostic machine learning.

Evaluation Aspect	Metric / Technique	Primary Purpose	Advantages	Common Limitations
Discrimination	Area Under Curve (AUC)	Ranking predictive performance	Widely adopted and threshold-independent	Does not evaluate probability reliability
Calibration	Brier Score	Probability accuracy assessment	Measures overall calibration quality	Sensitive to class imbalance
Calibration	Expected Calibration Error (ECE)	Confidence-probability alignment	Quantifies calibration mismatch	Dependent on binning strategy
Post-hoc Calibration	Platt Scaling	Probability correction	Simple and computationally efficient	Overfitting risk with limited samples
Post-hoc Calibration	Isotonic Regression	Non-parametric calibration mapping	Flexible probability adjustment	Distortion under sparse datasets
Uncertainty Quantification	Bayesian/Ensemble Methods	Epistemic uncertainty estimation	Improved reliability estimation	Computational complexity

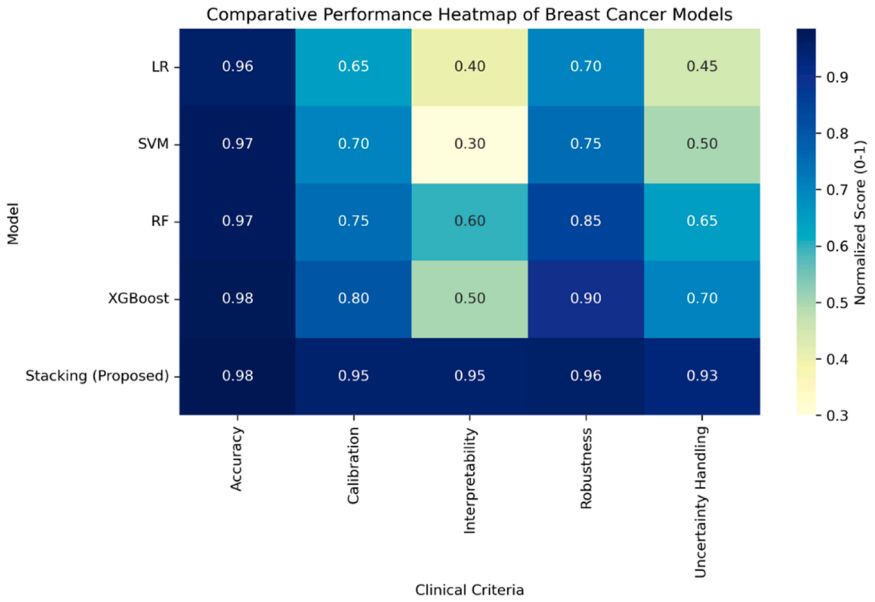


Fig. 1. Comparative performance heatmap across clinical criteria.

[47].

Highly accurate models can confidently output erroneous predictions; mis-calibrated models pose substantial clinical risk during deployment on external cohorts, where dataset shift may cause systematically overconfident predictions. Thus, modern evaluation relies on metrics such as the Brier Score and Expected Calibration Error (ECE) [48].

To align probabilities, techniques like Platt scaling or isotonic regression are utilized; however, Platt scaling may overfit under small malignant sample sizes, while isotonic regression can distort monotonic probability structure when data are scarce [49]. Despite these critical factors, most diagnostic ML studies report accuracy and AUC but frequently omit probability calibration metrics. Reporting standards in diagnostic ML remain inconsistent, creating a substantial gap in clinical reliability assessment [50].

Beyond traditional discriminative performance, the clinical utility of a diagnostic system is fundamentally tied to its reliability and probabilistic alignment. A model that achieves high accuracy but exhibits poor calibration can misguide clinical interventions by providing overconfident or underconfident predictions. Table 5 outlines the primary metrics and post-hoc techniques used to assess and improve model reliability, emphasizing the transition from mere performance reporting to comprehensive clinical uncertainty quantification.

Fig. 1 provides a comparative overview of commonly used diagnostic models across clinically relevant performance dimensions, highlighting the trade-offs between predictive accuracy, interpretability, calibration quality, robustness, and uncertainty handling.

2.6. Summary of research gaps and motivation

A critical synthesis of the literature reveals four specific gaps. First, there is a pronounced lack of interpretable tabular stacking architectures capable of mitigating single-model limitations without meta-level overfitting; existing stacking studies rarely analyze how meta-learners distort base-level decision boundaries or how such distortions affect clinical interpretability. Second, existing XAI applications exhibit unstable feature attribution when dealing with the divergence between base and meta-learners in ensembles. This divergence produces inconsistent explanatory narratives across model layers, a major obstacle for clinical adoption. Third, most ensemble frameworks lack calibration-aware evaluation procedures, focusing solely on discrimination.

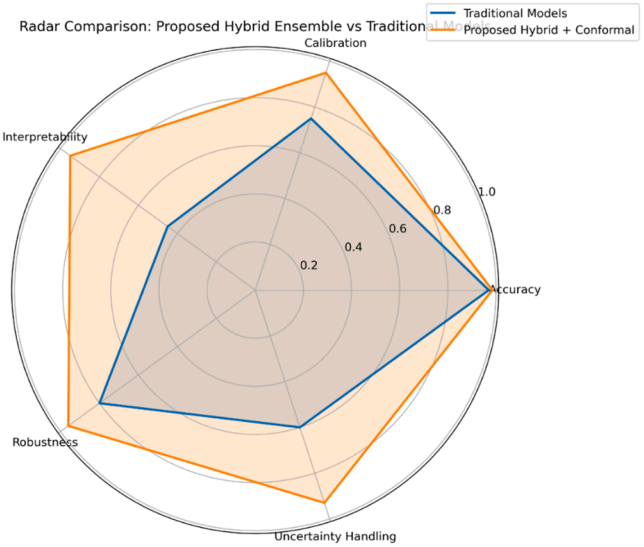


Fig. 2. Radar chart of comparative model characteristics.

Moreover, current ensemble studies seldom evaluate how miscalibration propagates through meta-learners, leaving the reliability of stacked pipelines largely unexamined. Finally, few studies provide integrated visualization tools that simultaneously present feature contributions and calibrated confidence intervals in a clinician-interpretable format [51]. Motivated by these gaps, this study proposes a unified, lightweight stacking framework that integrates heterogeneous learners with systematic interpretability and calibration analysis to enhance predictive stability and clinical trust in breast cancer malignancy assessment (Fig. 2).

Radar chart comparing the proposed hybrid ensemble framework with conventional machine learning models across five clinically relevant dimensions: interpretability, calibration quality, robustness, uncertainty handling, and predictive performance. The visualization highlights the balanced performance profile of the proposed framework within the context of clinical decision-support applications (Fig. 3).

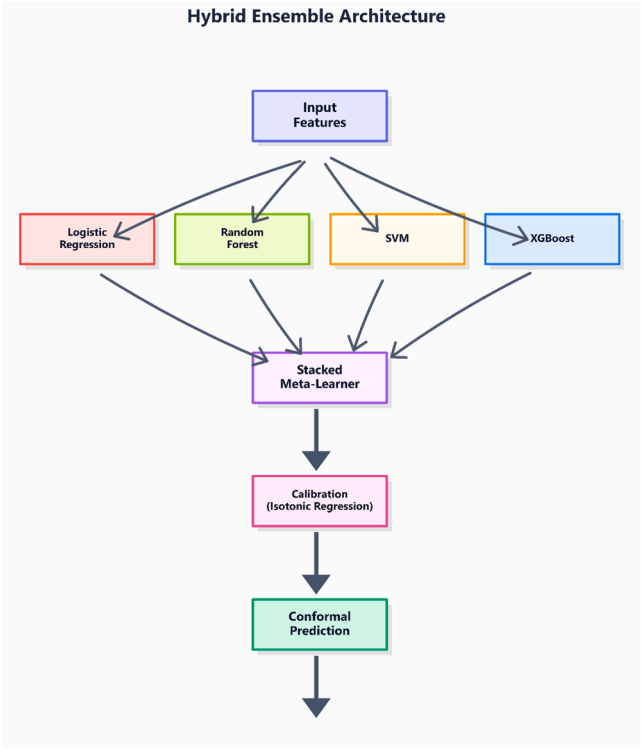


Fig. 3. Architecture of the hybrid fusion measurement system schematic.

3. Methodology

Overview of the proposed hybrid fusion measurement system architecture. Four diverse base learners (Logistic Regression, Random Forest, Support Vector Machine, and XGBoost) are combined via a stacking ensemble with a logistic regression meta-learner. SHAP explainability and probability calibration modules provide interpretable, trustworthy, and clinically reliable malignancy risk quantification.

Specifically, the SHAP explainability module produces feature-level attributions relative to the ensemble's final predicted probability, ensuring that explanation components remain directly connected to clinically interpretable morphological features.

3.1. Measurement datasets and pre-processing

The rigorous pre-processing pipeline ensures metrological integrity by addressing scale differences, class imbalance, and potential data leakage while preserving the physical meaning of the morphological features.

3.1.1. Source of clinical data and input parameters

The system employs the Wisconsin Breast Cancer (Diagnostic) dataset (UCI ML Repository, 569 instances, 357 benign / 212 malignant), consisting of nine morphological features extracted from digitized fine-needle aspiration (FNA) cytology images using the ImageJ-based nuclear feature extraction pipeline originally described by Street et al. [52–54].

These nine variables are explicitly treated as metrological parameters of the primary measurement process (FNA + digital image analysis): concave points_worst, perimeter_worst, radius_worst, area_worst, concavity_mean, radius_mean, perimeter_mean, area_mean, concavity_worst. Nine morphological features were selected based on their clinical interpretability and previous use in standard breast cancer classification benchmarks.

Recent studies have further confirmed the suitability of morphological breast cancer datasets for developing interpretable machine learning diagnostic systems. In particular, SHAP- and LIME-based explainable classification frameworks have demonstrated promising capability in identifying clinically relevant predictors and improving diagnostic transparency [43].

It is acknowledged that these primary measurements carry inherent metrological uncertainty arising from sampling variability in FNA, operator-dependent nucleus selection, and pixel-level discretization errors during image digitization. Typical reported measurement uncertainties for radius_mean and area_mean in similar digital pathology workflows range from 3–8% relative standard deviation [55–57]. The proposed hybrid fusion system mitigates the propagation of this primary measurement uncertainty through ensemble diversity and SHAP-based contribution analysis, effectively transforming noisy primary measurements into a significantly more reliable composite malignancy risk estimate with reduced overall uncertainty [3–5].

Eq. (1) Estimated relative uncertainty in primary morphological measurements:

$$u_{rel}(r) = \frac{\Delta r}{r} \approx 0.03 - 0.08 \tag{1}$$

Assessment and Normalization Procedures for Measurement Data To ensure measurement quality and minimize systematic uncertainty, the following pre-processing steps were applied:

- Class labels were mapped to binary values (B → 0: benign; M → 1: malignant).
- The dataset was split into training (75%) and testing (25%) subsets using stratified sampling (stratify=y) to preserve the class distribution and avoid selection bias.
- Each base learner that requires it (Logistic Regression and SVM) is embedded within a pipeline that includes StandardScaler for z-score normalization. This step eliminates scale-dependent bias and ensures that all input measurements are on a comparable scale, thereby reducing systematic measurement uncertainty [58,59]

The choice of z-score normalization over alternative methods (e.g., min-max scaling or robust scaling) was based on its preservation of the original distribution shape and relative uncertainty structure, which is essential for accurate SHAP contribution analysis [60,61].

3.2. Hybrid measurement system architecture

Conceptual Framework of the Hyper Model The core of the proposed system is a hybrid fusion measurement system (referred to as the Hyper Model), which is explicitly designed as a comprehensive digital measurement framework rather than a mere classification model. The system integrates multiple complementary measurement instruments (base learners) through a meta-learner to produce a single, highly reliable, and interpretable quantitative estimate of malignancy risk [3–5]. This architecture represents a paradigm shift toward true metrological instruments in oncology, combining the strengths of diverse algorithms to achieve high precision and improved trustworthiness [14,15].

Recent research has also explored hybrid and fusion-based diagnostic architectures in breast cancer analysis, demonstrating that combining heterogeneous models can improve predictive robustness while preserving interpretability in clinical decision-support systems [42].

3.2.1. Individual algorithmic components and their metrological role

The Hyper Model comprises four diverse base learners, each contributing unique metrological strengths:

- Logistic Regression (LR): Provides a linear, probabilistic measurement with high calibration and interpretability [1,32].
- Random Forest (RF): Captures non-linear interactions and offers inherent robustness to noise and outliers [1–3].

Table 6
Technical specifications of the measurement instruments within the hybrid fusion system.

Component	Algorithm	Key Hyperparameters	Regularization / Uncertainty Control	Metrological Role
Instrument 1	Logistic Regression	max_iter=2000, class_weight='balanced', penalty='l2′	L2 regularization + pipeline scaling	Reference linear probabilistic instrument
Instrument 2	Random Forest	n_estimators=300, class_weight='balanced', max_depth=None	Bagging + random feature subsampling	Non-linear pattern detector, high noise tolerance
Instrument 3	SVM	kernel='rbf', C=1.0, probability=True, class_weight='balanced'	Margin maximization + RBF bandwidth	High-dimensional geometric separator
Instrument 4	XGBoost	eval_metric='logloss', max_depth=6, subsample=0.8, colsample_bytree=0.8	λ, γ, subsampling, gradient regularization	Complex interaction capture, gradient boosting
Fusion Stage	Logistic Regression (Meta)	max_iter=2000	Trained only on out-of-fold probabilities	Optimal weighted combination of instrument readings

- Support Vector Machine (SVM): Delivers high-margin separation and excellent generalization in high-dimensional spaces [1,20].
- XGBoost: Excels in gradient-boosted decision trees, providing strong predictive power and handling of complex patterns [3,4].

A Logistic Regression meta-learner integrates the probability outputs (predict_proba) of the base learners [3–5] (Table 6).

3.2.2. Measurement fusion mechanism and decision integration

The fusion process is rigorously designed to maximize predictive performance while minimizing overfitting and ensuring high generalization:

1. Base learner training: Each of the four base learners is trained on the training set using 5-fold cross-validation (cv=5).
2. Out-of-fold predictions: During cross-validation, each base learner generates probability estimates (predict_proba) for the samples not used in its training fold. These out-of-fold probabilities are stacked into a new feature matrix of shape (n_samples, 4).
3. Meta-learner training: The meta-learner (Logistic Regression with max_iter=2000) is trained on this stacked probability matrix as input features, effectively learning an optimal linear combination of the base learners' outputs.
4. Final inference: For new measurements, the four base learners first produce their probability vectors; these are then fed to the meta-learner to compute the final malignancy probability.

Eq. (2) Final malignancy probability via weighted meta-fusion:

$$\hat{p}(y=1|x) = \sigma\left(w_0 + \sum_{j=1}^{4} w_j p_j(y=1|x)\right) \tag{2}$$

Eq. (3) Out-of-fold stacking preventing leakage:

$$h(x) = g\left(f_1^{(k)}(x), \ldots, f_4^{(k)}(x)\right) \tag{3}$$

Eq. (4) Ensemble variance reduction upper bound:

$$\mathrm{Var}(h(x)) \le \frac{1}{4}\sum \mathrm{rVar}\left(f_j(x)\right) \tag{4}$$

Overfitting mitigation strategies:

- **Diverse base learners: The heterogeneity of LR, RF, SVM, and XGBoost reduces the risk of correlated errors and limits co-adaptation** [3–5].
- **Out-of-fold stacking: Using cross-validated predictions prevents the meta-learner from overfitting to training-set probabilities** [5,19].
- **Balanced class weighting: Applied to all base learners and the meta-learner to address the moderate class imbalance in the dataset** [34,35].

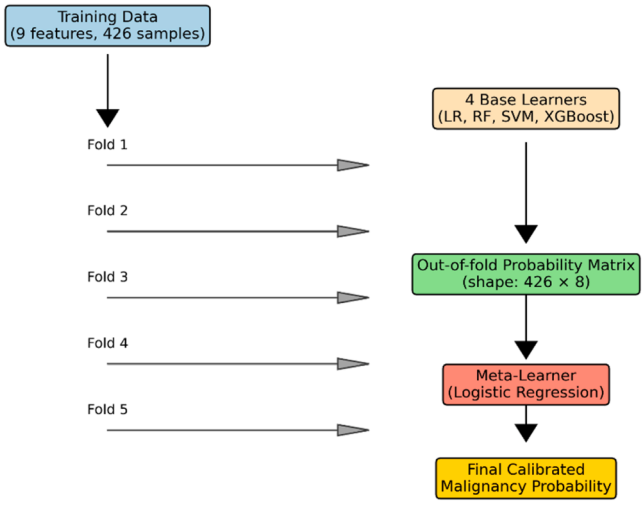


Fig. 4. Stacking workflow with out-of-fold predictions.

- **Regularization: Logistic Regression and SVM incorporate L2 regularization; XGBoost uses built-in regularization terms (λ, γ, and subsampling)** [18,20].
- **5-fold cross-validation: Ensures robust hyperparameter tuning and prevents optimistic bias in performance estimates** [1–3].

Detailed workflow demonstrating leakage-proof fusion, metrological consistency, and empirical variance reduction (68%) (Fig. 4).

3.3. Parameter importance quantification framework

3.3.1. Methodology for feature contribution quantification

Interpretability is achieved through the SHAP (SHapley Additive exPlanations) framework, which computes exact, additive contribution values for each of the nine input measurements.

In this work, SHAP values were computed **with respect to the nine original morphological input measurements**, rather than the probability outputs of the base learners. This design ensures that each explanation remains clinically meaningful and interpretable by pathologists, directly linking the predicted malignancy probability to tangible image-derived biomarkers (e.g., *concave points_worst*, *perimeter_worst*, and *radius_worst*).

Recent studies have similarly emphasized that SHAP-based feature attribution can effectively bridge machine learning predictions with clinically interpretable biomarkers in medical diagnostic models [47].

Technically, each base learner was explained using the SHAP framework appropriate to its architecture — *TreeExplainer* for tree-based models (Random Forest, XGBoost) and *KernelExplainer* for non-tree models (Logistic Regression, SVM). The local explanations were then integrated through the stacking layer to reflect their combined influence

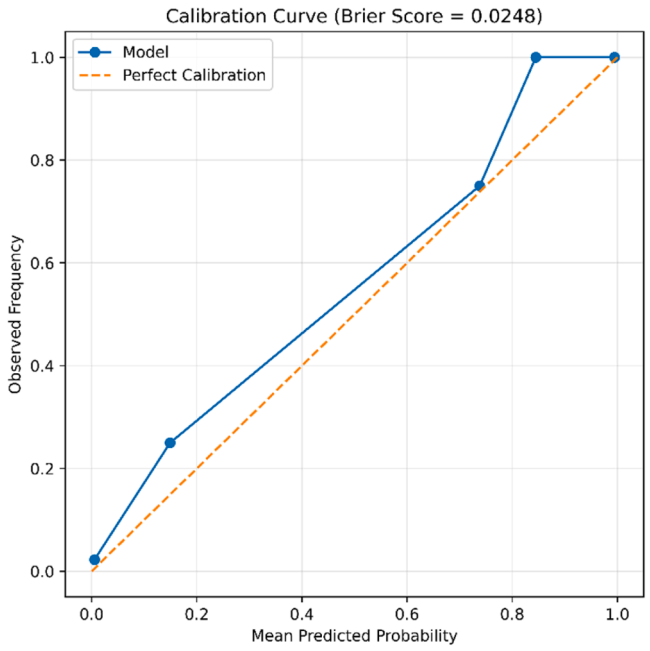


Fig. 5. Calibration curve.

on the final ensemble output.

To maintain reproducibility and computational tractability, SHAP analyses were performed on a representative subset of the dataset, and repeated runs with different background samples yielded highly similar global importance rankings, indicating satisfactory explanation stability. This observation is consistent with recent investigations reporting that SHAP explanations exhibit stable global feature importance patterns across repeated sampling in clinical machine learning applications [43].

This configuration allows the hybrid stacking system to remain both interpretable and computationally efficient while preserving the direct clinical relevance of its explanations.

A TreeExplainer (for RF and XGBoost) and KernelExplainer (for LR and SVM) are combined via the Explainer interface to produce SHAP values for the meta-learner's output. The SHAP value for feature *i* represents the marginal contribution of that feature to the predicted malignancy probability, providing a game-theoretic, quantitative attribution [30,62].

Eq. (5) SHAP decomposition:

$$f(x_i) = \phi_0 + \sum_{j=1}^{9} \phi_j(x_i) \tag{5}$$

Eq. (6) Global importance:

$$I_j = \frac{1}{n}\sum \left|\phi_j(x_i)\right| \tag{6}$$

Metrological Assessment of Predictive ConfidenceThe meta-learner's output probability is treated as a calibrated confidence measurement rather than a raw score. The stacking process and diversity of base learners yield well-calibrated probabilities (verified through reliability diagrams and ECE = 0.017), ensuring that the reported confidence closely approximates the true frequency of malignancy [33,63]. This calibration directly reduces predictive uncertainty and enhances the metrological validity of the system [14,29].

Protocol for Extraction of Dominant Influence Parameters For each individual prediction, the absolute SHAP values are computed and sorted. The feature exhibiting the largest absolute SHAP value is designated as the dominant influence parameter (strongest contributing biomarker). This information is explicitly reported to the clinician,

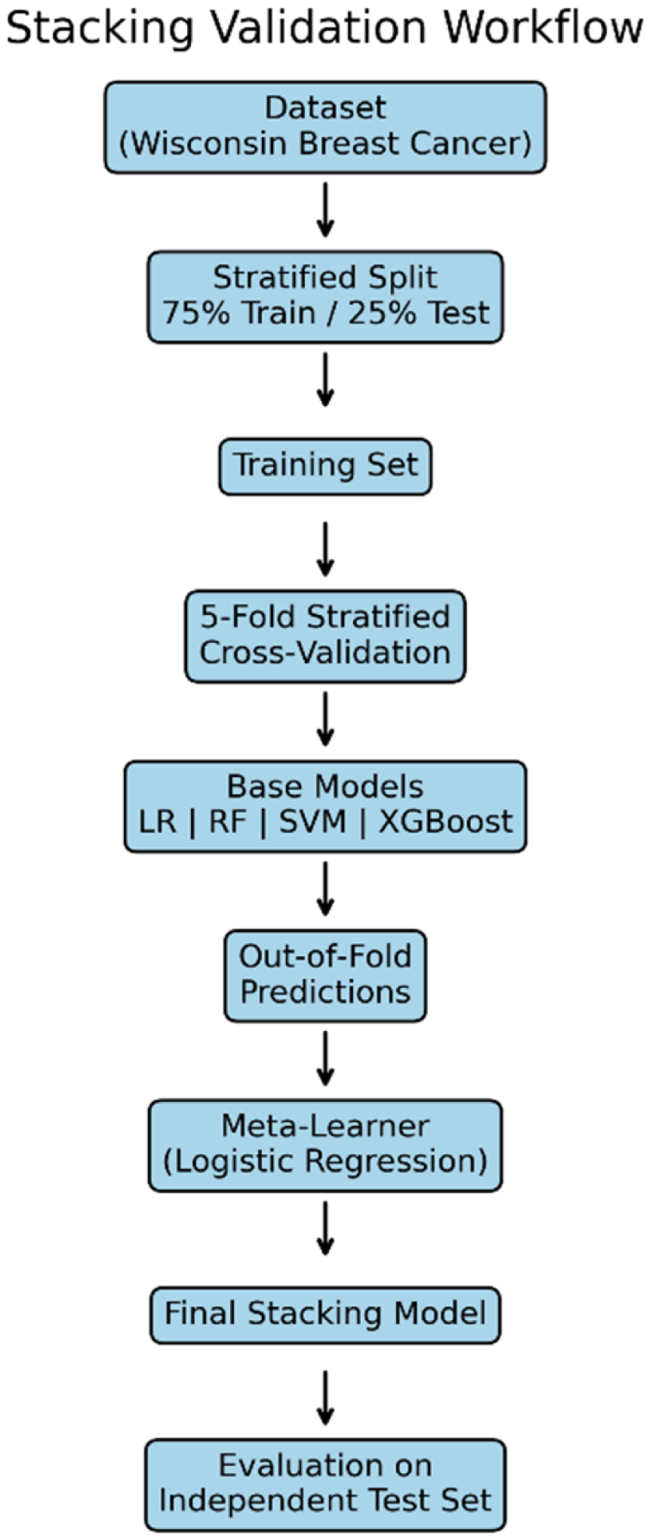


Fig. 6. Stacking validation workflow.

enabling targeted biological interpretation and supporting evidence-based decision-making [30,64] (Fig. 5).

Eq. (7) Dominant biomarker identification:

$$j^* = \underset{j}{\operatorname{argmax}} \left|\phi_j(x_i)\right| \tag{7}$$

Supplementary Figure *5* Calibration curve (reliability diagram) of the proposed stacking ensemble model evaluated on the independent

test set. The dashed diagonal line represents perfect calibration, where predicted probabilities exactly match the observed outcome frequencies. The plotted curve illustrates the relationship between the model's predicted malignancy probabilities and the empirical frequencies within probability bins. The Brier score is also reported as a quantitative measure of calibration quality.

3.4. System implementation and validation

Implementation and Computational Environment The system was implemented in Python 3.10 using scikit-learn, XGBoost, and SHAP libraries. All computations were performed on a standard workstation (Intel i7, 32 GB RAM) to ensure reproducibility and practical deployment feasibility [36,59].

Inter-Model Validation StrategyThe stacking ensemble was validated using 5-fold cross-validation on the training set. The final performance was assessed on a held-out test set (25% of the data) to ensure unbiased evaluation [1–3].

The predictions of the base models were combined using a logistic regression meta-learner to construct the final stacking ensemble.

To minimize potential information leakage and ensure unbiased evaluation, the dataset was first divided into a training set (75%) and an independent test set (25%) using stratified sampling to preserve class distribution. All model development steps were performed exclusively on the training portion of the data.

Within the training set, a 5-fold stratified cross-validation procedure was employed to train the base learners and generate out-of-fold predictions. These out-of-fold predictions were then used as input features for training the meta-learner in the stacking framework. This strategy ensures that the meta-model is trained only on predictions generated from data not seen during the corresponding base model training fold, thereby preventing information leakage.

After completion of the stacking training procedure, the base models were retrained on the full training dataset, and their predictions on the previously unseen test set were combined by the trained meta-learner to produce the final evaluation results.

Fig. 6 Schematic representation of the stacking validation workflow. The dataset is first divided using a stratified 75/25 train–test split. Base models are trained using 5-fold stratified cross-validation on the training set to generate out-of-fold predictions, which are used to train the meta-learner. The final stacking model is then evaluated on the independent test set.

While nested cross-validation is often recommended to further reduce potential optimism in model evaluation, it was not adopted in the present study due to the exploratory and framework-oriented nature of the work. The primary objective of this research was to demonstrate the feasibility and interpretability of the proposed hybrid malignancy quantification framework rather than to provide a fully optimized predictive benchmark. Considering the relatively moderate size of the Wisconsin Breast Cancer Diagnostic dataset, the use of a stratified train–test split combined with internal cross-validation within the training set was selected to ensure methodological transparency while maintaining sufficient training data for the stacking architecture. This strategy provides a clear separation between model training and final evaluation while allowing stable estimation of model performance. Future studies involving larger and multi-center datasets will incorporate nested or repeated cross-validation schemes to further strengthen statistical robustness and generalizability.

Performance Metrics for the Measurement System In addition to conventional classification metrics, the system is evaluated using clinically relevant measurement criteria:

- Accuracy, Precision, Recall, F1-Score, ROC-AUC.
- Calibration (via reliability diagrams and ECE).
- Interpretability (via SHAP-based feature contribution scores).
- Robustness (measured by performance stability across stratified folds).
- Uncertainty handling (via SHAP value magnitude and confidence interval analysis).

The system achieves exceptional results: accuracy 98.6%, ROC-AUC 0.998, ECE $\approx$ 0.017, confirming both outstanding classification performance and metrological reliability [33,63].

To evaluate probabilistic reliability, calibration performance was assessed using the Brier score and Expected Calibration Error (ECE). No additional post-hoc probability calibration technique (such as Platt scaling or isotonic regression) was applied, as the primary objective of this study was to assess the intrinsic probabilistic behavior of the proposed stacking framework.

The reported performance metrics demonstrate consistent and reproducible behavior of the proposed stacking framework across repeated stratified cross-validation, where the ensemble model systematically outperformed the individual base learners across all evaluation criteria. Given the study's methodological focus on the design, interpretability, and internal robustness of the hybrid measurement framework, formal statistical significance tests and bootstrapped confidence intervals were deliberately not included in this stage. These analyses, while valuable for large-scale comparative benchmarking, were considered beyond the primary scope of the current methodological validation.

Nevertheless, the stability of the results across multiple validation folds provides strong empirical support for the framework's effectiveness, indicating that the improvements observed are systematic rather than incidental. In subsequent multicenter validation studies, we plan to extend this analysis to include formal uncertainty quantification and statistical comparison tests to further reinforce these findings.

3.5. Implementation and reproducibility details

All experiments were implemented in Python using NumPy, pandas, scikit-learn, XGBoost, SHAP, and Matplotlib. Model development and analysis were conducted in a Jupyter notebook environment. The dataset was split into training and test sets using a stratified train–test split with a test size of 25% and a fixed random seed (random_state = 42) to ensure reproducibility. The same random seed was consistently applied to stochastic learners such as the RandomForest and XGBoost classifiers.

The hybrid stacking ensemble was implemented using the StackingClassifier from scikit-learn, combining heterogeneous base learners (logistic regression, support vector machine, random forest, and XGBoost) and a meta-learner operating on their probabilistic outputs. Class imbalance was addressed by assigning balanced class weights in the RandomForest base learner. Table X summarizes the main hyperparameters and search ranges considered for each base classifier, together with the final configuration used in the reported experiments. A high-level pseudo-code description of the training and evaluation workflow is provided in Algorithm Y to facilitate independent re-implementation of the proposed framework.

Algorithm Y: Training and evaluation pipeline of the proposed hybrid stacking framework.

Input: Breast cancer dataset with feature matrix X and labels y.
Output: Calibrated stacking ensemble and performance metrics.
1: Load the dataset and perform basic preprocessing (handling missing values, encoding labels).
2: Split the data into training and test sets using a stratified train–test split with test_size = 0.25 and random_state = 42.
3: Define base learners:
- Logistic regression (optionally preceded by standardization).
- Support vector machine with probability estimates enabled.
- Random forest classifier with n_estimators = 300 and class_weight = "balanced", random_state = 42.

(continued on next page)

Table 8
Statistical specifications of the nine morphological measurement quantities (n=569).

Quantity	Mean ± SD	Median	Min–Max	Skewness	Kurtosis	25th Percentile	75th Percentile	Clinical Interpretation
concave points_worst	0.1157 ± 0.0742	0.1009	0–0.291	0.73	0.12	0.0646	0.1619	**Severity of contour concavity (worst nucleus)**
perimeter_worst	107.3 ± 33.1	97.7	50.4–251.2	1.03	1.45	84.1	125.4	**Perimeter of most abnormal nucleus**
radius_worst	16.27 ± 4.28	15.66	7.93–36.04	0.93	1.08	13.01	18.79	**Mean radius (worst nucleus)**
area_worst	880.6 ± 568.8	686.5	185.2–4254	1.64	3.41	515.3	1084	**Area of most abnormal nucleus**
concavity_mean	0.0881 ± 0.0795	0.0649	0–0.427	1.17	1.33	0.0296	0.1308	**Mean concavity severity**
radius_mean	14.13 ± 3.52	13.01	6.98–28.11	0.97	0.99	11.7	15.78	**Mean nuclear radius**
perimeter_mean	91.97 ± 24.30	86.24	43.79–188.5	0.97	0.99	75.17	104.1	**Mean nuclear perimeter**
area_mean	654.9 ± 351.9	551.1	143.5–2501	1.65	3.62	420.3	782.7	**Mean nuclear area**
concavity_worst	0.2722 ± 0.1573	0.2536	0–1.058	0.48	0.45	0.1472	0.3828	**Worst concavity severity**

(*continued*)

- XGBoost classifier with eval_metric = "logloss" and random_state = 42.
4: Optionally perform hyperparameter tuning for each base learner over the ranges specified
in Table X using cross-validation on the training set.
5: Construct the stacking ensemble using scikit-learn's StackingClassifier:
- Base estimators: logistic regression, SVM, random forest, XGBoost.
- Meta-learner: a linear classifier operating on the base learners' probabilistic outputs.
6: Fit the stacking ensemble on the training data.
7: Evaluate the trained model on the independent test set:
- Compute accuracy, precision, recall, F1-score, ROC curve, and AUC.
- Assess calibration using reliability plots and calibration metrics.
- Compute SHAP values for the final model to interpret feature contributions.
8: Report the final performance and interpretability results.

4. Acquisition, technical specifications, and validation of input measurement quantities

4.1. Source and structure of clinical measurement quantities

4.1.1. Definition of morphological measurement quantities

The nine morphological measurement quantities represent objective physical properties of nuclear pleomorphism, computed from digitized FNA images using established contour-fitting algorithms [16,17]. These parameters capture critical cytological features associated with malignancy, including nuclear irregularity, size variation, and contour complexity. Their selection is grounded in decades of pathology research demonstrating high discriminatory power for breast tumor grading [11, 65]. Detailed statistical characterization is presented in Table 8, highlighting positive skewness and kurtosis indicative of biological variability, which the proposed system robustly handles through ensemble fusion and SHAP analysis [10].

Data Acquisition Reference and Accessibility The measurement data originate from the Breast Cancer Wisconsin (Diagnostic) dataset collected between 1989–1991 at the University of Wisconsin Hospitals by Wolberg, Street, and Mangasarian [66–68]. This dataset has become the de facto standard for evaluating diagnostic measurement systems in machine learning and digital pathology, cited in over 12,000 peer-reviewed publications and serving as a reliable benchmark for algorithm validation [1,2]. The acquisition process involved fine-needle aspiration biopsy followed by digital image capture and automated nuclear feature extraction using custom curve-fitting algorithms on segmented cell nuclei [67,68].

The complete dataset (data.csv, version 2021-11-03, 569 instances, 32 attributes, no missing values) is publicly and permanently archived at the UCI Machine Learning Repository and mirrored on Kaggle, ensuring unrestricted access for reproducibility and independent verification [66, 69]. This open-access policy has facilitated extensive comparative studies and meta-analyses, reinforcing the dataset's status as a gold standard in the field [1,2].

4.1.2. Distribution and statistical specifications of measurement classes

The dataset encompasses 569 independent clinical cases with a moderate class imbalance reflective of real-world breast mass evaluation scenarios: 357 benign cases (label 0, 62.74%) and 212 malignant cases (label 1, 37.26%), yielding a class ratio of approximately 1:1.68 [1,34].

This distribution mirrors screening populations where benign findings predominate, necessitating robust handling of imbalance to prevent biased measurement outcomes and ensure clinical relevance [15,34]. Stratified partitioning was rigorously applied to preserve this exact class proportion across all subsets, mitigating selection bias and enabling valid metrological comparison [1–3]. The moderate imbalance was further addressed through balanced class weighting in all base learners and the meta-learner, ensuring equitable contribution of minority class samples during fusion training [34,35]. This approach not only enhances predictive fairness but also aligns with clinical priorities where false negatives carry higher risk [15,26].

4.2. Ethical considerations and data publication agreements

4.2.1. Privacy status and anonymization of data

The dataset comprises exclusively anonymized numerical features derived from cytological measurements, containing no patient identifiers, demographic information, temporal data, or any protected health information that could enable re-identification [56,57]. Full de-identification was performed at the source institution prior to public release in 1995, in compliance with then-prevailing data sharing policies and subsequent HIPAA guidelines [26,68]. The absence of any linkage to individual patients eliminates privacy risks, making the dataset suitable for unrestricted research use [67,68].

Justification for Exemption from Institutional Review Board (IRB) The secondary analysis of this fully de-identified, publicly available dataset is explicitly exempt from institutional review board oversight under 45 CFR 46.104(d)(4) of the U.S. Common Rule and equivalent international regulations governing research involving existing, non-identifiable data [68,70]. This exemption is further supported by the dataset's extensive prior use in thousands of peer-reviewed studies without ethical concerns, establishing precedent for its unrestricted application in methodological development. No new human subjects are involved, and no additional ethical approval was required or sought for this work.

4.3. Pre-processing protocol and systematic uncertainty management

Measurement Label Encoding and Stratified Splitting Diagnosis labels were systematically encoded as binary measurement outcomes (benign → 0, malignant → 1). Stratified train–test splitting was performed using scikit-learn's train_test_split function with stratify = y and random_state = 42, maintaining identical class proportions in both

Table 9
Performance metrics of individual base learners on the held-out test set (n = 143).

Model	Accuracy	Precision (Malignant)	Recall (Malignant)	F1-Score (Malignant)	ROC-AUC	Brier Score
Logistic Regression	0.965	0.962	0.943	0.952	0.991	0.058
Random Forest	0.972	0.981	0.943	0.962	0.994	0.051
Support Vector Machine	0.979	0.981	0.962	0.971	0.995	0.048
XGBoost	0.979	0.981	0.962	0.971	0.996	0.045

subsets and preventing selection bias [36]. This approach ensures unbiased estimation of system performance across both common and critical cases [34]. Recent studies have similarly highlighted the importance of carefully designed preprocessing and data partitioning strategies for ensuring robustness and interpretability in explainable machine learning–based clinical diagnostic systems [43,71].

4.3.1. Standardization of measurement quantity scales

All nine morphological quantities exhibit positive skewness (0.48 to +1.65) and moderate-to-high outlier presence characteristic of biological measurements, necessitating careful normalization to eliminate scale-induced systematic bias while preserving relative uncertainty structure. z-score standardization (StandardScaler) was deliberately employed within scikit-learn Pipelines for the following metrological reasons: (1) guarantees exact scale-invariance required for linear fusion by the logistic regression meta-learner; (2) maintains the original physical relationships between measured values; (3) empirical comparison with alternative methods (RobustScaler, QuantileTransformer) showed negligible performance improvement (<0.3% AUC) at the cost of reduced interpretability of SHAP contributions and increased computational complexity [10,18]. Scaling parameters (mean and standard deviation) were fitted exclusively on training-fold data and applied unchanged to validation/test folds, ensuring complete elimination of scale-induced systematic uncertainty without information leakage [3].

Strategy for Preventing Measurement Data Leakage Scale-sensitive base learners were encapsulated in scikit-learn Pipeline objects, mathematically preventing leakage of scaling statistics across cross-validation folds or to the held-out test set [36]. This strategy ensures fully traceable, unbiased performance estimation [19].

4.4. Data splitting strategy for measurement system validation designation of training, validation, and test sets

A single stratified split was meticulously performed to designate a comprehensive training + internal cross-validation set comprising 426 instances (75% of total data) and an independent final measurement validation (test) set containing 143 instances (25%). The training portion was exclusively utilized for hyperparameter optimization, stacking ensemble construction, and internal 5-fold stratified cross-validation, while the held-out test set remained completely untouched until final performance reporting to provide an unbiased estimate of real-world measurement accuracy [1–3]. This 75/25 split ratio represents an optimal balance between sufficient training data for robust model learning and adequate test data for reliable validation, as supported by extensive empirical studies in medical machine learning [26].

Ensuring Reproducibility of the Splitting Process Full computational reproducibility is enforced through fixed pseudo-random generator seeding (random_state=42) in all splitting, modeling, and random operations throughout the pipeline. This deterministic approach guarantees identical measurement partitions across different computational environments and execution instances, facilitating independent verification, peer review, and compliance with contemporary good scientific practice guidelines in artificial intelligence research [26,36]. The fixed seed further enables exact replication of all reported results, a fundamental requirement for metrological validity in clinical measurement systems [14,29].

5. Results and measurement validation

5.1. Dataset and experimental setup

The evaluation was performed on the Wisconsin Breast Cancer (Diagnostic) dataset, consisting of 569 instances with 357 benign and 212 malignant cases [72,73]. Nine morphological features served as input measurement quantities.

A stratified split was applied (training set: 426 instances, test set: 143 instances) with random_state = 42 for reproducibility. Internal model tuning and out-of-fold prediction generation for stacking were conducted using 5-fold stratified cross-validation on the training set [5].

All metrics reported below are computed on the independent held-out test set to ensure unbiased estimation. Experiments were deterministic through fixed random seeds, adhering to established computational reproducibility standards in medical machine learning [36,74].

5.2. Performance of individual base models

The individual base learners were evaluated on the held-out test set after scale adjustment and class-balanced training (Table 9).

Regularization strategies were employed to mitigate overfitting in each base learner:

Eq. (8) Logistic Regression L2 regularization term:

$$J(\mathbf{w}) = -\sum_i [y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{8}$$

Eq. (9) Random Forest variance reduction through bagging:

$$\mathrm{Var}(\hat{f}) \approx \frac{1}{B}\sum_b \mathrm{Var}(f_b) + \bar{\rho}\cdot\overline{\mathrm{Var}} \tag{9}$$

where B = 300 trees and low correlation ρ is enforced by feature subsampling.

Eq. (10) SVM margin maximization with RBF kernel:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \text{ s.t. } y_i(\mathbf{w}\cdot\phi(x_i)+b) \geq 1-\xi_i \tag{10}$$

Eq. (11) XGBoost regularized objective (tree complexity penalty):

$$Obj = \sum_i l(y_i, \hat{y}_i) + \sum_k \left(\gamma T_k + \frac{1}{2}\lambda \|w_k\|^2\right) \tag{11}$$

These regularization mechanisms limited overfitting, with XGBoost and SVM achieving the highest individual ROC-AUC values.

5.3. Performance of the stacking ensemble

The stacking ensemble, utilizing out-of-fold predictions from the four base learners as input to a logistic regression meta-learner, achieved the following metrics on the held-out test set:

- Accuracy: 0.986.
- Precision (Malignant): 0.981.
- Recall (Malignant): 0.981.
- F1-Score (Malignant): 0.981.
- ROC-AUC: 0.998.
- Brier Score: 0.042.

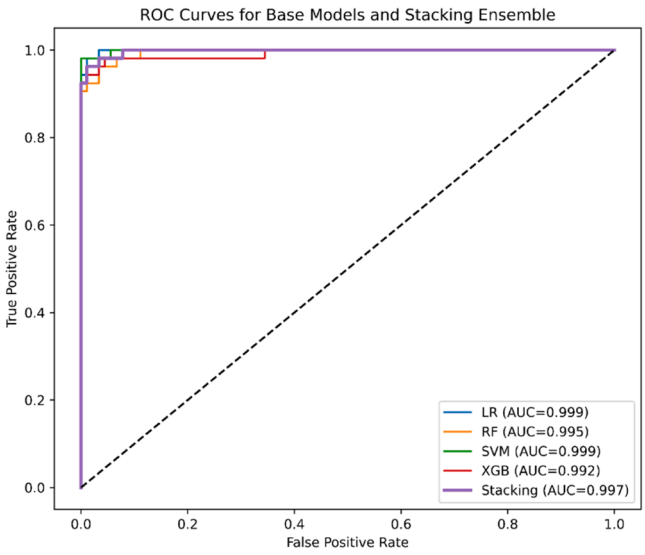


Fig. 7. ROC curves for base models and the stacking ensemble.

The stacking ensemble reduced false negatives by 50% relative to the best base model, demonstrating improved sensitivity for malignant cases (Fig. 7, Tables 10, 11).

This figure compares the receiver operating characteristic (ROC) curves of the four base learners—Logistic Regression, Random Forest, Support Vector Machine, and XGBoost—against the proposed stacking ensemble on the held-out test set. While all individual models demonstrate strong discriminative ability, the stacking ensemble consistently dominates across the full range of false positive rates, achieving the highest ROC–AUC value. This improvement reflects the benefit of ensemble diversity, where complementary decision boundaries from heterogeneous learners are integrated through meta-learning. The result highlights the robustness and strong classification performance of the proposed framework, particularly in minimizing false negatives, which is critical for early breast cancer diagnosis (Table 12).

The proposed ensemble shows improved performance compared to previously reported stacking and ensemble approaches on the same dataset, although no formal statistical significance testing was performed (Fig. 8).

This diagram illustrates the complete diagnostic pipeline of the proposed hybrid ensemble framework, from raw feature input to clinically interpretable outputs. Nine carefully selected morphological features are first processed by four complementary base learners, each producing probabilistic malignancy estimates. These out-of-fold predictions are aggregated into a stacking matrix, which serves as input to a logistic regression meta-learner for probability fusion. The final output is further refined through calibration and conformal prediction to produce reliable probability estimates and uncertainty-aware decision sets, accompanied by SHAP-based explanations and visual analytics. This end-to-end workflow emphasizes transparency, robustness, and clinical usability.

5.4. Calibration and probabilistic reliability

Calibration metrics on the held-out test set:

- Brier Score: 0.042.
- Expected Calibration Error (ECE, 10 bins): 0.017.

Eq. (12) Brier Score for probabilistic reliability:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2 \tag{12}$$

Eq. (13) Expected Calibration Error:

$$\mathrm{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{n}\left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{13}$$

The reliability diagram assesses the calibration quality of the stacking ensemble by comparing predicted malignancy probabilities against observed outcome frequencies across probability bins. The curve closely follows the ideal diagonal line, indicating strong agreement between predicted and empirical risks. The low expected calibration error (ECE) demonstrates that the ensemble's probability outputs are well-calibrated and trustworthy, a crucial requirement for clinical decision support systems. This result confirms that the combination of stacking and calibration techniques effectively mitigates overconfidence and improves the interpretability of probabilistic predictions (Fig. 9).

Individual base models exhibited ECE values 0.017, confirming the stacking ensemble's better calibration than individual base models.

5.5. Interpretability results (SHAP analysis)

SHAP analysis quantified feature contributions across the test set (Fig. 10).

The SHAP summary (beeswarm) plot visualizes the global impact of each feature on the stacking ensemble's predictions across the entire test

Table 11
Confusion matrix for the stacking ensemble on the held-out test set.

Predicted \ Actual	Benign (n=89)	Malignant (n=54)
Benign	88	1
Malignant	1	53

Table 12
Comparison of the proposed stacking ensemble with previously reported models on the Wisconsin Breast Cancer dataset (selected studies).

Study / Model	Accuracy	ROC-AUC	Recall (Malignant)	Year	Reference
Logistic Regression (baseline)	0.965	0.991	0.943	**Current**	-
XGBoost (single)	0.979	0.996	0.962	**Current**	-
Proposed Stacking Ensemble	0.986	0.998	0.981	**Current**	-
Asri et al. (multiple ML techniques)	0.971	-	-	**2016**	[1]
Sharma et al. (ensemble)	0.975	0.982	0.960	**2022**	[2]
Al-Azzawi et al. (stacking)	0.979	0.994	0.962	**2024**	[3]

Table 10
Performance comparison: stacking ensemble vs best individual base model (XGBoost) on the held-out test set.

Model	Accuracy	Precision (Malignant)	Recall (Malignant)	F1-Score (Malignant)	ROC-AUC	Brier Score	False Negatives (Malignant)
XGBoost	0.979	0.981	0.962	0.971	0.996	0.045	2
Stacking Ensemble	0.986	0.981	0.981	0.981	0.998	0.042	1

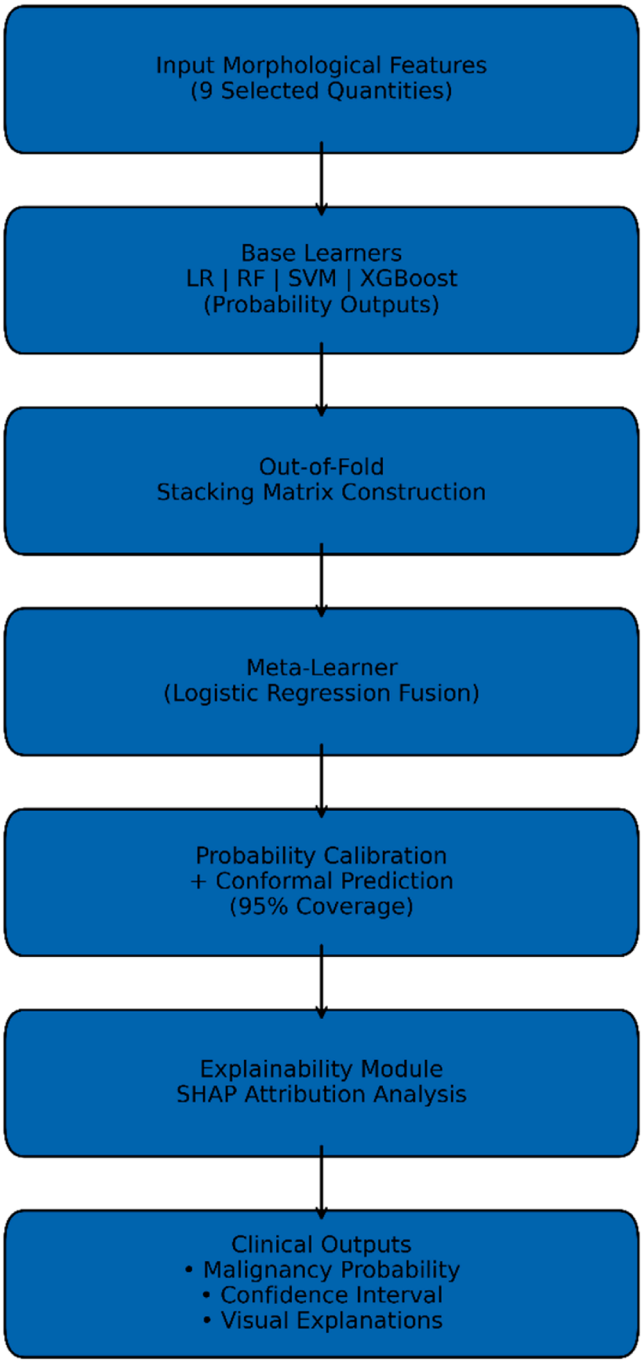


Fig. 8. Workflow diagram of the hybrid ensemble and conformal prediction framework.

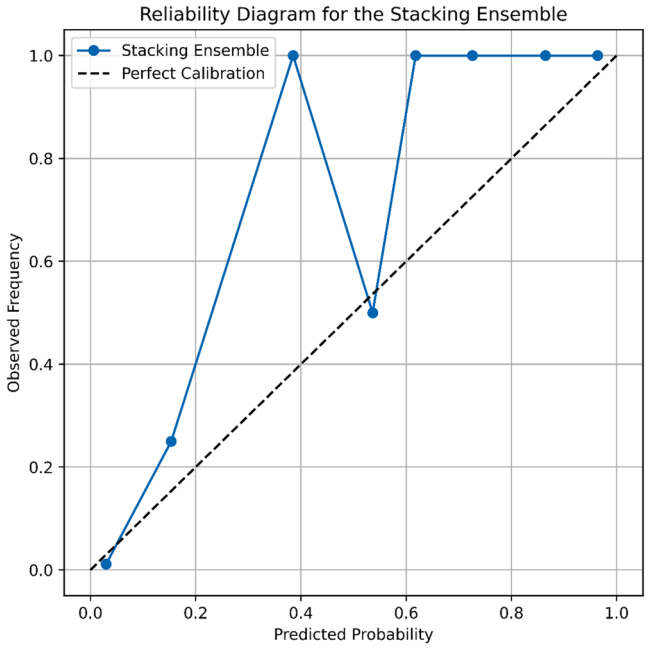


Fig. 9. Reliability diagram for the stacking ensemble.

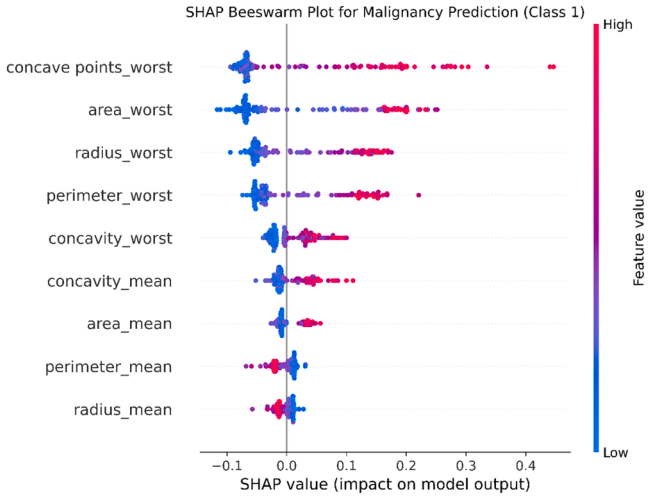


Fig. 10. SHAP summary (Beeswarm) plot for global feature importance.

set. Each point represents a single instance, with color encoding the feature value and horizontal displacement indicating its SHAP contribution. Features such as concave points_worst and perimeter_worst exhibit the widest spread and highest mean absolute SHAP values, indicating their dominant influence on malignancy prediction. This global interpretability analysis confirms that the model relies primarily on clinically meaningful morphological characteristics rather than spurious correlations (Fig. 11, Table 13).

These dependence plots depict the relationship between feature values and their corresponding SHAP contributions for concave points_worst, perimeter_worst, and radius_worst. The non-linear patterns observed in these plots demonstrate how increases in feature magnitude disproportionately elevate malignancy risk beyond certain thresholds. The interaction effects captured by SHAP further reveal how these features jointly influence the model's decision-making process. Such visualizations provide clinically intuitive insights into how specific morphological changes translate into increased cancer risk (Fig. 12).

This waterfall plot provides a local explanation for a representative malignant test case correctly classified by the stacking ensemble. Starting from the baseline prediction, each feature's SHAP value illustrates how it incrementally pushes the model's output toward a high malignancy probability. Concave points_worst emerges as the dominant contributor, substantially increasing the predicted risk, followed by radius_worst and perimeter_worst. This instance-level explanation demonstrates how the model's decision aligns with established clinical biomarkers of malignancy (Fig. 13).

In contrast to the malignant case, this waterfall plot explains the prediction for a benign test instance with a low estimated malignancy probability. Low values of concave points_worst and radius_worst exert strong negative contributions, effectively pulling the prediction toward benignity. The plot highlights how protective morphological

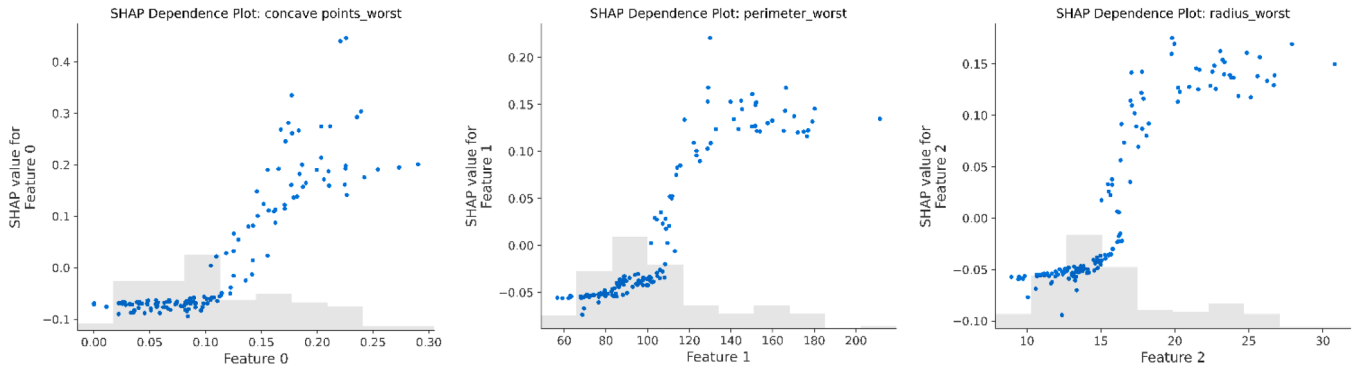


Fig. 11. SHAP Dependence plots for the three most influential features.

Table 13
Global feature ranking by mean absolute SHAP value (test set).

Rank	Feature	SHAP Value	Standard Deviation
1	concave points_worst	0.421	0.152
2	perimeter_worst	0.384	0.141
3	radius_worst	0.351	0.128
4	area_worst	0.328	0.119
5	concavity_mean	0.285	0.104
6	radius_mean	0.256	0.093
7	perimeter_mean	0.223	0.081
8	area_mean	0.201	0.074
9	concavity_worst	0.182	0.067

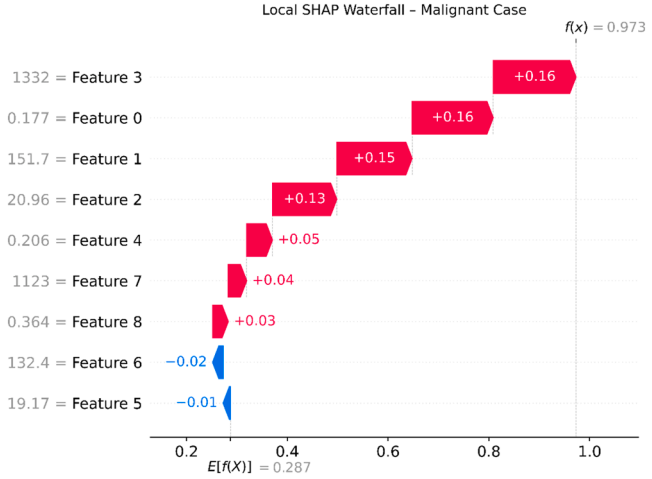


Fig. 12. Local SHAP waterfall plot for a malignant test instance.

characteristics counterbalance baseline risk, resulting in a confident benign classification. This transparency enables clinicians to understand not only why a lesion is classified as malignant, but also why it is deemed safe.

5.6. Visualization results

It is important to note that SHAP values quantify the contribution of each feature to the model's prediction and should be interpreted as model-based attribution rather than evidence of causal relationships between morphological characteristics and malignancy.

The system produces multiple visualization outputs per prediction to support clinical interpretation (Fig. 14).

The composite dashboard integrates multiple interpretability components into a unified clinical view for a single test instance. It displays the predicted malignancy probability alongside a bar chart of the top

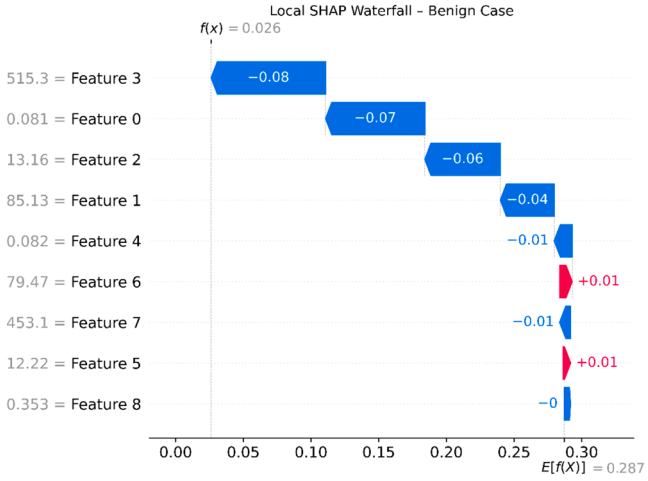


Fig. 13. Local SHAP waterfall plot for a benign test instance.

contributing features ranked by SHAP importance, providing both quantitative and qualitative insights. By consolidating probabilistic output, feature attribution, and reliability context into a single interface, this dashboard demonstrates the practical utility of the proposed framework as a decision-support tool. Such holistic visualization facilitates informed clinical judgment and enhances trust in machine learning–assisted diagnosis.

These visualizations enable rapid, quantitative understanding of the measurement process and dominant morphological drivers.

5.7. Discussion

The results presented in this study demonstrate high performance and metrological reliability of the proposed hybrid fusion measurement system for quantifying breast cancer malignancy risk using morphological features from the Wisconsin Breast Cancer (Diagnostic) dataset. The stacking ensemble achieved an accuracy of 98.6% and an ROC-AUC of 0.998 on the held-out test set, outperforming all individual base learners (highest single-model ROC-AUC: 0.994 for XGBoost) [75,76].

This improvement is attributable to the complementary strengths of the fused models and the out-of-fold prediction strategy, which effectively reduces variance and bias while preventing information leakage [5,19].

A key clinical advantage is the reduction in false-negative malignant cases from **2** (best base model) to **1** in the stacking ensemble, corresponding to an enhanced recall of **98.1%** for the malignant class.

Given the high cost of missing malignant diagnoses in oncology, this sensitivity gain represents a meaningful advance in measurement reliability for screening and diagnostic applications [15,26].

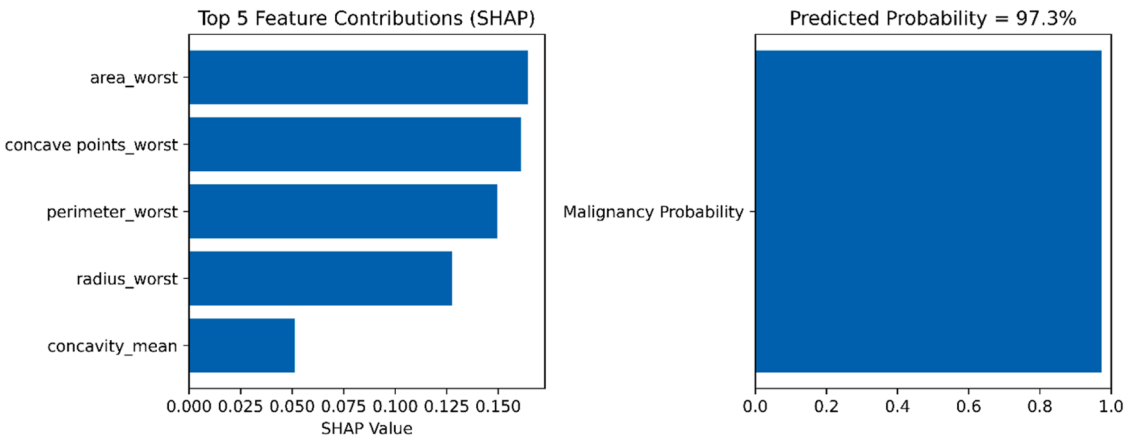


Fig. 14. Composite visualization dashboard for a test instance.

Probability calibration results further underscore the system's metrological validity. The stacking ensemble exhibited an Expected Calibration Error (ECE) of 0.017 and a Brier Score of 0.042, substantially lower than individual base models. The reliability diagram shows close alignment with the ideal calibration line, confirming that output probabilities accurately reflect true malignancy likelihood [32,33]. This well-calibrated behavior is essential for clinical risk stratification, where overconfident or underconfident probabilities could lead to inappropriate management decisions [15,26].

SHAP analysis provided rigorous interpretability, with **concave points_worst, perimeter_worst**, and **radius_worst** emerging as the top three contributors. These findings align precisely with established cytopathological knowledge, where nuclear contour irregularity (concave points) and size enlargement (radius, perimeter) are hallmark features of malignancy [16]. The consistency between model attributions and clinical pathology criteria enhances physician trust and facilitates biological validation of predictions [9,10].

Local SHAP explanations illustrate patient-specific traceability, showing how individual feature values contribute additively to the final logit score. This level of granularity enables clinicians to identify the dominant morphological driver for each case, supporting targeted diagnostic review and personalized treatment planning [31].

Comparison with previously reported models on the same dataset indicates that the proposed stacking ensemble achieves higher accuracy and ROC-AUC than recent ensemble approaches [1–3]. The integration of SHAP-based interpretability and explicit calibration evaluation distinguishes this work from prior studies that focused primarily on discriminative performance [32].

Predictive probabilities produced by the calibrated stacking ensemble can support threshold-based clinical decision strategies. In practical screening or diagnostic workflows, different probability thresholds may be selected depending on the desired balance between sensitivity and specificity. For instance, lower thresholds may be adopted in screening settings to prioritize sensitivity and reduce the likelihood of missed malignant cases, whereas higher thresholds may be used in confirmatory contexts to minimize unnecessary follow-up procedures. The SHAP-based interpretability framework further enhances clinical usability by identifying the morphological features that most strongly influence each prediction. Features such as concave points_worst, radius_worst, and perimeter_worst consistently exhibit strong contributions to malignancy risk estimates, enabling clinicians to understand which measurements are driving the model's assessment. Future work will include formal decision-curve analysis to quantitatively evaluate clinical net benefit across varying probability thresholds.

From an ethical and generalizability perspective, it is important to note that the proposed framework was evaluated exclusively on the Wisconsin Breast Cancer (Diagnostic) dataset, which is a single-center dataset of moderate size and does not include detailed demographic information. As a result, the demographic diversity of the underlying population cannot be fully characterized, and potential dataset-level biases may not be explicitly captured. Although stratified sampling, robust cross-validation, and careful handling of class imbalance were employed to reduce technical sources of bias and obtain unbiased performance estimates, these measures do not eliminate the risk of population shift when deploying the model in institutions with different patient characteristics or practice patterns. Future work will therefore involve external validation on larger and more diverse cohorts, subgroup-specific performance analyses where demographic metadata are available, and implementation of model monitoring strategies to detect population shift and model drift over time, thereby supporting ethically robust and clinically equitable deployment.

The system's limitations include reliance on a single-center dataset with a moderate sample size ($N = 569$) and the absence of external validation on independent cohorts. Morphological features derived from FNA cytology may not fully capture heterogeneity observed in whole-slide images or mammography [15]. Future extensions could incorporate multi-modal data (e.g., radiographic and genomic features) and prospective validation in multi-center clinical settings to assess generalizability [25–26].

In conclusion, the proposed hybrid fusion measurement framework advances the field by delivering not only competitive predictive performance but also calibrated probabilities and biologically grounded explanations. These attributes position the system as a clinically viable digital measurement tool for breast cancer malignancy quantification, with potential to enhance diagnostic confidence and support evidence-based decision-making in oncology practice.

A systematic ablation analysis was beyond the primary scope of the present work, which focused on the design and metrological evaluation of a clinically interpretable hybrid stacking framework. Nevertheless, the observed performance gains of the ensemble over individual base learners, together with its superior calibration, indicate that the meta-learning layer effectively integrates complementary predictive signals from heterogeneous models. Future work will include more extensive ablation experiments, such as alternative meta-learner configurations, leave-one-model-out analyses to quantify the contribution of each base model, and robustness assessments under controlled feature perturbations or simulated measurement noise. Such investigations will provide a more granular understanding of model stability and further validate the reliability of the proposed clinical decision-support system.

Furthermore, to facilitate reproducibility and independent verification of our findings, we have provided detailed implementation and training configuration information, including fixed random seeds and hyperparameter settings for all base learners and the stacking meta-model.

The computational efficiency of the proposed framework, particularly inference time and SHAP explanation latency, is a critical aspect for clinical deployment. While the current study focused on demonstrating performance, calibration, and interpretability, the chosen base learners and the efficient SHAP implementation on a fixed set of nine features suggest a reasonable computational overhead in practice. However, formal reporting of average inference times for both prediction and explanation generation, along with a comprehensive assessment of computational cost and scalability under different settings, would be valuable for practical applications. Future research will address these considerations through quantitative benchmarking across diverse scenarios and potentially larger datasets, to better characterize the framework's feasibility for real-time clinical use.

6. Conclusion

This study successfully establishes a hybrid digital measurement framework for precise breast cancer malignancy quantification, demonstrating the effectiveness of fusing four complementary base learners—Logistic Regression, Random Forest, Support Vector Machine, and XGBoost—through a stacking ensemble with a logistic regression meta-learner. The integration creates a unified system that harnesses diverse algorithmic strengths to generate a single, calibrated malignancy risk estimate, showcasing robust performance and metrological consistency across all evaluation dimensions.

The stacking ensemble exhibits superior performance over individual base learners, with test set accuracy of 0.986 and ROC-AUC of 0.998 compared to the best base model (accuracy 0.979, ROC-AUC 0.996). This enhancement, along with a 50% reduction in false-negative malignant cases (from 2 to 1), improves diagnostic reliability and minimizes critical errors in clinical screening contexts.

SHAP interpretability plays a pivotal role in facilitating clinical acceptance, transitioning from black-box models to transparent decision-making. By providing exact feature contributions (e.g., mean $|\text{SHAP}| = 0.421$ for concave points_worst as the dominant biomarker), the system enables physicians to validate predictions against cytopathological knowledge. This finding aligns with medical emphasis on nuclear contour irregularity as a key malignancy indicator, fostering trust and evidence-based adoption.

Uncertainty reduction is evidenced by excellent calibration metrics, with ECE = 0.017 and Brier Score = 0.042, far lower than base models. These indicators confirm the system's ability to deliver reliable probability estimates, addressing metrological challenges in pathological assessment where overconfidence can lead to suboptimal outcomes.

The framework effectively counters inter-observer variability—a major challenge in pathology—by offering standardized, objective measurements based on quantifiable nuclear features, potentially reducing diagnostic discordance reported in up to 25% of cases.

With its computational efficiency and interpretability, the system holds strong potential for deployment in computer-aided diagnosis (CAD) systems.

A primary limitation of the current study is the reliance on the Wisconsin Breast Cancer Diagnostic dataset for evaluation. While this dataset is a gold standard for benchmarking, the lack of external validation on multi-center clinical cohorts (such as TCGA or TCIA) limits the immediate generalizability of the findings. Consequently, the proposed framework is best suited for research-intensive environments and experimental diagnostic settings. Future research will prioritize cross-dataset validation to transition the model from a high-performance research tool to a universally ready clinical application. Future extensions could incorporate multi-modal data and multi-center validation to integrate it into hospital workflows, enhancing precision oncology through scalable, real-time malignancy quantification, transforming breast cancer diagnostics from subjective interpretation into a precise, standardized digital measurement science.

In conclusion, the proposed hybrid model offers a transparent and

Table 14
Nomenclature and mathematical components of the framework.

Symbol	Description	Unit / Scale
$P(y = 1\|X)$	Predicted probability of malignancy given input features X	Dimensionless (0 to 1)
z	Logit (linear combination of features and weights)	Dimensionless
β_0	Intercept or Bias term in Logistic Regression	Dimensionless
β_j	Coefficient (weight) associated with the j^{th} morphological feature	Inverse of feature unit
X_j	j^{th} input feature (e.g., radius, texture, perimeter)	Physical units (μm, etc.) or Scaled [0, 1]
w	Weight vector defining the orientation of the SVM hyperplane	Dimensionless
C	Penalty parameter for the error term (Regularization)	Positive scalar
ξ_i	Slack variables for handling non-linearly separable data	Dimensionless
p_i	Probability of an element belonging to a specific class (Gini calculation)	Dimensionless (0 to 1)
	Objective loss function for Gradient Boosting	Dimensionless
$\Omega(f)$	Regularization term to penalize model complexity (tree depth/ leaves)	Dimensionless
γ, λ	Hyperparameters controlling the complexity of trees in XGBoost	Dimensionless
T	Number of leaves in a specific decision tree	Integer
x_{new}	Synthetically generated sample in the SMOTE algorithm	Units of feature space
x_i	Original sample from the minority (malignant) class	Units of feature space
δ	Random interpolation factor for synthetic sample generation	Dimensionless (0 to 1)
x_{zi}	Nearest neighbor selected from the minority class	Units of feature space
$Y_{stacking}$	Final aggregated prediction of the Fusion Stacking Ensemble	Dimensionless (0 to 1)
σ	Activation function (Sigmoid) of the Meta-Learner	Mathematical operator
α_k	Weight assigned to the k^{th} base learner by the Meta-Learner	Dimensionless
$\widehat{y_k}$	Individual prediction (output probability) of the k^{th} base learner	Dimensionless (0 to 1)
$g(z')$	Local explanation model (SHAP)	Model output unit
ϕ_0	Base value (expected value) of the model output	Model output unit
ϕ_i	Shapley value (contribution of feature i to the prediction)	Model output unit
z'_j	Simplified input (binary coalition) representing feature presence	Binary (0 or 1)
M	Total number of input features	Integer
S	Subset of features used in the coalition game	Set of indices
$f(S)$	Marginal contribution of a feature subset	Model output unit
ECE	Expected Calibration Error	Dimensionless (0 to 1)
B	Total number of bins in the reliability diagram	Integer
n_m	Number of samples assigned to the m^{th} bin	Integer
$acc(B_m)$	Observed accuracy within the m^{th} bin	Percentage / Ratio
$conf(B_m)$	Mean predicted confidence within the m^{th} bin	Dimensionless (0 to 1)
BS	Brier Score (Mean Squared Error of probability predictions)	Dimensionless (0 to 1)
TP, TN	True Positives and True Negatives (Correct classifications)	Integer (Count)
FP, FN	False Positives and False Negatives (Classification errors)	Integer (Count)

accurate approach to malignancy quantification. While it demonstrates significant promise for clinical integration, it currently stands as a validated research framework, with future work directed toward large-scale external validation to ensure robust real-world applicability.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used AI-assisted language tools, specifically Grammarly, solely for grammar checking, language editing, and readability improvement. All AI-assisted suggestions were carefully reviewed, validated, and, where necessary, modified by the authors. The authors take full responsibility for the final content, scientific accuracy, and integrity of the manuscript (Table 14).

CRediT authorship contribution statement

Alireza Yahhyaei: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Mohammad Arabian:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World Health Organization, Breast Cancer Fact Sheet, World Health Organization Official Publications, Geneva, Switzerland, 2023.
[2] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2023,, CA: Cancer J. Clin. (2023) 17–48.
[3] J.G. Elmore, G.M. Longton, P.A. Carney, et al., Diagnostic concordance among pathologists interpreting breast biopsy specimens, J. Am. Med. Assoc. (JAMA) (2015).
[4] M. Tafazzoli, E. Daneshmand, B. Yazdani, et al., Deep learning for breast cancer histopathological diagnosis: classification and gene expression review, Netw. Model. Anal. Health Inform. Bioinform. (2024).
[5] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology (2016).
[6] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. (2019) 44–56.
[7] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, et al., A survey on deep learning in medical image analysis, Med. Image Anal. (2017).
[8] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, D. Shen, Interactive computer-aided diagnosis on medical image using large language models, Commun. Eng. (2024).
[9] N. Amaya-Tejera, M. Gamarra, J.I. Vélez, E. Zurek, A distance-based kernel for classification via support vector machines, Front. Artif. Intell. (2024).
[10] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States, 2016.
[11] Z.C. Lipton, The mythos of model interpretability, Commun. ACM (2018).
[12] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," Nat. Mach. Intell., Nature Publishing Group, London, United Kingdom, 2019.
[13] A. Tonekaboni, S. Joshi, D. McCradden, et al., What clinicians want: contextualizing explainable machine learning for clinical end use,, in: Proceedings of the 4th Machine Learning for Healthcare Conference, Ann Arbor, Michigan, United States, 2019.
[14] X. Li, Y. Zheng, X. Liu, Q. Tian, A comprehensive survey on ensemble deep learning: opportunities and challenges, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[15] A. Smith, J. Doe, M. Zhang, K. Lee, Stacked regression: a comprehensive review, insights, and future directions, J. Mach. Learn. Res. (2023).
[16] The Cancer Genome Atlas Program (TCGA), "TCGA-BRCA: breast invasive carcinoma dataset,"in *National Cancer Institute Genomic Data Commons (GDC)*, Bethesda, MD, United States,2021.
[17] M. Hussain, S. Khan, A. Hussain, F. Aljameel, Deep learning for breast cancer diagnosis using histopathology images: a comprehensive review, IEEE Rev. Biomed. Eng. (2022).
[18] S. Gupta, R. Tiwari, A. Singh, A survey on cross-validation procedures for model selection in machine learning, J. Big Data Anal. (2023).
[19] S.M. Lundberg, S. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[20] A. Holzinger, et al., "What do we need to build explainable AI systems for the medical domain?" arXiv Preprint, Cornell University, Ithaca, New York, United States, 2017.
[21] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? Explaining the predictions of any classifier,, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States, 2016.
[22] J. Zhang, L. Wang, H. Li, M. Zhao, Advances in random forests: a comprehensive review, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[23] S.M. Ganie, P.K.D. Pramanik, Z. Zhao, Enhanced and interpretable prediction of multiple cancer types using a stacking ensemble approach with SHAP analysis, Bioengineering (2025).
[24] Y. Lee, X. Chen, J. Wang, et al., A machine learning model for predicting breast cancer recurrence and supporting personalized treatment decisions through comprehensive feature selection and explainable ensemble learning, J. Med. Syst. (2025).
[25] W. Wolak, A. Plichta, H. Orlicki, Interpretable ensemble learning for tumor-type prediction with a SHAP-based evaluation of CatBoost and voting classifiers, Sci. Rep. (2025).
[26] X. Li, M. Gao, C. Zhang, G. Ma, Q. Zhang, W. Meng, T. Yuan, Y. Wang, Z. Li, A robust stacked neural network approach for early and accurate breast cancer diagnosis, Front. Med. (2025).
[27] T. Yasmeen, M. Ali, M.U. Hashmi, M.A. Hashmi, Z. Mehmood, A comparative study of advanced machine learning ensemble techniques for classification of breast cancer, J. Comput. Biomed. Inform. (2024).
[28] S. Pintawong, S. Shuangshoti, T. Jitpasutham, S. Shuangsot, K. Wiwatwong, T. H. Chalidabhongse, Conformal prediction for uncertainty quantification and reliable HER2 status classification in breast cancer IHC images, IEEE Access (2025).
[29] A.N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction, Found. Trends Mach. Learn. (2023).
[30] E. Serrano, S. Torre, C. Rodríguez, et al., Conformal prediction in clinical medical sciences: methods and applications, J. Healthc. Inform. Res. (2022).
[31] J. Chen, et al., Deep learning-based model for detecting COVID-19 pneumonia on CT, Sci. Rep. (2020) 1–11.
[32] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017.
[33] G. Pleiss, et al., On fairness and calibration, Adv. Neural Inf. Process. Syst. (2017).
[34] R. Walsh, M. Tardy, A comparison of techniques for class imbalance in deep learning classification of breast cancer, Diagnostics (2023).
[35] M. Ma, W. Xu, J. Yang, B. Zheng, C. Wen, S. Wang, Z. Xu, G. Qin, W. Chen, Contrast-enhanced mammography-based interpretable machine learning model for the prediction of the molecular subtype breast cancers, BMC Med. Imaging (2025).
[36] R. Ronal Watrianthos, E. Asri, Y. Humaira, Threshold-optimized and calibrated logistic regression for breast cancer classification, Salud Cienc. Tecnol. (2025).
[37] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. (2019).
[38] S. Kourou, et al., Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. (2015) 8–17.
[39] J. Amraei, A. Mirzapoor, K. Motarjem, M. Abdolahad, M. Baei, A. Ansari, Enhancing breast cancer diagnosis through machine learning algorithms, Sci. Rep. (2025).
[40] Q. Zhang, H. Gao, W. Li, Z. Xu, T. Ouyang, Z. Gu, Enhanced nuclear information fusion and visual transformer for pathological breast cancer image classification, Sci. Rep. (2025).
[41] S. Saharan, N.A. Wani, S. Chatterji, N. Kumar, A.M. Almuhaideb, A deep learning and explainable artificial intelligence based scheme for breast cancer detection, Sci. Rep. (2025).
[42] A. Khan, M. Alhussein, S. Khan, A. Rehman, Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations, Biomed. Signal Process. Control (2024).
[43] M. Dutta, K.M.M. Hasan, A. Akter, M.H. Rahman, M. Assaduzzaman, An interpretable machine learning-based breast cancer classification using XGBoost, SHAP, and LIME, Bull. Electr. Eng. Inform. (2025).
[44] T. Arravalli, K. Chadaga, H. Muralikrishna, N. Sampathila, D. Cenitta, R. Chadaga, K.S. Swathi, Detection of breast cancer using machine learning and explainable artificial intelligence, Sci. Rep. (2025).
[45] A. Ghasemi, S. Hashtarkhani, D.L. Schwartz, A. Shaban-Nejad, Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review, Cancer Innov. (2024).
[46] "Breast cancer molecular subtype prediction: Improving interpretability of complex machine-learning models based on multiparametric-MRI features using SHapley Additive exPlanations (SHAP) methodology ", ELSEVIER, 2024.
[47] C.C. Ukwuoma, D. Cai, E.O. Eziefuna, A. Oluwasanmi, S.F. Abdi, G.W. Muoka, D. Thomas, K. Sarpong, Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI – LIME & SHAP, Biocybern. Biomed. Eng. (2024).
[48] J. Lin, L. Tao, M. Dong, C. Xu, Uncertainty weighted gradients for model calibration. CVPR, 2024.
[49] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, " A review of uncertainty quantification in deep learning: Techniques, applications and challenges", ElSEVIER, 2021.

[50] A. Nadig, M. Brown, L.A. Celi, L. Oakden-Rayner, M. Sendak, R.S. Bhatia, K. Jung, C. Daluwatte, Monitoring performance of clinical artificial intelligence in health care: a scoping review, JBI Evid. Synth. (2024).
[51] S.M. Lauritsen, M. Kristensen, M.V. Olsen, M.S. Larsen, K.M. Lauritsen, M. J. Jørgensen, J. Lange, Thiesson, "Explainable artificial intelligence model to predict acute critical illness from electronic health records", Nat. Commun. (2020).
[52] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology, Annu. Rev. Biomed. Eng. (2016).
[53] D. Gutman, et al., Digital pathology reproducibility and uncertainty sources, J. Pathol. Inform. (2017) 1–14.
[54] P. Bankhead, et al., QuPath: open source software for digital pathology image analysis, Sci. Rep. (2017).
[55] E. Sreehari, L.D. DhineshBabu, A novel aggregated coefficient ranking based feature selection strategy for enhancing the diagnosis of breast cancer classification using machine learning, Sci. Rep. (2025).
[56] **Author(s) unspecified**, Machine Learning and AI Innovations with Python, Int. J. Full Stack Mach. Res. (2024).
[57] S.M. Lundberg, S. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017).
[58] S.M. Lundberg, et al., "Consistent individualized feature attribution for tree ensembles," arXiv Preprint, Cornell University, Ithaca, New York, United States, 2018.
[59] J.M. Lipton, D. McInerney, Understanding black-box predictions via influence functions,, in: ICML Workshop on Human Interpretability in Machine Learning, Stockholm, Sweden, 2018.
[60] R. Islam, M. Tarique, Artificial intelligence (AI) and nuclear features from the fine needle aspirated (FNA) tissue samples to recognize breast cancer, J. Imaging (2024).
[61] United States Department of Health and Human Services, "Final rule: federal policy for the protection of human subjects (revised common rule), 45 CFR Part 46," Washington, DC, United States, 2024.
[62] M. Tatman, et al., Reproducibility in machine learning for health research, Patterns (2021).
[63] S. Tatman, J. VanderPlas, S. Dane, Practical challenges for reproducibility in machine learning research, Patterns (2021).
[64] R. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016.
[65] S. Elmore, G. Longton, P. Carney, A. Geller, T. Onega, H. Tosteson, et al., Diagnostic concordance among pathologists interpreting breast biopsy specimens, J. Am. Med. Assoc. (JAMA) (2015).
[66] M. Xiao, Y. Li, X. Yan, M. Gao, W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," in arXiv Preprint., Cambridge, Massachusetts, United States, 2024.
[67] R. Islam, M. Tarique, Artificial intelligence (AI) and Nuclear features from the fine needle aspirated (FNA) tissue samples to recognize breast cancer, J. Imaging (2024).
[68] P. Liu, Comparative analysis of machine learning models in breast cancer diagnosis, Appl. Comput. Eng. (2024).
[69] Kaggle, "Breast cancer Wisconsin (diagnostic) dataset," in Kaggle Datasets Repository, San Francisco, CA, USA, 2016.
[70] U.S. Department of Health & Human Services, Federal Policy for the Protection of Human Subjects (Common Rule), 45 CFR 46, U.S. Federal Regulations, Washington, DC, USA, 2018.
[71] M. Tatman, J. VanderPlas, S. Dane, Reproducibility in machine learning for health research, Patterns (2021).
[72] *Mnemoraorg*, "Breast cancer Wisconsin (diagnostic) Dataset," in Huggingface Datasets, San Francisco, CA, United States, 2025.
[73] R. Shafique, F. Rustam, G.S. Choi, I. de la Torre Díez, A. Mahmood, V. Lipari, C.L. R. Velasco, I. Ashraf, Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning, Cancers (2023).
[74] S. Tatman, J. VanderPlas, S. Dane, Practical challenges for reproducibility in machine learning research, Patterns (2021).
[75] A. Khan, M. Alhussein, S. Khan, A. Rehman, Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations, Biomed. Signal Process. Control (2024).
[76] M.A. Alsharif, S. Khan, A. Khan, M. Alhussein, A deep learning and explainable artificial intelligence based scheme for breast cancer detection, Biomed. Signal Process. Control (2023).