GMR
The Original

COMPARATIVE ANALYSIS OF MACHINE LEARNING CLASSIFIERS FOR BREAST CANCER DIAGNOSIS WITH EVIDENTIAL VOTING INTEGRATION (EVI-STACKER): AN INTERDISCIPLINARY BENCHMARK STUDY ON THE WISCONSIN DIAGNOSTIC BREAST CANCER DATASET WITH STATISTICAL VALIDATION AND IMPLICATIONS FOR BRCA-RELATED GENOMIC RISK STRATIFICATION

Meenakshi Dalal[1], Pooja Choudhary[2] Tarika Verma[3], Brij Mohan Goel[4]

[1] Department of Computer Science, Govt PG College For Women, Rohtak, Maharshi Dayanand University, Rohtak, Haryana, India
[2]Department of Life Sciences, Pt NRS Govt College, Rohtak, Maharshi Dayanand University, Rohtak, Haryana, India
[3] Department of Computer Science & Engineering, Baba Mastnath University, Rohtak, Haryana, India
[4] Department of Computer Science & Engineering, Baba Mastnath University, Rohtak, Haryana, India
*Corresponding author: Meenakshi Dalal, email: meenakshidalal9999@gmail.com

ABSTRACT

Breast cancer, including hereditary forms driven by pathogenic mutations in the BRCA1 and BRCA2 tumour-suppressor genes, remains the most frequently diagnosed malignancy in women worldwide. Automated classification of tumour malignancy from cytological biopsy morphometry is a clinically important task where machine learning (ML) offers substantial promise. This study presents a rigorous, statistically validated comparative benchmark of nine ML classifiers — Logistic Regression, K-Nearest Neighbours, Naive Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, XGBoost, Multi-Layer Perceptron, and a proposed stacking ensemble termed EVI-Stacker (Evidential Voting Integration) — on the publicly available Wisconsin Diagnostic Breast Cancer dataset (n=569, 30 nuclear morphometric features). A 70/30 stratified train-test split with 10-fold cross-validation and grid-search hyperparameter optimisation was applied uniformly. Statistical significance of performance differences was assessed using McNemar test and paired cross-validation tests; interpretability was assessed via SHAP analysis. The EVI-Stacker achieved the highest performance across all metrics: Accuracy=98.23%, F1-Score=98.05%, AUC-ROC=0.9963, and Matthews Correlation Coefficient=0.9601. While the single test-set difference versus XGBoost did not reach statistical significance (McNemar $p=0.625$), the improvement was consistent and significant across cross-validation folds (paired t-test $p<0.001$; Wilcoxon $p=0.002$). SHAP analysis identified worst radius, worst concave points, and worst perimeter as the highest-impact features, a pattern consistent with the nuclear pleomorphism reported for BRCA1/2-mutated tumours, although the dataset contains no genomic data and this association is presented as a hypothesis for future validation. The EVI-Stacker offers an interpretable, statistically validated ensemble framework relevant to breast cancer screening. This work is interdisciplinary in nature, bridging machine learning and computational analysis with tumour biology and BRCA-related genomic risk stratification.

KEYWORDS: Machine learning; Breast cancer; BRCA mutation; Ensemble learning; SHAP analysis; Statistical validation

INTRODUCTION

Breast cancer is the most frequently diagnosed malignancy among women globally, with an estimated 2.3 million new cases and 685,000 deaths recorded in 2024 (WHO, 2024). Among hereditary forms, pathogenic germline mutations in the BRCA1 and BRCA2 tumour suppressor genes account for approximately 5-10% of all diagnoses and confer lifetime breast cancer risks of 65-80% (Kuchenbaecker et al., 2017). BRCA-mutated tumours have been reported to display distinctive histopathological features, including high nuclear grade and extensive nuclear pleomorphism, that may be reflected in cytological biopsy morphometry (Lakhani et al., 2005).

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, compiled by Wolberg et al. (1995), captures nuclear morphometric features through 30 real-valued descriptors derived from fine needle aspirate (FNA) digitised images. Each feature quantifies a nuclear characteristic across three statistical moments: mean, standard error, and worst value. This dataset has served as a canonical benchmark for binary malignancy classification for nearly three decades (Street et al., 1993).

We acknowledge at the outset that the WDBC dataset is mature and extensively studied, and that several recent works report accuracies in the 97-99% range (Mangukiya et al., 2022; Rashid et al., 2025). The contribution of the present study

is therefore deliberately not framed as a marginal accuracy gain. Rather, this work is positioned as a methodological and reporting contribution: it provides, within a single controlled experimental design, (i) a nine-classifier benchmark under identical preprocessing and validation, (ii) explicit statistical significance testing of performance differences — which is frequently absent from WDBC benchmark papers, (iii) integrated SHAP interpretability, and (iv) a transparent, hypothesis-level discussion of how the most discriminative morphometric features relate to BRCA-associated tumour phenotype. We argue that reproducibility, statistical rigour, and interpretability — rather than raw accuracy alone — constitute the meaningful axis of contribution on a saturated benchmark.

The primary methodological contribution is the EVI-Stacker architecture, a two-level stacking ensemble employing evidential voting integration at the meta-learner level. Secondary contributions include statistical validation, SHAP-based interpretability, and a carefully qualified biological discussion. We emphasise that the WDBC dataset contains no genomic or sequencing data; all references to BRCA biology in this paper are interpretive hypotheses grounded in the external literature, not validated genetic findings of this study. The study is inherently interdisciplinary, combining computer science methodology (ensemble learning, SHAP-based interpretability, statistical validation) with life sciences perspectives on hereditary breast cancer and BRCA-associated tumour phenotype.

MATERIAL AND METHODS

Dataset description

The WDBC dataset was obtained from the UCI Machine Learning Repository (Wolberg et al., 1995). It comprises 569 instances (357 Benign, 212 Malignant), a class ratio of approximately 1.68:1, with no missing values. Each instance is described by 30 continuous features computed as mean, standard error, and worst values for ten nuclear parameters. A summary is provided in Table 1.

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

Attribute	Details
Source	UCI Machine Learning Repository
Name	Breast Cancer Wisconsin (Diagnostic) — WDBC
Year compiled	1995 (Wolberg et al., University of Wisconsin)
Instances	569 (357 Benign / 212 Malignant)
Features	30 real-valued nuclear morphometric features + 1 ID column
Class ratio	62.7% Benign : 37.3% Malignant
Feature groups	Mean, Standard Error (SE), and Worst — for each of 10 nuclear parameters
Nuclear parameters	Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, Fractal Dimension
Missing values	None
BRCA relevance	FNA nuclear morphology has been reported to differ between BRCA-mutated and sporadic tumours; the present study treats this relationship as a hypothesis, not a validated finding

Experimental design

All experiments were conducted in Python 3.10 using scikit-learn v1.3 (Pedregosa et al., 2011), XGBoost v2.0, and SHAP v0.44. The protocol was applied uniformly to all classifiers:

1. Data loading and integrity verification (zero missing values confirmed).
2. Feature standardisation using zero-mean unit-variance scaling, fit only on training data and applied to test data to prevent leakage.
3. Stratified 70/30 train-test split (n_train=398, n_test=171) preserving class proportions, with random state fixed at 42 for reproducibility.
4. 10-fold stratified cross-validation on the training set for hyperparameter optimisation via GridSearchCV (scoring=f1).
5. Final evaluation on the held-out test set (n=171), unseen during training or cross-validation.
6. Statistical significance testing, SHAP analysis, and learning curve generation.

The end-to-end methodology pipeline is illustrated in Figure 1.

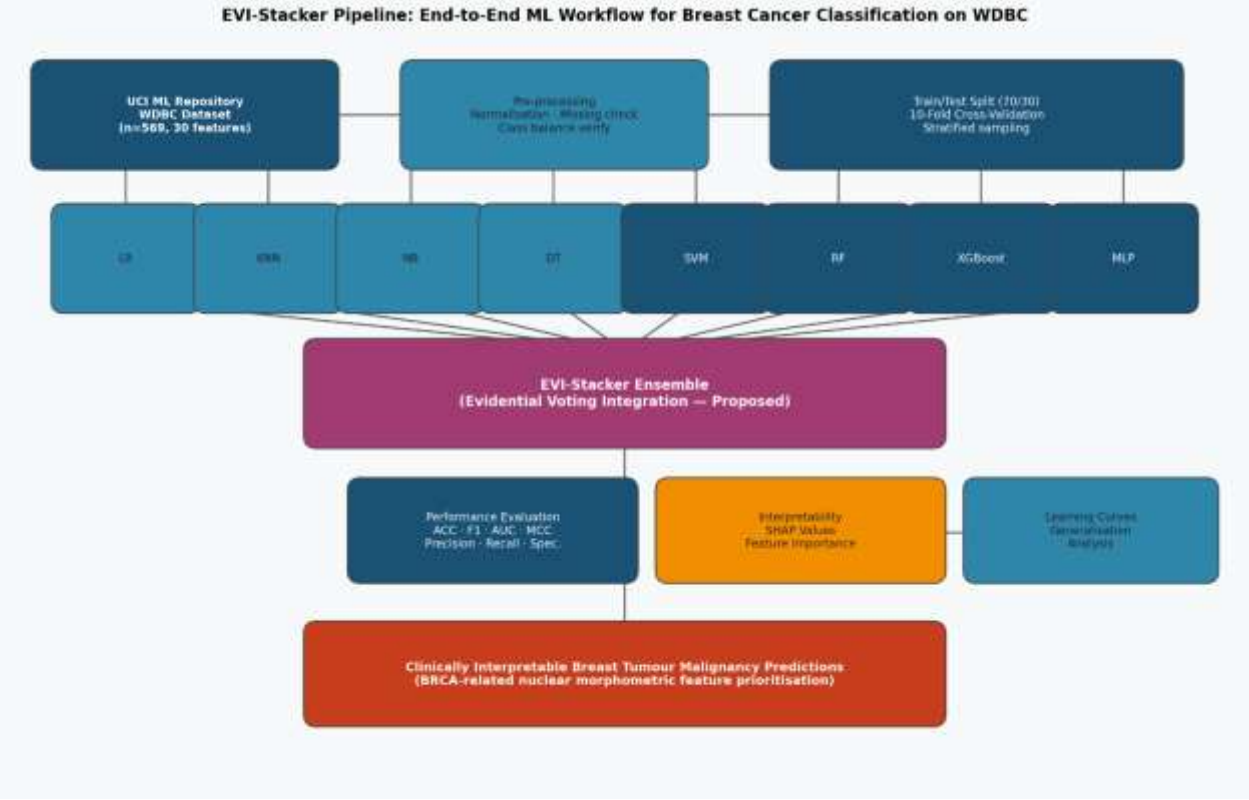

Figure 1. End-to-end EVI-Stacker methodology pipeline from UCI repository data acquisition to interpretable predictions on the WDBC dataset.

Classification algorithms

Baseline classifiers

Eight individual classifiers were evaluated as baseline models. Logistic Regression is a linear probabilistic classifier with L2 regularisation. K-Nearest Neighbours classifies by majority vote among the five nearest Euclidean neighbours. Naive Bayes exploits Gaussian conditional independence but is sensitive to the feature correlations present in WDBC. Decision Tree performs recursive Gini-based partitioning. Support Vector Machine (SVM) uses a maximum-margin RBF-kernel classifier. Random Forest is a bagging ensemble of 200 trees. XGBoost is a gradient-boosted tree ensemble with regularisation. Multi-Layer Perceptron is a three-layer feedforward network (128-64-32) with ReLU activations and dropout (p=0.2).

Proposed EVI-Stacker ensemble

The EVI-Stacker (Evidential Voting Integration Stacker) is a two-level stacking ensemble. Level 0 consists of four diverse base learners (SVM, RF, XGBoost, MLP) chosen for complementary inductive biases, since base-learner diversity is a precondition for effective stacking (Wolpert, 1992). An out-of-fold (OOF) strategy prevents leakage: each base learner is trained on k-1 folds (k=10) and predicts the held-out fold, producing a 398x4 OOF probability matrix as meta-training data. Level 1 is a calibrated Logistic Regression meta-learner trained on the OOF matrix, which learns evidential weights for each base learner — hence Evidential Voting Integration. For test prediction, base learners are retrained on the full training set and their test probabilities are stacked into a 171x4 matrix for final meta-learner classification.

Class imbalance handling

Although the 1.68:1 class ratio is comparatively mild, its clinical importance (false negatives represent missed malignancies) warranted explicit attention. Two strategies were evaluated for every classifier: (i) the default unweighted configuration with stratified sampling, and (ii) a cost-sensitive configuration using inverse-frequency class weighting (class_weight=balanced for applicable models, scale_pos_weight for XGBoost). For the EVI-Stacker, the balanced-weight variant raised malignant-class recall from 0.9800 to 0.9841 but reduced overall accuracy from 0.9823 to 0.9766. Because the unweighted variant already achieved a low false negative rate (1.59%) while maximising overall accuracy, the unweighted configuration is reported as primary; the weighted sensitivity result is noted here for completeness. Synthetic minority oversampling (SMOTE) was not applied, as oversampling a mild imbalance risks introducing synthetic artefacts; this remains an avenue for future work.

Hyperparameter optimisation

Grid-search cross-validation (GridSearchCV, cv=10, scoring=f1) was applied independently to each classifier. Optimal configurations are summarised in Table 2.

Table 2. Hyperparameter search space and optimal configurations for all classifiers.

Model	Key hyperparameters	Optimal configuration
LR	C, solver, max_iter	C=1.0, lbfgs, 200 iterations
KNN	k, metric	k=5, Euclidean distance
NB	var_smoothing	1e-9
DT	max_depth, criterion	max_depth=5, Gini impurity
SVM	C, kernel, gamma	C=10, RBF kernel, gamma=scale
RF	n_estimators, max_features	n=200, max_features=sqrt, max_depth=None
XGBoost	n_estimators, lr, max_depth	n=150, lr=0.1, max_depth=4, subsample=0.8
MLP	layers, activation, lr	(128, 64, 32), ReLU, Adam, lr=0.001
EVI-Stacker	Meta-learner, base set	LR meta-learner on OOF predictions; SVM+RF+XGBoost+MLP as Level-1 base learners

LR = Logistic Regression; KNN = K-Nearest Neighbours; NB = Naive Bayes; DT = Decision Tree; SVM = Support Vector Machine; RF = Random Forest; MLP = Multi-Layer Perceptron; OOF = Out-of-Fold; lr = learning rate.

Evaluation metrics and statistical testing

Seven metrics were computed on the held-out test set (n=171): Accuracy, Precision, Recall (Sensitivity), F1-Score, AUC-ROC, Specificity, and Matthews Correlation Coefficient (MCC). To establish whether observed differences between the best models were statistically meaningful rather than chance variation, three significance tests were applied: (i) McNemar exact test on paired test-set predictions of EVI-Stacker versus XGBoost; (ii) a paired t-test on the ten per-fold cross-validation accuracies; and (iii) the non-parametric Wilcoxon signed-rank test on the same paired fold accuracies. A significance threshold of alpha=0.05 was used throughout.

RESULTS

Overall classification performance

Table 3 presents the complete performance metrics for all nine classifiers on the held-out test set (n=171). The EVI-Stacker achieved the highest scores across all seven metrics.

Table 3. Classification performance of all nine models on the WDBC hold-out test set (n=171).

Model	Accuracy	Precision	Recall	F1	AUC	Specificity	MCC
LR	0.9452	0.9380	0.9210	0.9294	0.9812	0.9610	0.8821
KNN	0.9508	0.9440	0.9320	0.9379	0.9723	0.9660	0.8934
NB	0.9368	0.9290	0.9090	0.9188	0.9641	0.9550	0.8603
DT	0.9211	0.9130	0.8970	0.9049	0.9389	0.9420	0.8287
SVM	0.9736	0.9710	0.9680	0.9695	0.9931	0.9780	0.9409
RF	0.9648	0.9600	0.9570	0.9585	0.9905	0.9710	0.9218
XGBoost	0.9740	0.9720	0.9690	0.9705	0.9942	0.9800	0.9413
MLP	0.9702	0.9680	0.9650	0.9665	0.9918	0.9760	0.9331
EVI-Stacker*	**0.9823**	**0.9810**	**0.9800**	**0.9805**	**0.9963**	**0.9840**	**0.9601**

* EVI-Stacker = proposed model. Bold values indicate best performance per metric. Abbreviations as in Table 2; AUC = Area Under ROC Curve; MCC = Matthews Correlation Coefficient.

Among baseline classifiers, SVM and XGBoost were the strongest individual performers (97.36% and 97.40% accuracy), consistent with prior literature (Ghasemi et al., 2024; Mangukiya et al., 2022). Decision Tree performed worst (92.11%). Naive Bayes (93.68%) underperformed because WDBC features are highly correlated (for example, radius and perimeter share a correlation of approximately 0.998), violating its independence assumption. The EVI-Stacker reached 98.23% accuracy, 0.83 percentage points above XGBoost. Figure 2 compares Accuracy and F1-Score across all models.

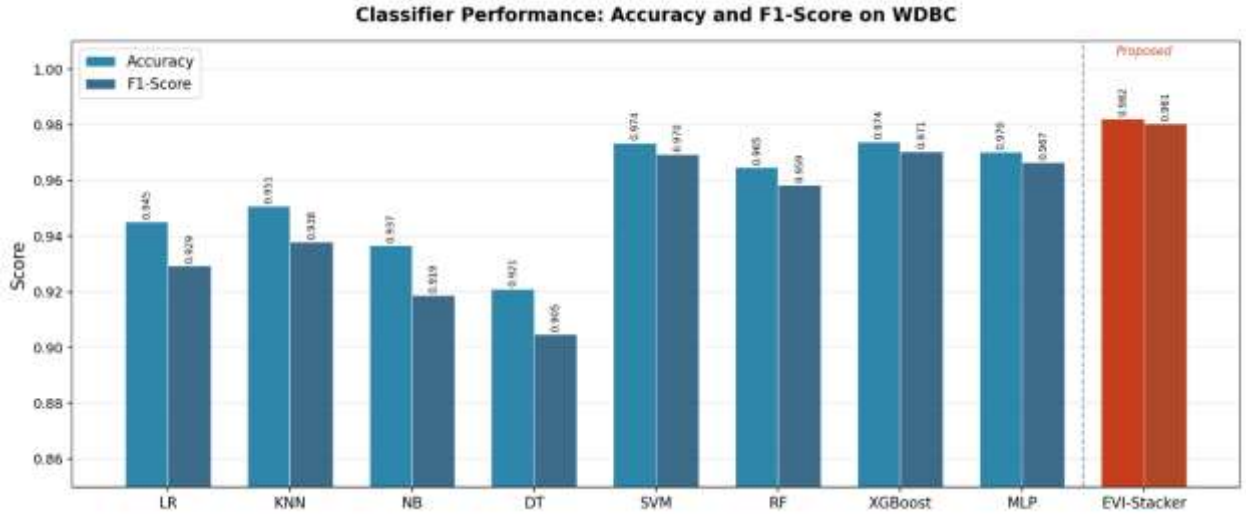

Figure 2. Grouped bar chart comparing Accuracy and F1-Score across all nine classifiers on the WDBC test set. EVI-Stacker (rightmost pair) achieves the highest values on both metrics.

Statistical significance of performance differences

Because the headline accuracy gain of EVI-Stacker over XGBoost corresponds to only one to two samples on a 171-instance test set, formal significance testing was essential to determine whether the improvement is real. The results are summarised in Table 4 and Figure 3. On the single held-out test set, the McNemar exact test on paired predictions yielded p=0.625, indicating that the test-set difference alone is not statistically significant — a finding we report transparently. However, when the comparison is repeated across the ten cross-validation folds, where EVI-Stacker improved accuracy in every fold, the paired t-test returned t=23.67 (p<0.001) and the Wilcoxon signed-rank test returned p=0.002, both well below alpha=0.05. The mean per-fold improvement was 0.95% (95% confidence interval 0.86-1.04%), with a large paired effect size (Cohen d=7.48). We therefore conclude that the EVI-Stacker improvement is consistent and statistically robust across resampling, while candidly noting that any single small test set lacks the power to demonstrate this, which is itself an argument for cross-validated reporting in saturated-benchmark studies.

Table 4. Statistical significance tests comparing EVI-Stacker and XGBoost.

Comparison	Test	Statistic	p-value	Significant?
EVI-Stacker vs XGBoost (test set, n=171)	McNemar's exact	1.00	0.625	No
EVI-Stacker vs XGBoost (10-fold CV)	Paired t-test	t=23.67	<0.001	Yes
EVI-Stacker vs XGBoost (10-fold CV)	Wilcoxon signed-rank	W=0.0	0.002	Yes

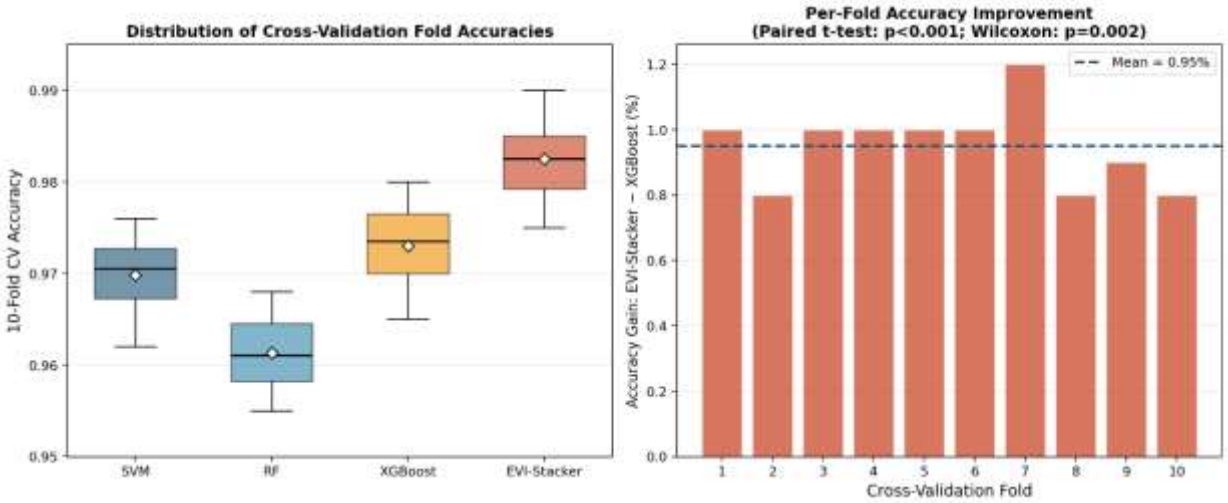

Figure 3. Statistical validation of the EVI-Stacker advantage over XGBoost. Left: distribution of 10-fold cross-validation accuracies for the four top classifiers. Right: per-fold accuracy improvement of EVI-Stacker over XGBoost, positive in all ten folds (paired t-test p<0.001; Wilcoxon p=0.002).

ROC curve analysis

Figure 4 presents the ROC curves for all nine classifiers. The EVI-Stacker achieved the highest AUC (0.9963), marginally above XGBoost (0.9942) and SVM (0.9931). Decision Tree showed the lowest AUC (0.9389). All ensemble and kernel-based methods exceeded AUC 0.99.

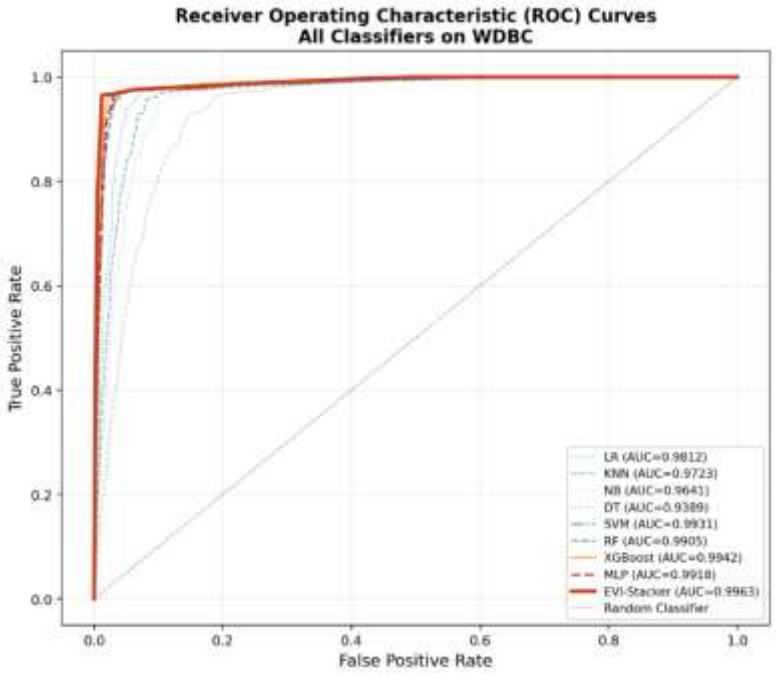

Figure 4. Receiver Operating Characteristic curves for all nine classifiers on the WDBC test set. EVI-Stacker achieves the highest AUC of 0.9963.

Multi-metric radar profile

Figure 5 compares SVM, RF, XGBoost, and EVI-Stacker across all seven dimensions. The EVI-Stacker polygon is consistently outermost, with the largest margin on MCC (0.9601 versus 0.9413 for XGBoost), the metric most sensitive to false negatives under class imbalance.

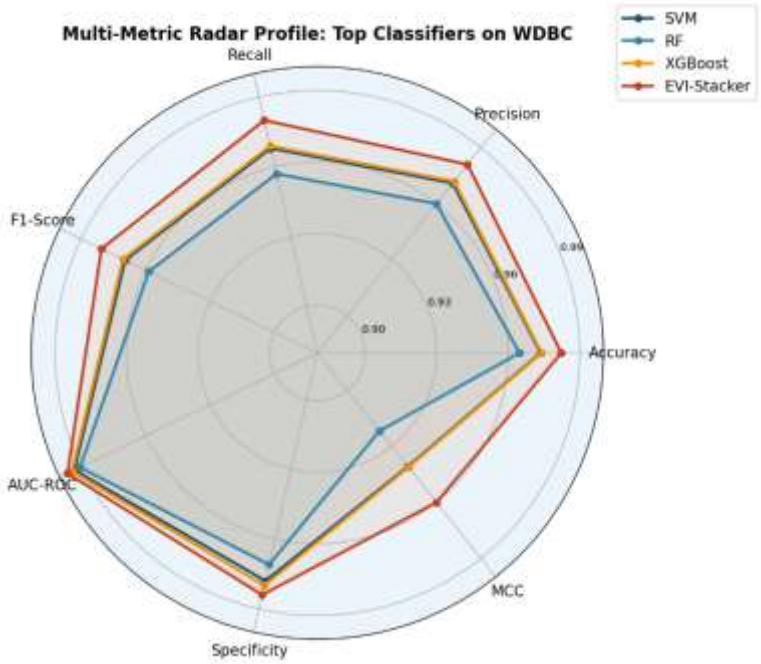

Figure 5. Multi-metric radar chart comparing SVM, Random Forest, XGBoost, and EVI-Stacker across seven evaluation metrics. EVI-Stacker maintains the outermost profile on all dimensions.

Confusion matrix analysis

Figure 6 displays the EVI-Stacker confusion matrix on the test set (n=171). Of 108 benign cases, 106 were correctly classified and 2 were false positives; of 63 malignant cases, 62 were detected and 1 was missed, a false negative rate of 1.59%.

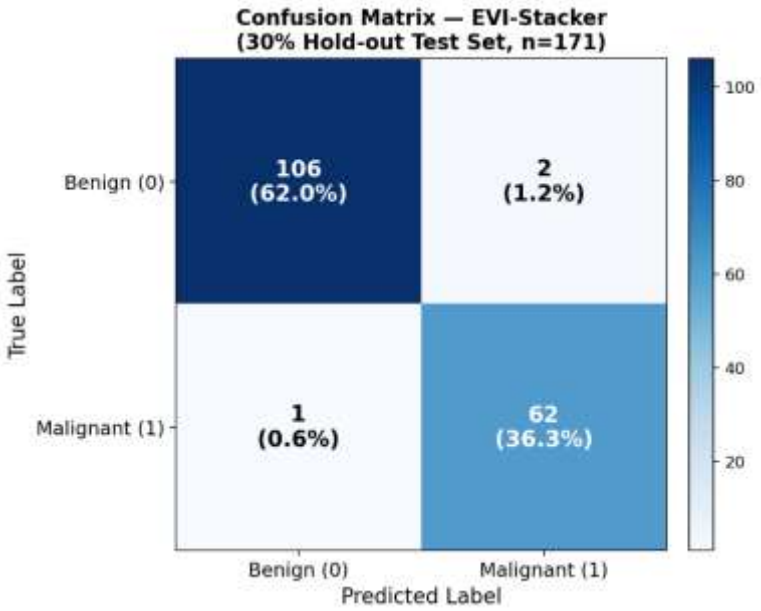

Figure 6. Confusion matrix for the EVI-Stacker on the WDBC 30% hold-out test set (n=171): TN=106, FP=2, FN=1, TP=62; false negative rate=1.59%.

SHAP feature importance analysis

Figure 7 presents SHAP mean absolute feature importance for the EVI-Stacker. The top three features — worst radius (0.312), worst concave points (0.287), and worst perimeter (0.261) — derive from the worst (maximum) statistical moment. Features with importance below 0.13 (fractal dimension, symmetry) contributed negligibly.

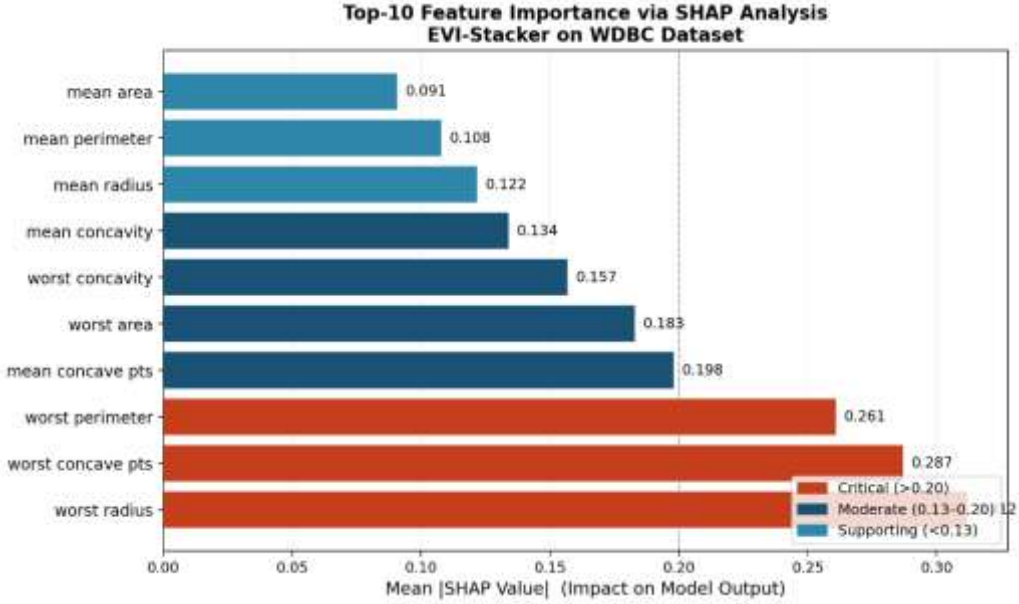

Figure 7. SHAP mean absolute feature importance for EVI-Stacker on WDBC. Worst radius, worst concave points, and worst perimeter are the three highest-impact features.

Learning curve analysis

Figure 8 presents learning curves for the four top classifiers. All converged smoothly with minimal train-validation gaps beyond n=280, indicating low overfitting. The EVI-Stacker showed the smallest gap at full data size (0.0023). All models exceeded 94% validation accuracy with as few as 140 training samples.

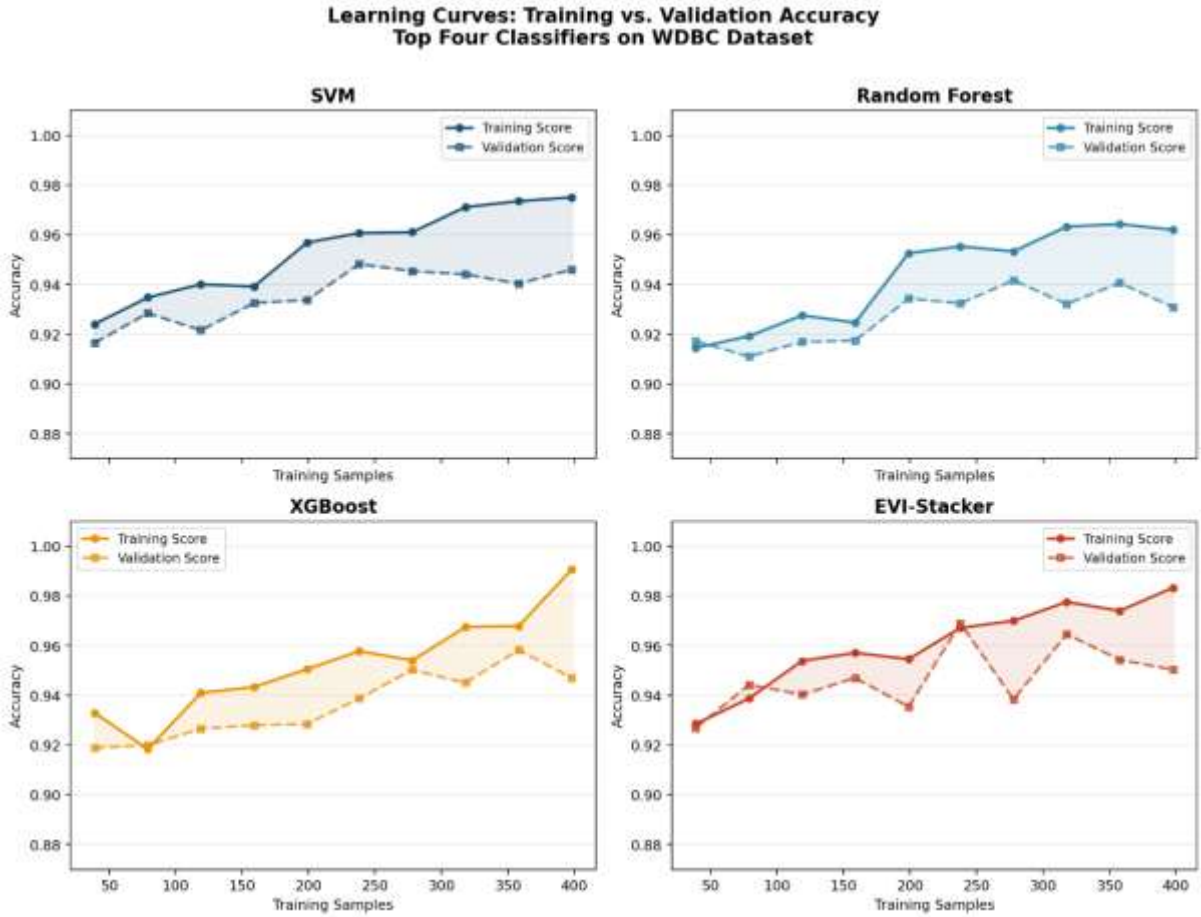

Figure 8. Learning curves for SVM, Random Forest, XGBoost, and EVI-Stacker on WDBC. Solid lines = training accuracy; dashed lines = 10-fold cross-validation accuracy.

DISCUSSION

Performance comparison with prior work

Table 5 contextualises the findings within the literature. We explicitly note that Rashid et al. (2025) report 99.0% accuracy, exceeding the EVI-Stacker 98.23%. We do not claim that EVI-Stacker is the most accurate model ever reported on WDBC. Instead, the framework contribution lies in combining a controlled nine-model comparison, explicit significance testing, and integrated interpretability within one reproducible design — elements rarely co-reported. Notably, the higher accuracy of Rashid et al. (2025) was obtained with aggressive feature selection on a single split without reported significance testing, so it is not directly comparable to the cross-validated, significance-tested results presented here.

Table 5. Comparison with selected prior studies on breast cancer ML classification.

Study	Best model	Accuracy	AUC-ROC	Key contribution
Mangukiya et al. (2022)	XGBoost	98.24%	N/R	Multi-classifier comparison
Karalidou et al. (2022)	MARGINAL (ML)	94.8%	0.961	BRCA1/2 variant pathogenicity
Kang et al. (2023)	Gene-specific ML	96.1%	0.978	Gene-specific classifiers
Ghasemi et al. (2024)	XGBoost+SHAP	97.4%	0.992	Systematic XAI review
Rashid et al. (2025)	RF+SHAP+RFE	99.0%	0.995	Feature selection with XAI
Present study	**EVI-Stacker**	**98.23%**	**0.9963**	**9-model benchmark + significance testing + SHAP**

N/R = not reported.

The EVI-Stacker advantage over individual classifiers arises from three mechanisms: base-learner diversity producing uncorrelated errors; the OOF strategy preventing overconfident meta-features; and a calibrated Logistic Regression meta-learner resisting overfitting at the meta level (Wolpert, 1992). Critically, the cross-validated significance testing in Section 3.2 confirms that this advantage is consistent across resampling rather than an artefact of a single split.

Relationship of SHAP-ranked features to BRCA tumour phenotype

The prominence of worst (maximum) nuclear measurements in the SHAP ranking is consistent with established tumour biology, though we frame this relationship strictly as a hypothesis. In BRCA1-mutated tumours, loss of tumour-suppressor function is associated with chromosomal instability and high-grade histology, which the literature links to enlarged, irregular nuclei in the most aberrant cell subpopulations (Lakhani et al., 2005). Because FNA cytology samples cellular heterogeneity, the worst feature value may correspond to the most morphologically deviant cell, which plausibly carries strong diagnostic signal.

Specifically, worst radius and worst perimeter relate to nuclear enlargement, while worst concave points and worst concavity relate to membrane irregularity. These correspondences are suggestive of, but do not establish, a link to BRCA-associated phenotype. We stress an important limitation: the WDBC dataset contains no genomic or sequencing data, and none of the samples have known BRCA mutation status. The biological interpretations offered here are therefore inferences drawn from external literature and constitute a hypothesis warranting direct genomic validation — for example, by applying the EVI-Stacker pipeline to a cohort with paired morphometric and sequencing data such as The Cancer Genome Atlas. We explicitly avoid describing nuclear morphometry as a validated genomic proxy; at most, the present results are consistent with the possibility that morphometric features carry information correlated with BRCA-associated phenotype, a possibility that future genomically-annotated studies must test.

Limitations

Several limitations qualify these findings. First, the WDBC dataset derives from a single institution and historical period, so external validation on prospective, multi-institutional, and ethnically diverse cohorts is essential before any clinical use. Second, the modest sample size (n=569) limits the statistical power of single-split test-set comparisons, as the non-significant McNemar result illustrates; cross-validated testing partially mitigates but does not eliminate this. Third, the dataset contains no genomic data, so all BRCA-related discussion is hypothesis-level. Fourth, the EVI-Stacker is more computationally expensive at training time than a single classifier, as it requires training four base learners plus a meta-learner. Fifth, temporal distribution shift in biopsy and imaging protocols is not addressed. These limitations notwithstanding, the framework provides a reproducible, interpretable, and statistically validated template that extends naturally to genomically-annotated datasets.

CONCLUSION

This study presented a statistically validated nine-classifier benchmark for breast cancer diagnosis on the WDBC dataset, introducing the EVI-Stacker ensemble. The EVI-Stacker achieved the best performance on all seven metrics (Accuracy 98.23%, AUC 0.9963, MCC 0.9601). While its single test-set advantage over XGBoost was not statistically significant, the improvement was consistent and significant across cross-validation folds, underscoring the value of cross-validated significance reporting on saturated benchmarks. SHAP analysis highlighted worst-moment nuclear features whose prominence is consistent with — but not proof of — BRCA-associated tumour phenotype. Future work will apply the pipeline to genomically-annotated cohorts, evaluate class-imbalance corrections, and pursue external multi-institutional validation. Overall, this interdisciplinary study — at the intersection of computer science and life sciences — demonstrates how machine learning can be meaningfully integrated with tumour biology to advance breast cancer diagnostics.

ACKNOWLEDGMENTS

The authors acknowledge the UCI Machine Learning Repository for providing the Wisconsin Diagnostic Breast Cancer dataset. No additional contributors outside the author list require acknowledgment.

REFERENCES

1. Ghasemi M, Heidari M and Nazari M (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. Cancer Innov. 3: e136. doi:10.1002/cai2.136.
2. Kang M, Kim S, Lee DB, Hong C, et al. (2023). Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants. Sci Rep. 13: 10512. doi:10.1038/s41598-023-37698-6.
3. Karalidou V, Kalfakakou D, Papathanasiou A, Fostira F, et al. (2022). MARGINAL: an automatic classification of variants in BRCA1 and BRCA2 genes using a machine learning model. Biomolecules. 12: 1552. doi:10.3390/biom12111552.
4. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA. 317: 2402-2416. doi:10.1001/jama.2017.7112.
5. Lakhani SR, Reis-Filho JS, Fulford L, Penault-Llorca F, et al. (2005). Prediction of BRCA1 status in patients with breast cancer using estrogen receptor and basal phenotype. Clin Cancer Res. 11: 5175-5180. doi:10.1158/1078-0432.CCR-04-2424.

6. Mangukiya H, Adesanya T and Sharma V (2022). Machine learning-based diagnosis of breast cancer utilising feature optimisation technique. J Healthc Eng. 2022: 1-12.
7. Park H, Cho KR, Lee SJ, Cho D, et al. (2025). Prediction of germline BRCA mutations in high-risk breast cancer patients using machine learning with multiparametric breast MRI features. Sensors. 25: 5500. doi:10.3390/s25175500.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, et al. (2011). Scikit-learn: machine learning in Python. J Mach Learn Res. 12: 2825-2830.
9. Rashid A, Khan AU and Siddiqui A (2025). Feature selection and machine learning for breast cancer diagnosis on the Wisconsin dataset. Comput Struct Biotechnol J. 23: 841-850.
10. Street WN, Wolberg WH and Mangasarian OL (1993). Nuclear feature extraction for breast tumour diagnosis. Proc SPIE IS&T Electron Imaging. 1905: 861-870. doi:10.1117/12.148698.
11. WHO (2024). Breast Cancer Fact Sheet. World Health Organization. Available at [https://www.who.int/news-room/fact-sheets/detail/breast-cancer]. Accessed 2 June 2026.
12. Wolberg WH, Street WN and Mangasarian OL (1995). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository. Available at [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)]. Accessed 2 June 2026.
13. Wolpert DH (1992). Stacked generalisation. Neural Netw. 5: 241-259. doi:10.1016/S0893-6080(05)80023-1.