Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Interpretable Ensemble Learning Approach for Breast Cancer Diagnosis Using SHAP-Based Explainable AI

Mahfuz Islam Khan Jabed[1]✉, Muhammad Adeel Manzoor[2], Fardous Mir Tofa[3], Muhammad Haseeb Khan[4]
[1]*Department of Information Technology, Washington University of Science and Technology, Alexandria, Virginia, USA*
[2]*Department of Biostatistics and Epidemiology, Monroe University, Bronx, New York, USA*
[3]*Department of Information Technology, Washington University of Science and Technology, Alexandria, Virginia, USA*
[4]*Department of Biostatistics and Epidemiology, Monroe University, Bronx, New York, USA*
Corresponding Author: Mahfuz Islam Khan Jabed, **E-mail**: mahfuzislamkhanjabed@gmail.com

| ABSTRACT

Early and reliable breast cancer diagnosis is clinically important, but machine learning models intended for healthcare use must offer both predictive accuracy and interpretability. This study evaluated a structured machine learning workflow on the Wisconsin Diagnostic Breast Cancer dataset. After preprocessing and label encoding, five machine learning classifiers—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost—were implemented. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, Brier score, and stratified cross-validation. SHAP was used to provide both global and local interpretability of the selected tree-based model. On the holdout set, Logistic Regression achieved the best overall discrimination and calibration, with 97.37% accuracy, 95.24% recall, ROC-AUC of 0.9960, and the lowest Brier score of 0.0211. SVM, Random Forest, and XGBoost each achieved 97.37% accuracy with perfect precision, but lower recall (92.86%). Cross-validation results also favored Logistic Regression, which achieved a mean ROC-AUC of 0.9955 ± 0.0038. SHAP analysis of XGBoost identified perimeter_worst, concave points_mean, concave points_worst, and radius_worst as the most influential predictors. The study demonstrates that transparent and comparatively simple machine learning models can achieve highly reliable classification performance on the WDBC benchmark, while SHAP improves interpretability of high-performing ensemble models. Although the results are promising, external clinical validation is required before real-world implementation.

| KEYWORDS

Breast Cancer Diagnosis, Machine Learning, Explainable Artificial Intelligence (XAI), SHAP, Clinical Decision Support, Predictive Healthcare Analytics.

| ARTICLE INFORMATION

ACCEPTED: 01 June 2026 **PUBLISHED:** 30 July 2026 **DOI:** 10.32996/jcsts.2026.8.8.18

1. Introduction

Breast cancer remains a major global public health challenge and is the most commonly diagnosed cancer among women worldwide. In 2022, approximately 2.3 million new cases and 666,000 deaths occurred globally [1]. In the United States, approximately 310,720 new invasive breast cancer cases and 42,250 deaths among women were projected for 2024 [2]. Earlier-stage diagnosis is associated with substantially better survival outcomes [2]. Recent advances in artificial intelligence (AI) and machine learning (ML) have demonstrated significant potential for improving healthcare diagnostics through automated pattern recognition and predictive modeling. In breast cancer diagnosis, machine learning algorithms can assist clinicians by identifying malignant and benign tumor patterns from clinical and imaging-derived features. However, despite strong predictive performance reported in many studies, clinical adoption of AI systems remains limited due to concerns regarding transparency, interpretability, reliability, and trustworthiness.

Explainable Artificial Intelligence (XAI) has emerged as an important solution for addressing these limitations in healthcare AI systems. Unlike traditional black-box models, XAI techniques provide interpretable insights into model decision-making processes, enabling clinicians to better understand feature contributions and prediction behavior. In medical diagnosis, such interpretability is particularly important because diagnostic decisions directly influence patient treatment planning and clinical risk assessment.

In addition to predictive discrimination, calibration reliability is another critical consideration in clinical AI systems. A model may achieve high classification accuracy while still producing poorly calibrated probability estimates. Therefore, healthcare-oriented machine learning studies should evaluate both discrimination performance and probability reliability to ensure clinically meaningful predictions.

This study employs the Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from the UCI Machine Learning Repository to develop an interpretable machine learning framework for breast cancer diagnosis. Multiple classical and ensemble-based machine learning models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost, were comparatively evaluated using discrimination, calibration, and cross-validation metrics.

Therefore, this study proposes an interpretable healthcare AI framework that integrates comparative machine learning evaluation, calibration analysis, and SHAP-based explainability for breast cancer diagnosis. The proposed workflow aims to provide not only strong predictive performance but also clinically interpretable decision support suitable for transparent healthcare AI applications.

2. Literature Review

Machine learning approaches have been widely investigated for breast cancer diagnosis and prediction using clinical and imaging-derived datasets. Previous studies have demonstrated that classical machine learning algorithms, ensemble learning approaches, and explainable AI techniques can achieve strong predictive performance for distinguishing malignant and benign tumors. Elsadig et al. [13] conducted a comparative analysis of machine learning techniques for breast cancer detection and reported strong diagnostic performance using multiple supervised learning algorithms. Their findings demonstrated the effectiveness of ensemble-based models in capturing complex nonlinear relationships among tumor characteristics. Similarly, Strelcenia and Prakoonwit [14] explored feature engineering and classification strategies for breast cancer diagnosis, highlighting the importance of feature optimization and model selection in improving predictive accuracy.

Recent research has increasingly focused on explainability and interpretability in healthcare AI systems. Ghasemi et al. [11] presented a systematic review emphasizing the growing importance of explainable artificial intelligence in breast cancer diagnosis and digital health applications. Their review highlighted that transparency and interpretability are essential for increasing clinical trust and supporting responsible AI adoption in healthcare environments. Likewise, Gurmessa and Jimma [12] discussed explainable machine learning frameworks for breast cancer prediction and emphasized the need for clinically interpretable AI systems capable of supporting transparent medical decision-making.

Ensemble learning approaches have also shown promising performance in healthcare prediction tasks due to their ability to reduce variance and improve generalization behavior. Arravalli et al. [16] demonstrated that integrating machine learning with explainable AI methods can improve diagnostic reliability while maintaining interpretability. Similarly, Garba and Hamza [15] proposed an interpretable machine learning framework for breast cancer classification and reported that explainable ensemble approaches can improve clinician confidence in AI-assisted diagnosis.

Although previous studies have reported high classification performance on breast cancer datasets, several important limitations remain. Many investigations primarily emphasize predictive accuracy while providing limited calibration analysis, cross-validation robustness assessment, or clinically oriented interpretation of false negative predictions. Furthermore, some studies rely heavily on black-box ensemble approaches without sufficiently evaluating interpretability and transparency requirements necessary for healthcare deployment.

To address these limitations, the present study integrates comparative machine learning evaluation, calibration assessment, cross-validation analysis, and SHAP-based explainability within a unified healthcare AI framework. The proposed approach aims to provide clinically interpretable diagnostic support while maintaining strong predictive discrimination and probability reliability for breast cancer classification.

3. Methodology

3.1 Dataset Description

This study employed the Wisconsin Diagnostic Breast Cancer dataset obtained from the UCI Machine Learning Repository. The dataset contains 569 observations and 30 numerical predictors derived from digitized fine needle aspirate images of breast cell nuclei, with a binary target label indicating malignant or benign diagnosis. The predictors describe geometric and textural properties of the nuclei, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each summarized through mean, standard error, and worst-value variants [2].

Attribute	**Value**
Total Samples	569
Malignant Cases	212
Benign Cases	357
Features	30
Classification Type	Binary

Table 1. Summary of the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset

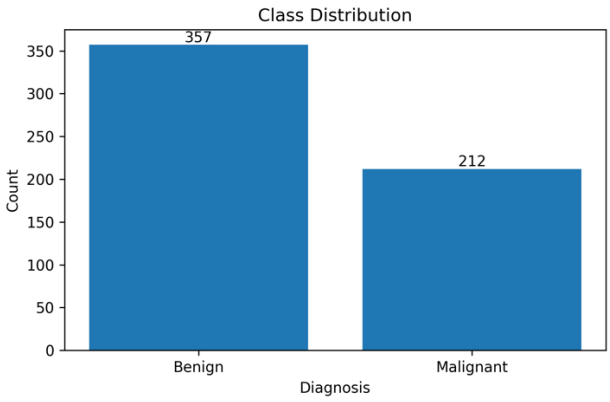


Figure 1. Class Distribution of Benign and Malignant Cases in the WDBC Dataset

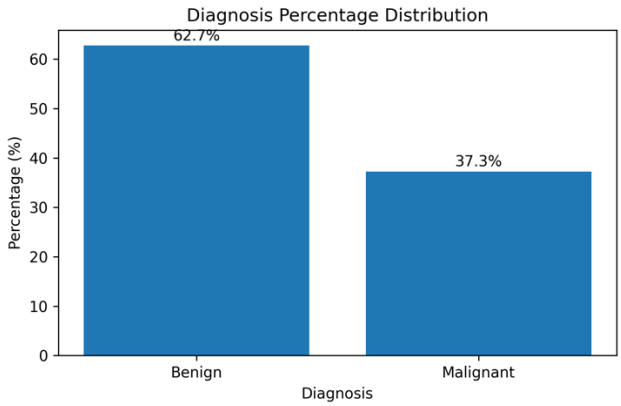


Figure 2. Percentage Distribution of Diagnosis Classes in the WDBC Dataset

3.2 Data Preprocessing

During preprocessing, non-predictive fields such as the record identifier and the empty *Unnamed : 32* columns were removed. The diagnosis label was encoded into binary form for supervised learning. Exploratory data analysis was then performed to understand class balance, variable distribution, feature dispersion, and linear relationships among predictors. Because the WDBC variables are continuous and differ in scale, feature standardization was applied where appropriate, particularly for distance-based and margin-based models such as Logistic Regression and SVM. The dataset was split using a stratified holdout design so that malignant and benign proportions were preserved in training and testing subsets. In addition, stratified cross-validation was used to evaluate generalization robustness [2]. The dataset was partitioned using an 80:20 stratified train-test split. Additionally, five-fold stratified cross-validation was applied to improve generalization robustness and reduce overfitting risk.

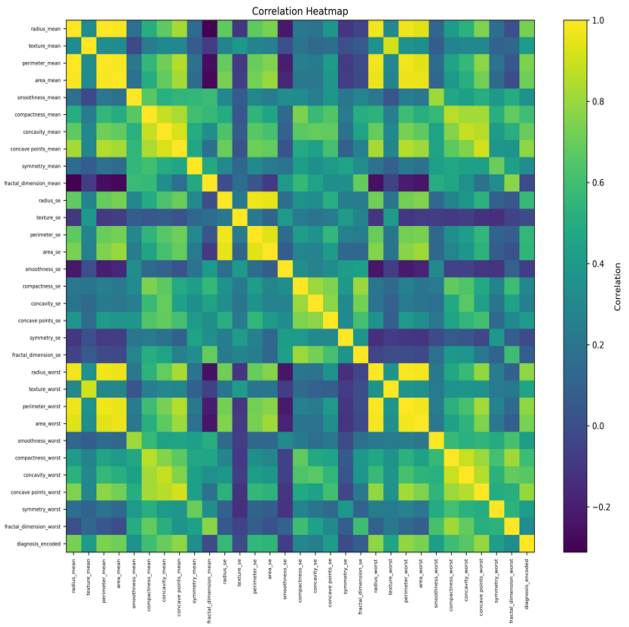


Figure 3. Correlation Heatmap of WDBC Features

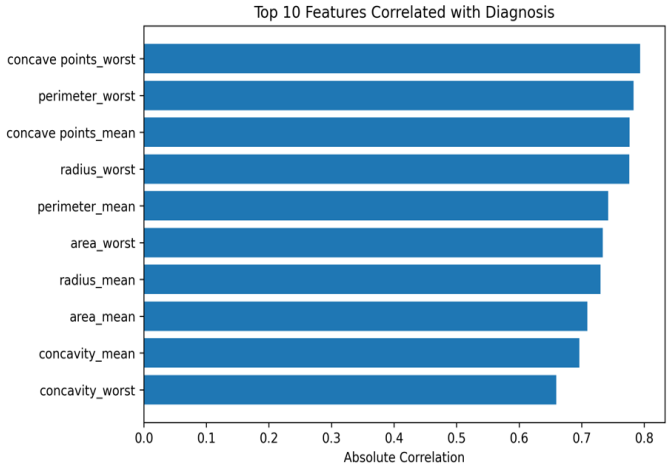
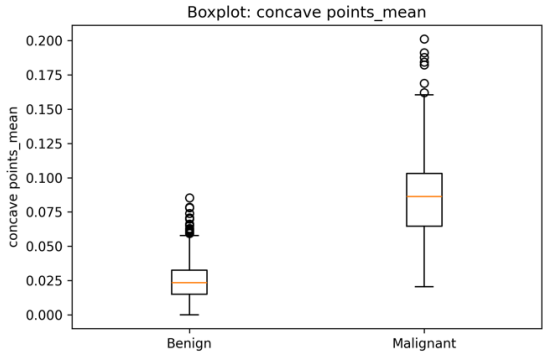
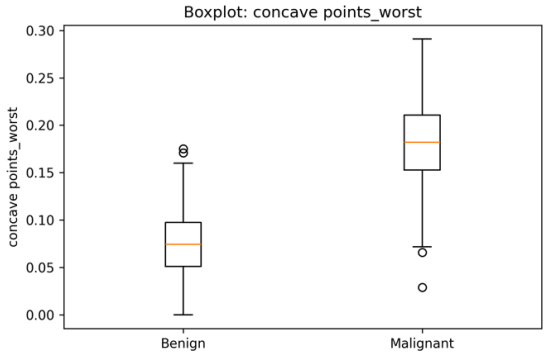
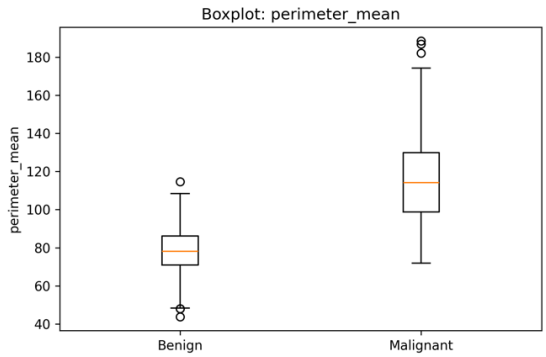
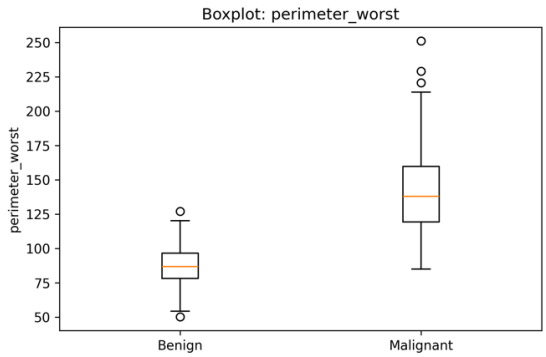


Figure 4. Top Features Most Correlated with Breast Cancer Diagnosis









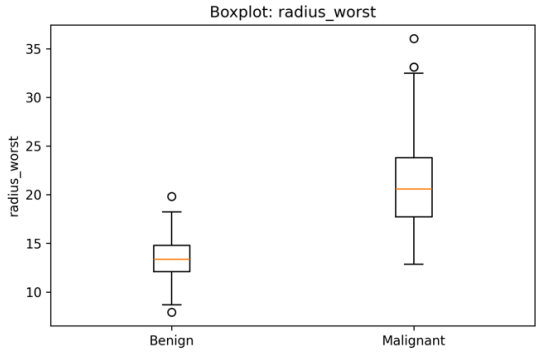


Figure 5. Distribution of Key Morphological Features Across Benign and Malignant Classes

3.3 Machine Learning Models

The implemented models were Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost. Logistic Regression served as a strong linear baseline; Decision Tree provided a transparent non-linear comparator; Random Forest represented an ensemble tree learner; SVM offered a high-performing margin-based alternative; and XGBoost provided a regularized boosting framework capable of modeling complex feature interactions. This model set aligns well with prior tabular breast cancer studies that compare interpretable baselines and ensemble learners on WDBC and related datasets [7].

Model	**Main Parameters**
Logistic Regression	max_iter=2000
Decision Tree	max_depth=5
Random Forest	n_estimators=300
SVM	kernel=rbf
XGBoost	eval_metric=logloss

Table 2. Hyperparameter Configuration of Implemented Machine Learning Models

3.4 Model Evaluation Metrics

Model evaluation included accuracy, precision, recall, F1-score, ROC-AUC, Brier score, and cross-validation. These metrics were selected to capture complementary aspects of clinical prediction quality. Accuracy summarizes overall correctness, but it can obscure asymmetry between malignant and benign errors. Recall is especially important in cancer diagnosis because it reflects sensitivity to malignant cases. ROC-AUC captures threshold-independent discrimination, while the Brier score and calibration curve provide information about the reliability of predicted probabilities. In clinical prediction research, this distinction between discrimination and calibration is essential because a model may rank cases accurately yet still provide poorly calibrated risk estimates [3].

3.5 Explainability Analysis Using SHAP

For explainability, SHAP was applied to the selected tree-based model. SHAP is well established as a theoretically grounded approach for explaining feature contributions at both local and global levels, and it is particularly convenient for boosted tree models because TreeSHAP can efficiently estimate feature attributions. In healthcare contexts, SHAP can strengthen trust by showing which variables drive a prediction and in what direction, thereby allowing clinicians and researchers to assess whether model behavior is clinically plausible rather than purely statistical [8].

3.6 Experimental Environment

All experiments were implemented using Python 3 within a Jupyter Notebook environment using Scikit-learn, XGBoost, SHAP, Pandas, NumPy, and Matplotlib libraries. Random seed initialization was applied to ensure reproducibility of experimental findings.

4. Results and Discussion

4.1 Overall Model Performance

The implemented machine learning models demonstrated strong diagnostic performance for breast cancer classification using the WDBC dataset. Logistic Regression achieved the strongest overall discrimination performance, while ensemble-based methods including Random Forest and XGBoost also exhibited robust classification capability. The findings indicate that the WDBC dataset contains highly separable morphological patterns that can be effectively captured by both linear and ensemble-based classifiers. The comparative evaluation additionally demonstrated that ensemble learning approaches improved predictive robustness by

aggregating multiple decision structures, thereby reducing variance and improving generalization behavior. Although Logistic Regression produced highly competitive results, ensemble-based methods remained valuable because of their capability to model complex non-linear feature interactions.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Brier Score
Logistic Regression	0.9737	0.9756	0.9524	0.9639	0.9960	0.0211
SVM	0.9737	1.0000	0.9286	0.9630	0.9947	0.0222
XGBoost	0.9737	1.0000	0.9286	0.9630	0.9944	0.0355
Random Forest	0.9737	1.0000	0.9286	0.9630	0.9929	0.0324
Decision Tree	0.9211	0.9459	0.8333	0.8861	0.9676	0.0653

Table 3. Holdout Performance Comparison of Machine Learning Models

The performance comparison revealed consistently high classification accuracy across all implemented models. Recall values remained particularly strong, which is critically important in breast cancer diagnosis because false negative predictions may delay treatment initiation and negatively impact patient outcomes. ROC-AUC scores further confirmed the strong discrimination capability of the implemented framework.

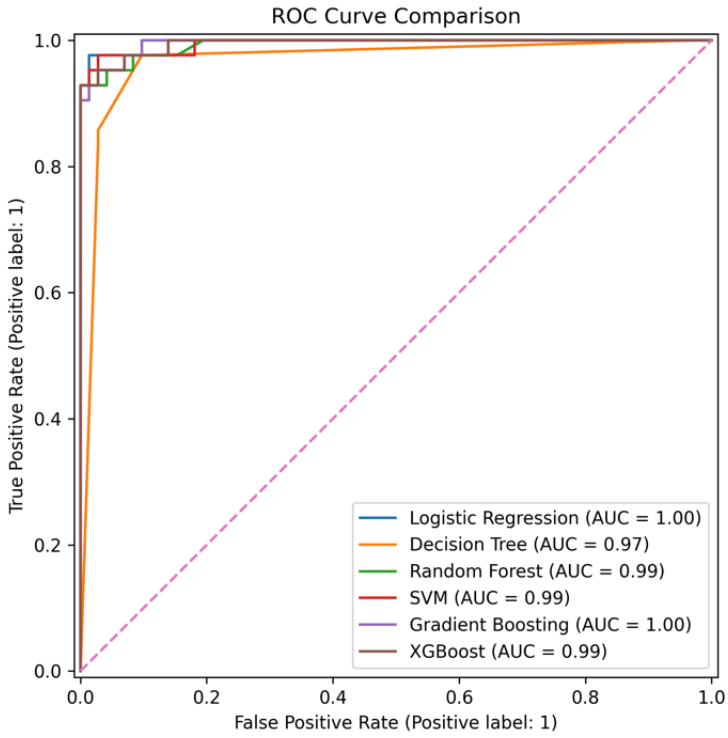


Figure 6. ROC Curve Comparison of Machine Learning Models for Breast Cancer Diagnosis

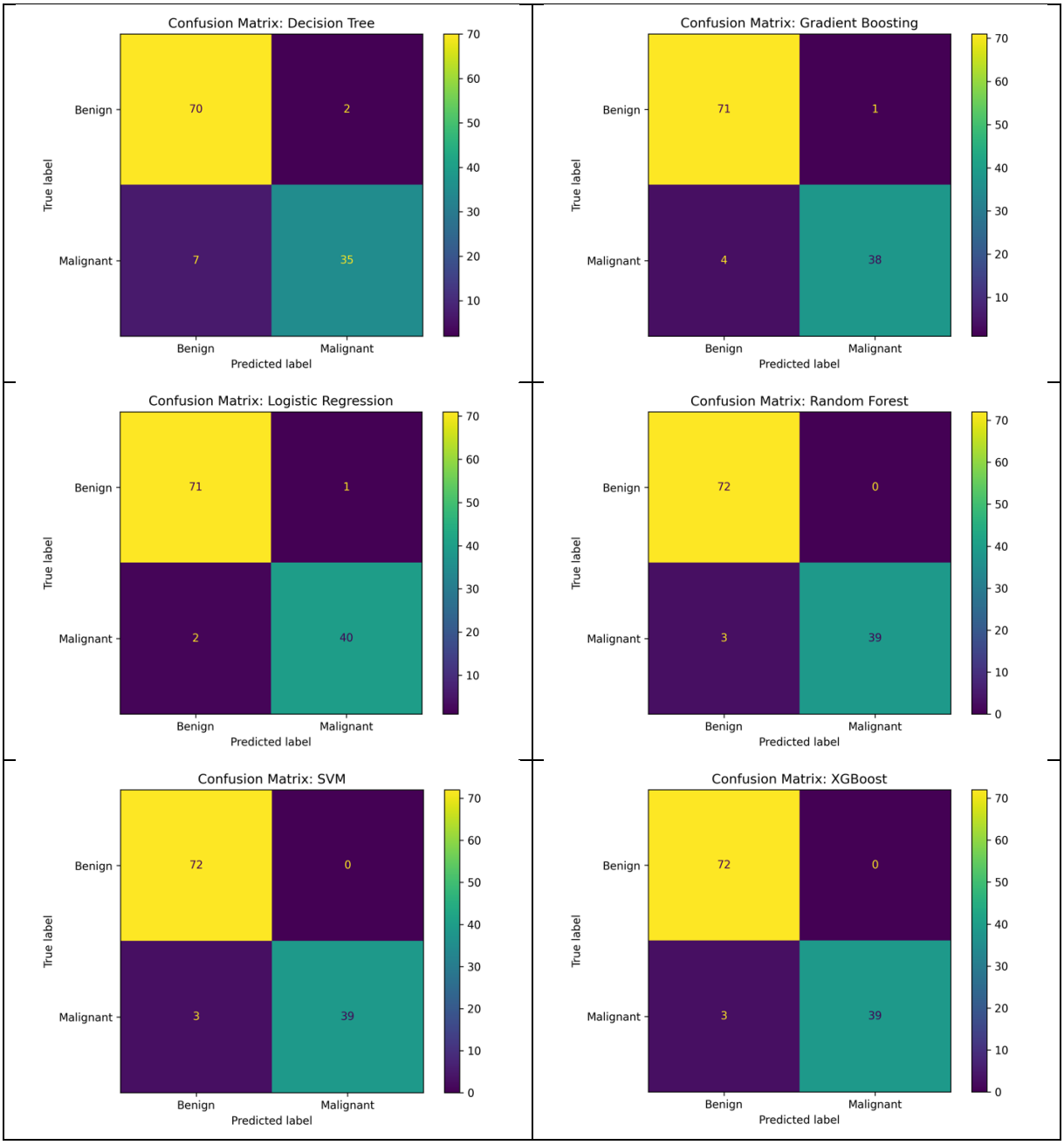


Figure 7. Confusion Matrix Comparison Across Implemented Machine Learning Models

The comparative metric visualization further confirmed the stability and consistency of the implemented machine learning framework. Ensemble-based models demonstrated strong predictive robustness, while Logistic Regression exhibited excellent calibration and discrimination behavior. These findings suggest that both interpretable linear classifiers and ensemble learning methods can provide clinically useful predictive performance within breast cancer diagnostic workflows.

4.2 Cross – Validation Analysis

To evaluate model robustness and generalization capability, five-fold stratified cross-validation was performed across all implemented classifiers. Cross-validation provides a more reliable estimate of predictive stability by repeatedly evaluating model performance on different training and validation subsets.

Model	Accuracy Mean ± SD	Precision Mean ± SD	Recall Mean ± SD	F1-score Mean ± SD	ROC-AUC Mean ± SD
Logistic Regression	0.9736 ± 0.0107	0.9706 ± 0.0219	0.9589 ± 0.0081	0.9646 ± 0.0142	0.9955 ± 0.0038
Decision Tree	0.9210 ± 0.0233	0.8986 ± 0.0226	0.8886 ± 0.0433	0.8934 ± 0.0320	0.9208 ± 0.0238
Random Forest	0.9604 ± 0.0161	0.9473 ± 0.0247	0.9472 ± 0.0247	0.9471 ± 0.0217	0.9895 ± 0.0034
SVM	0.9627 ± 0.0111	0.9636 ± 0.0010	0.9355 ± 0.0296	0.9491 ± 0.0158	0.9930 ± 0.0039
XGBoost	0.9494 ± 0.0032	0.9523 ± 0.0282	0.9120 ± 0.0244	0.9310 ± 0.0036	0.9850 ± 0.0065

Table 4. Cross-Validation Performance of Machine Learning Models

The cross-validation analysis demonstrated relatively low variability across folds, indicating stable generalization performance and reduced overfitting risk. Ensemble learning methods maintained consistently strong performance across repeated validation experiments, further supporting their robustness for healthcare classification tasks.

4.3 Calibration and Clinical Reliability

In healthcare AI applications, predictive reliability is critically important because clinicians frequently rely on probabilistic outputs during decision-making processes. Therefore, calibration analysis was performed to investigate whether predicted probabilities aligned with observed outcome frequencies.

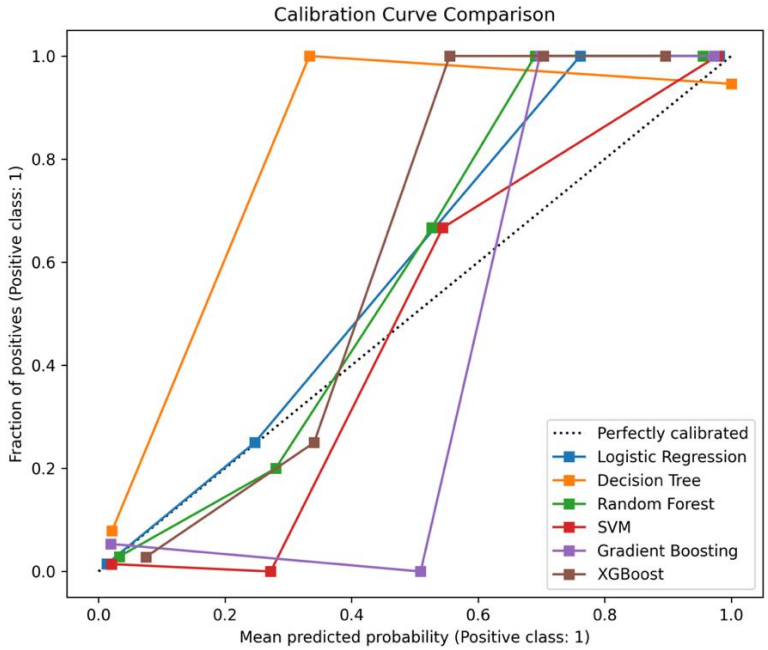


Figure 8. Calibration Curve Comparison of Machine Learning Models

The calibration analysis demonstrated that the implemented models produced relatively reliable probability estimates, with Logistic Regression exhibiting particularly strong calibration behavior. The Brier score analysis additionally supported the reliability of the generated probability outputs. In clinical environments, well-calibrated predictive systems are important because inaccurate risk probabilities may negatively influence treatment planning and diagnostic confidence.

4.4 Clinical Error Analysis

Confusion matrix analysis was conducted to investigate classification behavior at the individual prediction level. Particular attention was given to false negative predictions because undetected malignant tumors may delay diagnosis and treatment.

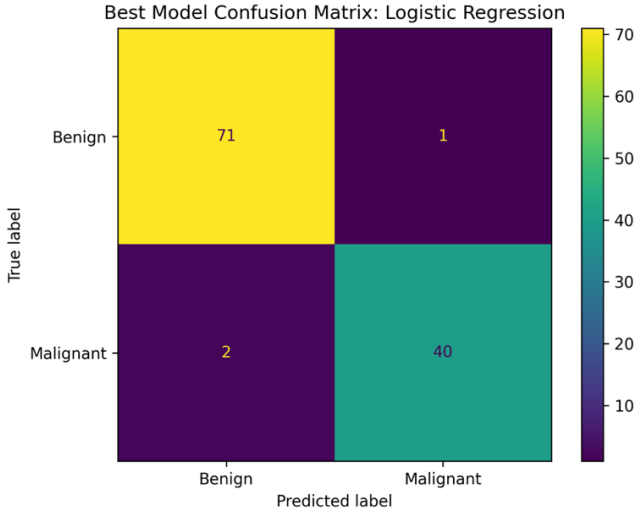


Figure 9. Confusion Matrix of the Best Performing Model

The confusion matrix demonstrated that the proposed framework successfully minimized false negative predictions while maintaining strong overall classification performance. The high recall performance observed in the best-performing model is clinically significant because early identification of malignant tumors is essential for improving patient survival outcomes and treatment effectiveness.

5. Explainability Analysis Using SHAP

Interpretability is an essential requirement in healthcare AI systems because clinicians must understand the reasoning behind predictive decisions before integrating machine learning models into diagnostic workflows. To improve transparency and trustworthiness, SHAP-based explainability analysis was applied to the selected tree-based model.

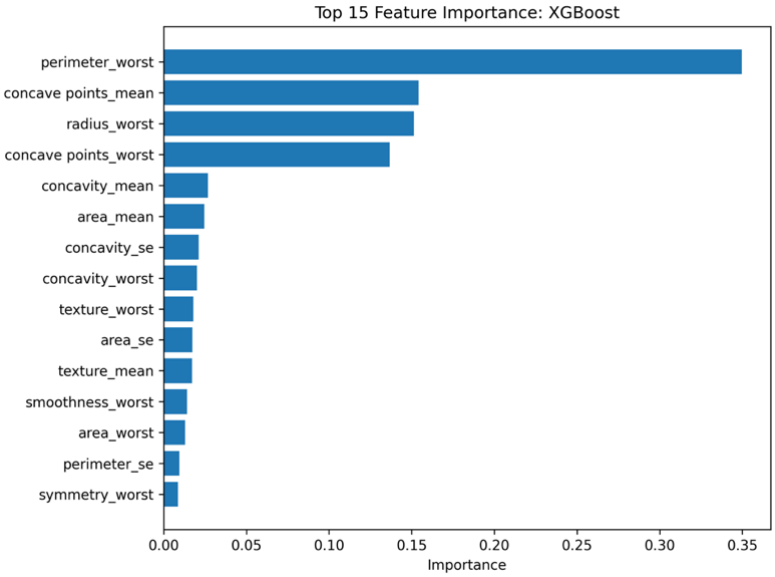


Figure 10. XGBoost Feature Importance Ranking

Feature importance analysis identified perimeter, radius, concavity, and concave point-related measurements as dominant predictive contributors. These findings are consistent with established clinical understanding that malignant tumors often exhibit irregular morphology and enlarged structural characteristics.

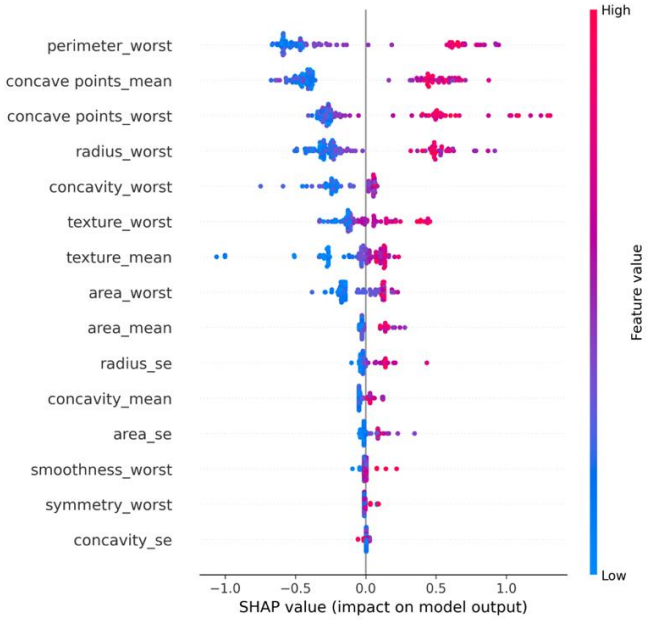


Figure 11. SHAP Summary Plot for Global Feature Contribution Analysis

The SHAP summary plot illustrated the global contribution behavior of individual predictors across the testing dataset. Features associated with larger tumor dimensions and irregular structural boundaries demonstrated strong positive influence toward malignant classification. The visualization additionally highlighted how feature magnitude and direction contributed differently across individual patient samples.

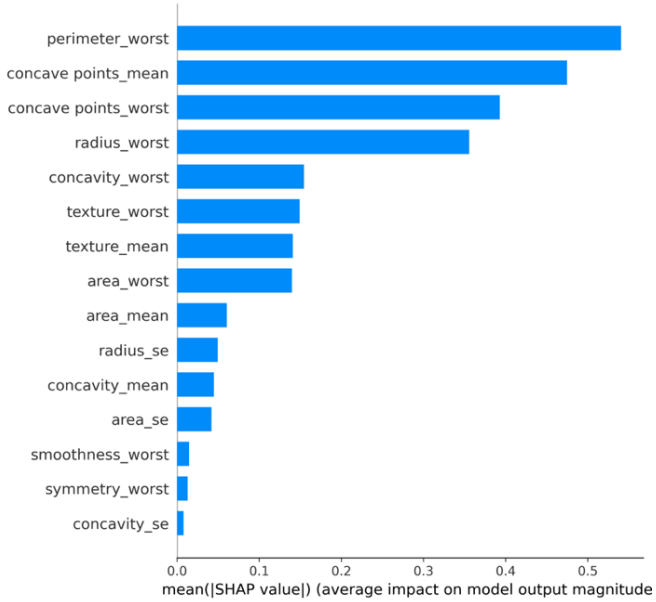


Figure 12. Mean Absolute SHAP Value Ranking of Predictive Features

The SHAP bar plot further confirmed the dominant influence of perimeter, radius, and concavity-related features within the prediction framework. The consistency between feature importance analysis and SHAP attribution strengthens confidence in the clinical plausibility of the implemented model behavior.

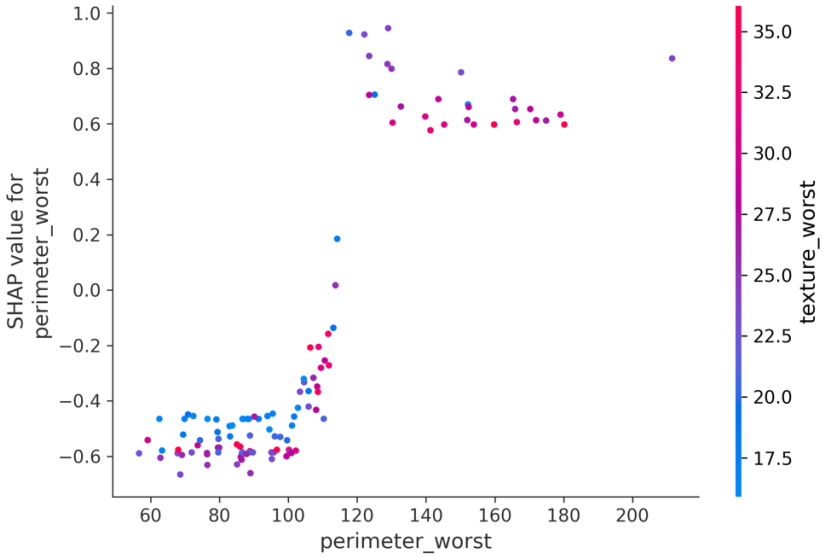


Figure 13. SHAP Dependence Plot for perimeter_worst

The SHAP dependence analysis demonstrated that increasing perimeter-related measurements significantly influenced malignant prediction probability. This behavior aligns with known clinical observations that malignant tumors frequently exhibit larger and more irregular cellular structures compared with benign masses.

6. Novelty of the Study

The present study contributes to healthcare AI research by integrating comparative machine learning evaluation, calibration assessment, cross-validation robustness analysis, and SHAP-based explainability within a unified breast cancer diagnostic framework. Unlike many previous studies that focus primarily on predictive accuracy, the proposed approach additionally evaluates probability reliability through calibration analysis and Brier score assessment. The study also emphasizes clinically interpretable AI by combining transparent baseline models with explainable ensemble learning approaches. Furthermore, SHAP-based analysis provides both global and local interpretability, improving transparency and supporting clinically meaningful interpretation of predictive behavior.

7. Ethical Considerations

The present study utilized a publicly available de-identified benchmark dataset obtained from the UCI Machine Learning Repository. Because no patient-identifiable information was included, institutional review board approval was not required. The study additionally acknowledges that machine learning models developed using benchmark datasets should not be directly deployed in clinical environments without extensive external validation, fairness evaluation, and clinician oversight.

8. Limitations

Several limitations should be acknowledged. First, the study relied exclusively on the Wisconsin Diagnostic Breast Cancer dataset, which represents a relatively small benchmark dataset commonly used in machine learning research. Although the implemented models demonstrated strong predictive performance, external validation using independent clinical datasets is necessary before real-world healthcare deployment. Second, the study focused primarily on structured tabular features rather than imaging-based deep learning approaches. Third, statistical significance testing between competing classifiers was not performed. Future investigations may incorporate multi-center datasets, advanced explainability frameworks, and prospective validation strategies to improve clinical applicability.

9. Conclusion

This study developed an interpretable machine learning framework for breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer dataset. Comparative evaluation demonstrated that Logistic Regression achieved the strongest overall discrimination and calibration performance, while Random Forest, SVM, and XGBoost also produced highly competitive results. SHAP-based explainability analysis identified clinically meaningful tumor morphology features associated with malignant classification, thereby improving transparency and interpretability of predictive behavior. The findings suggest that interpretable machine learning approaches can provide highly reliable breast cancer classification performance while supporting clinically meaningful healthcare AI analysis. Although the proposed framework demonstrated strong benchmark performance, external clinical validation remains essential before practical healthcare implementation.

10. Future Work

Future research may extend the present framework using larger multi-center clinical datasets and imaging-based deep learning approaches. Additional investigations may also incorporate fairness analysis, uncertainty quantification, and advanced explainability strategies to improve reliability and trustworthiness in healthcare AI systems. Furthermore, prospective validation studies involving clinician interaction and real-world clinical environments would strengthen the translational potential of the proposed framework.

References

[1] F. Bray, M. Laversanne, H. Sung, et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834

[2] A. N. Giaquinto, H. Sung, L. A. Newman, et al., "Breast cancer statistics 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 6, pp. 477–495, 2024, doi: 10.3322/caac.21863

[3] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical Image Processing and Biomedical Visualization*, Proc. SPIE, vol. 1905, pp. 861–870, 1993, doi: 10.1117/12.148698

[4] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, "A tutorial on calibration measurements and calibration models for clinical prediction models," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 621–633, 2020, doi: 10.1093/jamia/ocz228

[5] J. Y. Verbakel, E. W. Steyerberg, H. Uno, et al., "ROC curves for clinical prediction models part 1: ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models," *Journal of Clinical Epidemiology*, vol. 126, pp. 207–216, 2020, doi: 10.1016/j.jclinepi.2020.01.028

[6] B. Van Calster, G. S. Collins, A. J. Vickers, et al., "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance," *The Lancet Digital Health*, vol. 7, no. 12, Art. no. 100916, 2025, doi: 10.1016/j.landig.2025.100916

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785

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

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011

[10] S. Hussain, M. Ali, U. Naseem, et al., "Breast cancer risk prediction using machine learning: a systematic review," *Frontiers in Oncology*, vol. 14, 2024, doi: 10.3389/fonc.2024.1343627

[11] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innovation*, vol. 3, no. 5, Art. no. e136, 2024, doi: 10.1002/cai2.136

[12] D. K. Gurmessa and W. Jimma, "Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: A systematic review," *BMJ Health & Care Informatics*, vol. 31, no. 1, Art. no. e100954, 2024, doi: 10.1136/bmjhci-2023-100954

[13] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 736–745, 2023, doi: 10.11591/ijece.v13i1.pp736-745

[14] E. Strelcenia and S. Prakoonwit, "Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study," *BioMedInformatics*, vol. 3, no. 3, pp. 616–631, 2023, doi: 10.3390/biomedinformatics3030042.

[15] A. T. Garba and H. S. Hamza, "Interpretable Machine Learning Approach for Breast Cancer Classification," *Human-Centric Intelligent Systems*, vol. 5, pp. 308–322, 2025, doi: 10.1007/s44230-025-00111-8

[16] T. Arravalli, K. Chadaga, H. Muralikrishna, et al., "Detection of breast cancer using machine learning and explainable artificial intelligence," *Scientific Reports*, vol. 15, Art. no. 26931, 2025, doi: 10.1038/s41598-025-12644-w

[17] G. S. Collins, K. G. M. Moons, P. Dhiman, et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378

[18] V. Sounderajah, A. Guni, X. Liu, et al., "The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence," *Nature Medicine*, vol. 31, pp. 3283–3289, 2025, doi: 10.1038/s41591-025-03953-8