Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Machine Learning for Breast Cancer Risk Prediction

Patience Atukunda

patukunda2@gmail.com

Makerere University

Joyce Nakatumba-Nabende

Makerere University

Rhoda K Wanyenze

Makerere University

Ida-Maria Sintorn

Uppsala University

Research Article

Keywords: Breast cancer, Machine learning, XAI, Logistic Regression, XGBoost, Random Forest, ANN, Naive Bayes, SVM

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11291724/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable Machine Learning for Breast Cancer Risk Prediction

Patience Atukunda[1*], Joyce Nakatumba-Nabende[2], Rhoda K Wanyenze[3] and Ida-Maria Sintorn[4]
[1]School of Medicine, Makerere University, Kampala, Uganda. Email: patukunda2@gmail.com
[2]Department of Computer Science, Makerere University, Kampala, Uganda. Email: jnakatumba@gmail.com
[3]School of Public Health, Makerere University, Kampala, Uganda. Email: rwanyenze@musph.ac.ug
[4]Department of Information Technology, Uppsala University, Uppsala, Sweden. Email: ida.sintorn@it.uu.se
*Corresponding Author

Abstract

Breast cancer remains one of the leading causes of cancer-related morbidity and mortality among women worldwide. Early identification of women at increased risk is essential for timely intervention and improved clinical outcomes. Although machine learning has shown considerable promise in breast cancer prediction, comparative evaluations of diverse machine learning paradigms populations remain limited. This study compared probabilistic, linear, ensemble, and deep learning models for breast cancer risk prediction while incorporating explainable artificial intelligence to improve model interpretability. A clinical and epidemiological dataset of Cuban women was used to build six breast cancer prediction models; Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network. These were trained using inverse class weighting to address class imbalance. Model performance was evaluated using stratified 5-fold and 10-fold cross-validation and an independent hold-out test set. Performance was assessed using ROC-AUC, accuracy, sensitivity, specificity, precision, F1-score, and Matthews Correlation Coefficient (MCC). Statistical differences between models were evaluated using the Wilcoxon signed-rank test, while SHAP (SHapley Additive exPlanations) was employed to interpret feature importance. All six models demonstrated excellent predictive performance, achieving ROC-AUC values exceeding 0.97 across both cross-validation strategies. XGBoost consistently achieved the highest overall performance (), followed closely by Random Forest, with both models demonstrating superior stability and generalization across validation folds. These findings demonstrate that ensemble learning methods provide accurate, robust, and interpretable prediction of breast cancer risk and show promise for supporting personalized risk assessment and early detection of breast cancer.

Keywords: Breast cancer; Machine learning; XAI; Logistic Regression, XGBoost; Random Forest, ANN, Naive Bayes, SVM.

Introduction and Background

Breast cancer is one of the most prevalent malignancies and leading cause of cancer related mortality affecting women worldwide[1]. Early identification of individuals at high risk is critical for improving prevention, screening, and treatment outcomes. Numerous studies have identified important breast cancer risk factors, including age, family history, personal history, reproductive history, alcohol consumption, tobacco use, hormonal exposure, obesity, and genetic predisposition[2], [3], [4]. These factors have formed the foundation for many statistical and computational breast cancer prediction models. Advances in machine learning have significantly improved predictive modelling in healthcare. Unlike traditional statistical approaches, machine learning algorithms can automatically identify nonlinear relationships and hidden patterns from large nonlinear relationships and hidden patterns from large and complex datasets [13].

In breast cancer research, machine learning methods have increasingly been used for diagnosis, risk prediction, prognosis, and treatment outcome analysis. Comparative studies have shown that machine learning techniques frequently outperform conventional statistical methods in predictive accuracy [14]. Recent advances in artificial intelligence and machine learning have demonstrated substantial potential in improving breast cancer prediction, diagnosis, prognosis, and survival analysis [5]. Machine Learning is a field of study that gives computers the ability to learn without explicitly being programmed

XGBoost has emerged as one of the most effective machine learning algorithms for structured clinical data because of its high predictive accuracy and ability to model nonlinear relationships. Studies have shown that XGBoost frequently outperformed traditional classifiers such as logistic regression and support vector machines for breast cancer prediction [8]. Other studies have also shown that integrating techniques such as SHAP improves transparency and interpretability of the results of complex machine learning models in breast cancer research [9].

Several studies have applied machine learning techniques to breast cancer datasets. For example, in 2024, Islam et al. used XGBoost model to explain breast cancer classification outcomes and identify influential predictors[6], SHAP values were applied to RF, XGBoost and lasso regression to analyse the risk factors between two age groups using mammography [7]. Similarly, a study by Dianati-Nasab et al in 2024 used Random Forest, Neural Networks, Bagged CART, and XGBoost models to identify modifiable and non-modifiable breast cancer risk factors in an Iranian population[6]. Despite the growing body of literature on breast cancer prediction, several limitations remain. Many existing studies focus predominantly on imaging datasets such as mammograms and histopathological images, while relatively fewer studies investigate epidemiological and lifestyle-based breast cancer risk-factor datasets. Furthermore, limited research has specifically examined breast cancer prediction among despite possible demographic, genetic, and environmental differences that may influence disease patterns. [8].

The contribution of the paper is a comparison of Probabilistic, Linear, Ensemble and Deep Learning models. We compared the performance of Support Vector Machines, Random Forest, Naïve Bayes, Logistic Regression, Artificial Neural Networks and XGBoost models in predicting breast cancer among Cuban women. This study emphasizes rigorous preprocessing and

methodological validation, including careful handling of categorical variables, prevention of data leakage, balancing of classes using class weights and comparison of performance of seven machine learning models on the same dataset. Although the dataset used in this study originates from Cuban women, the findings contribute valuable insights that may inform future breast cancer prediction research in Sub-Saharan Africa and other resource-constrained settings. The methodology presented in this study can be adapted for African epidemiological datasets and may support the development of affordable and data-driven breast cancer risk assessment tools suitable for low-resource healthcare environments.

Methodology

Data preprocessing

The dataset consists of 1697 breast cancer patients and 18 variables comprising both numerical and categorical variables, including several columns with mixed textual and numeric information. These variables include; id, age, menarche, menopause, agefirst, children, breastfeedingmonths, nrelbc, biopsies, hyperplasia, race, year, imc, weight, exercise, alcohol, tobacco, allergies, emotional, depressive, histologicalclass, birads and the outcome variable cancer. Out of 1697 patients, 1160 have cancer and 537 have no cancer. Several preprocessing steps were applied to ensure data quality. Missing values were handled by replacing NaN with 0 where appropriate. All categorical variables such as hyperplasia, alcohol, tobacco, and cancer were converted into binary numerical representations. Additionally, in the nrelbc column, family history categories were converted into a count of affected relatives. Variables not intended for predictive modelling were excluded before model development. These include, Emotional, Depressive, Histological Class, BIRADS, and Biopsies. Histological class, BIRADS assessment, and biopsy findings represent diagnostic information that is typically unavailable during early risk prediction and could introduce data leakage into the learning process. Excluding these variables ensured that the developed models relied only on clinically available predictor variables. This aligns with concerns raised by[9], who emphasized that data leakage remains a major source of misleadingly high performance in machine learning research.[9]. Continuous variables requiring normalization were standardized where necessary, particularly for algorithms sensitive to feature scaling, such as Support Vector Machines and Artificial Neural Networks.

Machine Learning workflow

Six supervised algorithms representing different machine learning paradigms were used as indicated in Figure 1 below. Logistic Regression served as the conventional statistical baseline because of its widespread application in clinical prediction studies. Naive Bayes represented probabilistic classification, Support Vector Machine represented kernel-based learning, Random Forest represented ensemble bagging methods, XGBoost represented gradient boosting techniques, and the Artificial Neural Network represented deep learning approaches. To address class imbalance, class weights were calculated from the training split and incorporated into the training process.

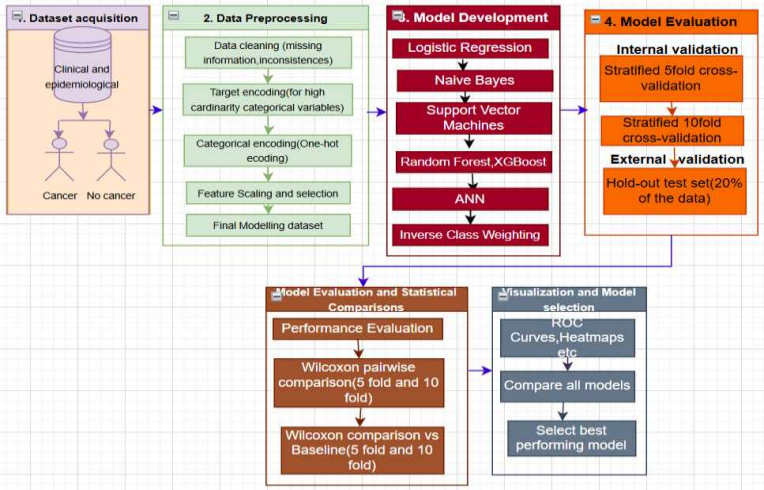

Figure 1:Machine Learning workflow for this study

Class imbalance handling

Since the dataset exhibited an unequal distribution of breast cancer and non-cancer cases, inverse-frequency class weighting was employed to reduce potential bias toward the majority class. Class weights were calculated from the training data as

$$W_k = \frac{N}{KN_k}$$

Equation 1: Class Imbalance Handling

where N denotes the total number of training observations, K the number of classes, and N_k the number of observations in class k.

The resulting weights were incorporated into the model-specific classification loss, such that observations from the minority class contributed proportionally more to the optimization objective. Importantly, class weights were calculated using the training data within each cross-validation fold to prevent information leakage.

Model Explainability

SHapley Additive exPlanations (SHAP) [10]was employed to quantify the contribution of individual predictors to model predictions. For each observation, the SHAP framework decomposes the model output into a baseline prediction and feature-specific contributions, such that

$$f(x) = \emptyset_o + \sum_{j=1}^{p} \emptyset_j$$

Equation 2: Model Explainability

Global feature importance was estimated using the mean absolute SHAP value,

$$I_J = N^{-1} \sum_{i=1}^{n} \emptyset_{ij}$$

Equation 3: Global Feature Importance

across all observations.

For visualization and comparison across models, feature importance values were normalized by dividing each feature's mean absolute SHAP value by the maximum importance within the corresponding model. Consequently, a normalized value of 1 represented the most influential predictor for a given model, whereas lower values represented relatively smaller contributions.

Model Evaluation Using Stratified Cross-Validation

To obtain robust estimates of predictive performance and minimize bias resulting from a single train-test split, model evaluation was performed using both stratified five-fold and stratified ten-fold cross-validation. Using both five-fold and ten-fold cross-validation enabled assessment of the consistency and stability of model performance under different validation schemes.

Performance Evaluation Metrics

Model performance was evaluated using ROC-AUC, Accuracy, Precision, Sensitivity (Recall), Specificity and F1-score. For each classifier, the mean and standard deviation of every evaluation metric were calculated across the cross-validation folds. Following cross-validation, model performance was further evaluated using an independent hold-out validation procedure. The dataset was randomly divided into 80% training data and 20% testing data. Statistical comparisons between machine learning algorithms were performed using the Wilcoxon signed-rank test. This was done using pairwise Wilcoxon signed-rank test to compare every machine learning algorithm against each of the remaining algorithms. This analysis determined whether observed differences in predictive performance were statistically significant. Furthermore, baseline Wilcoxon signed-rank tests were conducted using Logistic Regression as the reference model. Logistic Regression was selected because it represents the conventional statistical approach commonly employed in clinical prediction modelling. Each advanced machine learning algorithm was individually compared with Logistic Regression to determine whether statistically significant improvements in predictive performance were achieved. Statistical significance was assessed using a significance level of $\alpha = 0.05$.

Results

Model performance over 10-fold CV

In this section, we describe the results of the six Machine Learning models. Figure 2 below shows that XGBoost consistently achieved the highest scores across nearly all evaluation metrics, including ROC-AUC (0.994) accuracy (0.961), F1-score (0.971), precision (0.972), and sensitivity (0.971), specificity (0.938) and Matthews Correlation Coefficient (0.999), confirming its superiority as the best-performing model in this study. Random Forest exhibited performance comparable to XGBoost, with only marginally lower values, indicating that both ensemble

learning algorithms provided highly accurate and robust predictions. Artificial Neural Network and Naïve Bayes also showed strong predictive performance, although they were slightly less consistent across all metrics. In contrast, Logistic Regression and Support Vector Machine demonstrated comparatively lower scores, particularly for MCC and F1-score as is seen in figure 2, suggesting a reduced ability to correctly classify both breast cancer and non-cancer cases relative to the ensemble models. The uniformly high ROC-AUC values observed across all models indicate excellent discriminatory ability; however, the differences in F1-score and MCC reveal that ensemble methods achieved a better balance between sensitivity and specificity.

Nevertheless, the ten-fold ROC-AUC distribution in Figure 2 and 3 shows that XGBoost achieved the highest median ROC-AUC and the smallest variability, making it the most accurate and stable model followed by Random Forest although it exhibited slightly greater variability than XGBoost. ANN produced high ROC-AUC values, but its wider interquartile range and longer whiskers indicate greater variability across the validation folds. This suggests that the ANN produced excellent predictions overall but was more sensitive to changes in the training data than the ensemble methods. Additionally, Naïve Bayes, Logistic Regression, and SVM all achieved excellent discrimination but were less stable than the ensemble methods, particularly Logistic Regression and SVM, indicating that their predictive performance varied more across different data partitions. Such variability may reflect the inability of these conventional models to fully capture the complex nonlinear relationships among the demographic, reproductive, hereditary, lifestyle, and clinical risk factors included in the dataset. The generally narrow distributions across all models indicate that the 10-fold stratified cross-validation provided robust and reliable estimates of model performance.

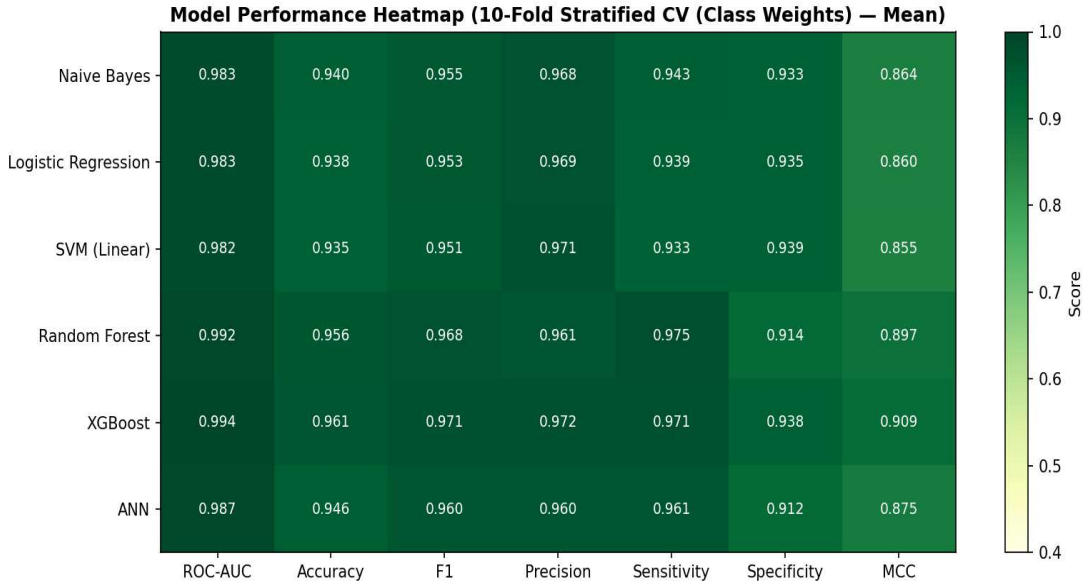

Figure 2: Mean Predictive Performance of Machine Learning Models Using 10-Fold Stratified Cross-Validation

Feature Importance Comparison Across all Models

The SHAP feature importance analysis revealed that the contribution of predictor variables varied across all the models. hyperplasia, alcohol consumption, tobacco use, body mass index (BMI), body weight, number of children, and breastfeeding duration consistently emerged as the most influential predictors of breast cancer. In figure 3 and 4 below, Hyperplasia is seen as the most important feature across Naïve Bayes (1.0), Logistic Regression (0.87), Random Forest (0.9), XGBoost (0.93), and the Artificial Neural Network (0.95).

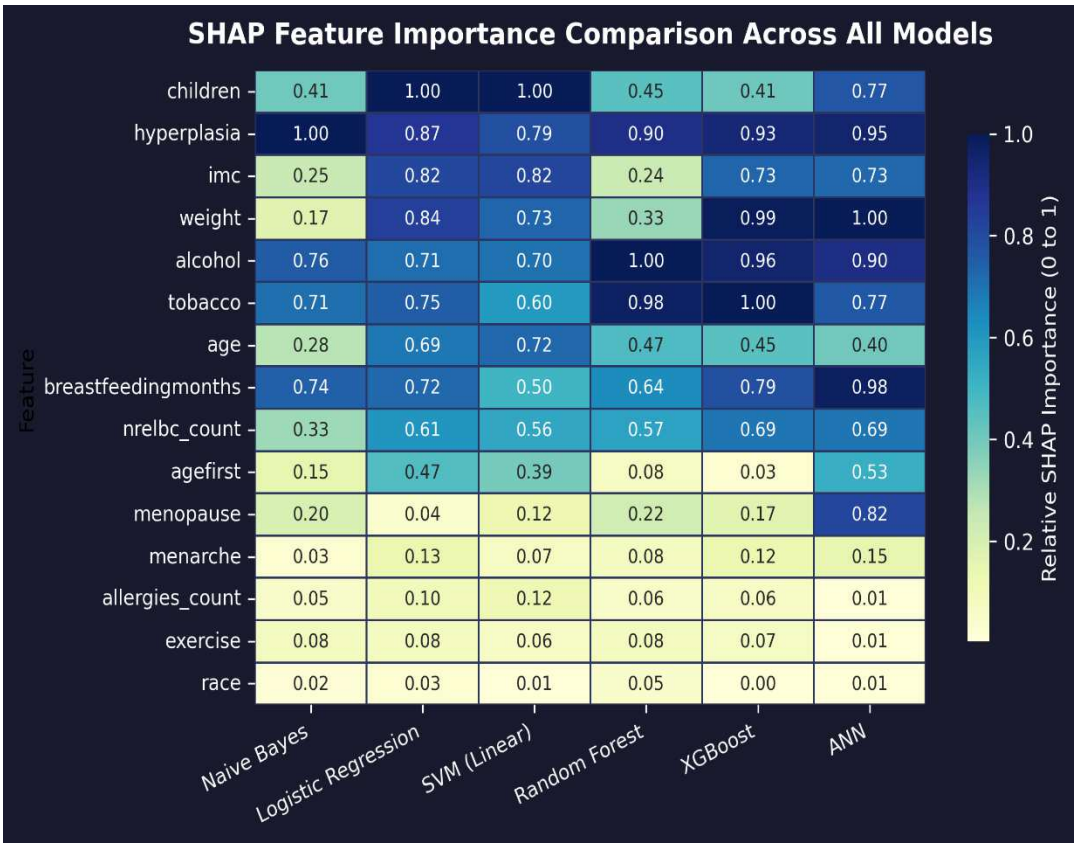

Figure 3: Relative SHAP Importance of Breast Cancer Predictors Across Machine Learning

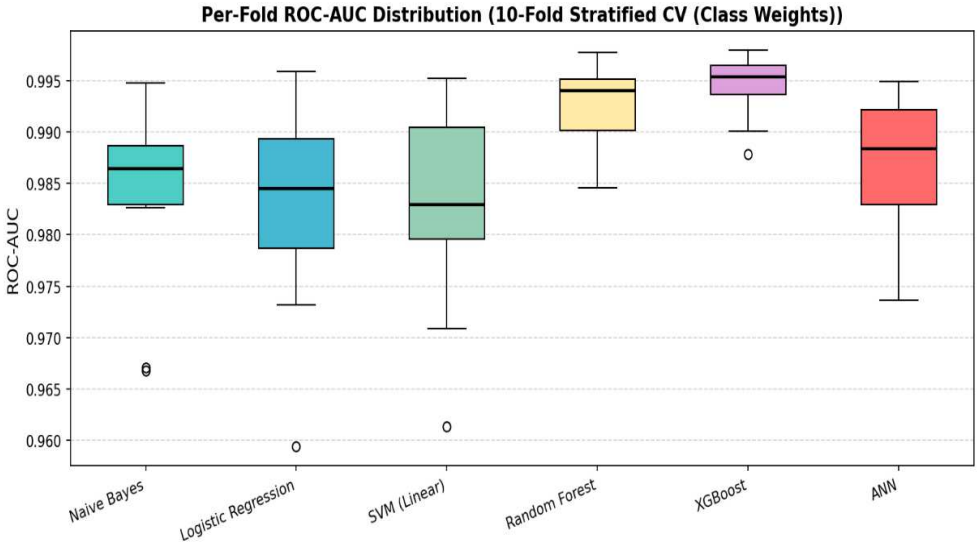

Figure4: Distribution and Variability of ROC-AUC Across 10-Fold Stratified Cross-Validation

Comparison of Five-Fold and Ten-Fold Cross-Validation Results

The comparison between stratified five-fold and ten-fold cross-validation demonstrated that the predictive performance was highly consistent across both validation strategies as seen in Table 1. Differences in ROC-AUC, accuracy, specificity and F1-score were negligible, and the ranking of the classifiers remained unchanged, with XGBoost consistently achieving the highest performance (0.9999), followed closely by Random Forest (0.9998). This consistency indicates that the models learned stable and generalizable patterns from the dataset rather than overfitting to specific training subsets. The small standard deviations observed across both validation schemes, particularly for XGBoost and Random Forest, further confirm the robustness and reliability of these ensemble models. Similar findings have been reported by [11] where ensemble learning algorithms maintained stable performance regardless of the cross-validation strategy employed, highlighting their superior ability to generalize to unseen data compared with conventional statistical and machine learning methods.

Model	**AUC**		**F1**		**Accuracy**		**Specificity**	
	5-Fold	**10-Fold**	**5-Fold**	**10-Fold**	**5-Fold**	**10-Fold**	**5-Fold**	**10-Fold**
Naïve Bayes	0.9986	0.988	0.936	0.935	0.9181	0.9169	**1**	**1**
Logistic Regression	0.9824	0.9823	0.953	0.9541	0.9369	0.9381	0.927	0.9312
ANN	0.9955	0.9967	0.98	0.9836	0.9729	0.9776	0.957	0.9665
Random Forest	**0.9998**	**0.9998**	**0.995**	**0.9944**	**0.9935**	**0.9923**	0.987	0.9832
XGBoost	**0.9999**	**0.9999**	**0.996**	**0.9974**	**0.9947**	**0.9965**	**0.993**	**0.9944**
SVM	0.9823	0.9823	0.993	0.9542	0.9369	0.9381	0.926	0.9293

Table 1: Comparison of 5-fold and 10-fold cross validation

Independent hold-out validation

In Table 2, the independent hold-out validation shows that XGBoost achieved the highest Matthews Correlation Coefficient (0.9048), indicating the best overall classification quality. XGBoost also had the lowest Brier score (0.0321), suggesting the most accurate probability calibration. Random Forest ranked second overall (0.856). ANN performed competitively (0.849). Logistic Regression (0.815) and SVM (0.808) showed comparatively lower performance. This demonstrates that the models generalize well beyond cross-validation.

Model	Paradigm	ROC-AUC	Accuracy	F1	Precision	Sensitivity	Specificity	MCC	Brier
ANN	Deep Learning	0.982	0.935	0.953	0.938	0.97	0.861	0.849	0.052
Random Forest	**Ensemble**	**0.991**	**0.938**	**0.956**	**0.924**	**0.991**	**0.824**	**0.856**	**0.045**
XGBoost	**Ensemble**	**0.993**	**0.959**	**0.97**	**0.95**	**0.991**	**0.889**	**0.905**	**0.032**
Logistic Regression	Linear	0.974	0.921	0.943	0.929	0.957	0.843	0.815	0.058
SVM	Linear	0.974	0.917	0.94	0.932	0.948	0.852	0.808	0.059
Naive Bayes	Probabilistic	0.979	0.926	0.946	0.944	0.948	0.88	0.83	0.052

Table2: Independent hold-out validation across all models

Wilcoxon comparisons versus Logistic Regression

Wilcoxon comparisons in Table 3 below using Logistic Regression as the baseline revealed differences in statistical significance between the two cross-validation strategies. In the 5-fold analysis, none of the alternative models differed significantly from Logistic Regression, although Naïve Bayes, Random Forest, XGBoost and ANN all produced p=0.0625. In contrast, the 10-fold analysis identified statistically significant differences between Logistic Regression and Naïve Bayes, Random Forest, XGBoost and ANN p=0.002 for each, whereas the difference between Logistic Regression and linear SVM remained non-significant p=0.846. These findings suggest that the larger number of paired validation estimates provided by 10-fold cross-validation improved the ability to detect systematic differences in model performance. When considered together with the higher mean performance observed for the ensemble models, particularly XGBoost and Random Forest, the findings provide evidence that their performance advantage over Logistic Regression is consistent across repeated validation folds

Model	P-Value(5-fold)	P-Value (10-fold)
Naïve Bayes	0.0625	0.002
SVM	**1**	**0.846**
Random Forest	0.0625	0.002
XGBoost	0.0625	0.002
ANN	0.0625	0.002

Table 3: Wilcoxon comparisons versus Logistic Regression as the baseline

Discussion of results

In this section, we provide a description of the results of this study. All the six machine learning algorithms achieved excellent predictive performance with ROC-AUC values exceeding 0.97 across both five-fold and ten-fold cross-validation. The narrowest interquartile range of XGBoost further indicates superior predictive accuracy and exceptional stability across different training and validation partitions. The superior performance of the ensemble learning algorithms suggests that the relationships between the demographic, reproductive, hereditary, lifestyle, and clinical risk factors in the dataset are complex and nonlinear, making tree-based ensemble methods more effective than conventional statistical models in capturing interactions among predictors [12]. Furthermore, the consistently small standard deviations and similar performance observed under both cross-validation strategies indicate that the models are robust and generalizable rather than overfitted to specific data partitions [13].

The comparison between stratified five-fold and ten-fold cross-validation demonstrated that the predictive performance of all six machine learning models was highly consistent across both validation strategies. Differences in ROC-AUC, accuracy, sensitivity, specificity, precision, F1-score, and Matthews Correlation Coefficient were negligible, and the ranking of the classifiers remained unchanged, with XGBoost consistently achieving the highest performance, followed closely by Random Forest. The small standard deviations observed across both validation schemes, particularly for XGBoost and Random Forest, further confirm the robustness and reliability of these ensemble models. Similar findings have been reported in previous breast cancer prediction studies, where ensemble learning algorithms maintained stable performance regardless of the cross-validation strategy employed, highlighting their superior ability to generalize to unseen data compared with conventional statistical and machine learning methods[12]. Furthermore, methodological studies on model evaluation have demonstrated that both five-fold and ten-fold cross-validation provide reliable estimates of predictive performance when models are well-specified and datasets are appropriately stratified (Rodríguez et al., 2010).

Feature Importance across all models

Hyperplasia was the most important feature across Naïve Bayes, Random Forest, XGBoost, and the ANN reflecting its well-established association with increased breast cancer risk[14]. Similarly, the high importance assigned to alcohol consumption, tobacco use, BMI, and body weight is consistent with epidemiological evidence identifying these factors as major modifiable risk factors for breast cancer [15], [16], while the prominence of parity and breastfeeding duration supports previous findings on the protective role of reproductive factors [17]. The differences in feature importance across models further suggest that different machine learning paradigms learn distinct patterns from the same clinical data, with ensemble and neural network models capturing more complex interactions among risk factors than conventional linear models. Overall, the SHAP analysis demonstrates that the models relied on biologically plausible predictors, thereby enhancing the transparency and interpretability in breast cancer prediction prediction[18].

Performance of Ensemble Learning Models

The superior performance of XGBoost and Random Forest may be explained by the characteristics of the predictor variables included in the Cuban breast cancer dataset. Unlike image-based breast cancer studies that rely on radiological or histopathological features, this study utilized demographic, reproductive, lifestyle, hereditary, and clinical characteristics to predict breast cancer risk. Variables such as age, family history of breast cancer, reproductive history, menopausal status, hormonal exposure, smoking and alcohol consumption do not act independently in determining breast cancer risk. Instead, these factors interact in complex and often nonlinear ways throughout a woman's lifetime. For example, increasing age may modify the influence of menopausal status, while family history may amplify the effect of hormonal or reproductive factors. Similarly, lifestyle factors may exert different effects depending on an individual's reproductive history or underlying clinical conditions. Such interactions are difficult to represent adequately using conventional linear statistical models.

Tree-based ensemble algorithms are particularly effective in modelling these complex relationships because they recursively partition the data into homogeneous subgroups without requiring prior assumptions regarding the functional relationships among predictor variables. Random Forest identifies multiple decision pathways by aggregating numerous independently generated trees, whereas XGBoost sequentially refines these decision pathways by correcting errors from preceding trees. Consequently, these algorithms can automatically identify clinically meaningful interactions among risk factors that may not be apparent using traditional regression techniques. The excellent predictive performance achieved by both ensemble models therefore suggests that breast cancer risk among Cuban women is influenced by combinations of interacting risk factors rather than isolated predictors.

This interpretation is supported by current epidemiological evidence indicating that breast cancer development results from the cumulative interaction of genetic susceptibility, reproductive history, hormonal influences, environmental exposures, lifestyle behavior, and ageing processes rather than from a single dominant determinant[19]. These findings therefore reinforce the value of ensemble learning methods for analyzing population-specific clinical datasets, particularly where disease risk is influenced by multiple interdependent determinants.

Performance of Machine Learning Models

Although Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Artificial Neural Network (ANN) all achieved excellent predictive performance (ROC-AUC > 0.97), they consistently performed slightly below XGBoost and Random Forest. Logistic Regression was limited by its assumption of linear relationships between predictor variables and breast cancer risk, while Naïve Bayes assumes conditional independence among predictors, an assumption that is rarely satisfied in clinical datasets. Similarly, the performance of SVM and ANN is highly dependent on parameter optimization and training data characteristics, which may explain their comparatively lower accuracy and stability. Nevertheless, the strong performance of these models demonstrates that the clinical and epidemiological variables used in this study are highly informative for breast cancer prediction, regardless of the learning algorithm. Similar observations have been reported

by[12] who found that Logistic Regression and SVM achieved high predictive performance but were consistently outperformed by ensemble learning methods due to their limited ability to model complex feature interactions. Similarly, [20] reported that although ANN effectively captured nonlinear relationships, XGBoost and Random Forest demonstrated superior discrimination and generalization on structured clinical data

Conclusion

In this paper, six machine learning algorithms were evaluated comprehensively using both five-fold and ten-fold stratified cross-validation, independent hold-out validation, multiple performance metrics, and statistical significance testing, providing a robust assessment of model performance. However, this study has limitations. The models were developed and evaluated using the Cuban dataset, which may limit their generalizability to other populations with different demographic and clinical characteristic. Future studies should evaluate these models in larger, multi-center and multi-ethnic populations, investigate additional risk factors and multimodal data sources, and assess the integration of the best-performing models into real-world clinical decision support systems to support personalized breast cancer risk prediction.

Declarations

Ethics approval

Not applicable since it is a publicly available dataset

Consent to participate

Not applicable since it is a publicly available dataset

Consent for publication

Not applicable. This manuscript does not contain any individual person's identifiable data. All the authors consented to the publication of this manuscript.

Availability of data and materials

Data used in this manuscript is publicly available on https://data.mendeley.com/datasets/7jhddnpz2p/1

Competing interests

The authors declare that they have no competing interests, financial or non-financial, that could have influenced the research, analysis, interpretation of the findings, or preparation of this manuscript.

Funding Declaration

The authors declare that there was no funding received for this manuscript

Authors' contributions

PA: Wrote the main manuscript and prepared the figures

JNN: Offered research guidance and reviewed the manuscript

IMS: Offered research guidance and reviewed the manuscript

RW: Offered research guidance and reviewed the manuscript

References:

[1] L. A. Torre, F. Islami, R. L. Siegel, E. M. Ward, and A. Jemal, "Global Cancer in Women: Burden and Trends," *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, vol. 26, no. 4, pp. 444–457, Apr. 2017, doi: 10.1158/1055-9965.EPI-16-0858.

[2] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine (Baltimore)*, vol. 103, no. 3, p. e36905, Jan. 2024, doi: 10.1097/MD.0000000000036905.

[3] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review," *Cancers*, vol. 13, no. 17, p. 4287, Aug. 2021, doi: 10.3390/cancers13174287.

[4] "Causes and Risk Factors of Breast Cancer, What Do We Know for Sure? An Evidence Synthesis of Systematic Reviews and Meta-Analyses." Accessed: May 12, 2026. [Online]. Available: https://www.mdpi.com/2072-6694/16/8/1583

[5] A. La Moglia and K. Mohamad Almustafa, "Breast cancer prediction using machine learning classification algorithms," *Intell.-Based Med.*, vol. 11, p. 100193, Jan. 2025, doi: 10.1016/j.ibmed.2024.100193.

[6] M. Dianati-Nasab *et al.*, "Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study," *Front. Oncol.*, vol. 13, Feb. 2024, doi: 10.3389/fonc.2023.1276232.

[7] J. Sun, C.-K. Sun, Y.-X. Tang, T.-C. Liu, and C.-J. Lu, "Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification," *Healthcare*, vol. 11, no. 14, p. 2000, Jul. 2023, doi: 10.3390/healthcare11142000.

[8] R. Gonzalez, P. Nejat, A. Saha, C. J. V. Campbell, A. P. Norgan, and C. Lokker, "Performance of externally validated machine learning models based on histopathology images for the diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer: A systematic review," *J. Pathol. Inform.*, vol. 15, p. 100348, Dec. 2024, doi: 10.1016/j.jpi.2023.100348.

[9] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," presented at the Neural Information Processing Systems, May 2017. Accessed: Aug. 10, 2026. [Online]. Available:

https://www.semanticscholar.org/paper/A-Unified-Approach-to-Interpreting-Model-Lundberg-Lee/442e10a3c6640ded9408622005e3c2a8906ce4c2
[11] "How Ensemble Learning Balances Accuracy and Overfitting: A Bias-Variance Perspective on Tabular Data," ResearchGate. Accessed: Jul. 30, 2026. [Online]. Available: https://www.researchgate.net/publication/398430759_How_Ensemble_Learning_Balances_Accuracy_and_Overfitting_A_Bias-Variance_Perspective_on_Tabular_Data
[12] T. Islam *et al.*, "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Sci. Rep.*, vol. 14, p. 8487, Apr. 2024, doi: 10.1038/s41598-024-57740-5.
[13] "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," ResearchGate. Accessed: Jul. 30, 2026. [Online]. Available: https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection
[14] L. C. Hartmann *et al.*, "Benign Breast Disease and the Risk of Breast Cancer," *N. Engl. J. Med.*, vol. 353, no. 3, pp. 229–237, Jul. 2005, doi: 10.1056/NEJMoa044383.
[15] "Alcohol, tobacco and breast cancer – collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease," *Br. J. Cancer*, vol. 87, no. 11, pp. 1234–1245, Nov. 2002, doi: 10.1038/sj.bjc.6600596.
[16] B. Lauby-Secretan *et al.*, "Body Fatness and Cancer--Viewpoint of the IARC Working Group," *N. Engl. J. Med.*, vol. 375, no. 8, pp. 794–798, Aug. 2016, doi: 10.1056/NEJMsr1606602.
[17] Collaborative Group on Hormonal Factors in Breast Cancer, "Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease," *Lancet*, vol. 360, no. 9328, pp. 187–195, Jul. 2002, doi: 10.1016/S0140-6736(02)09454-0.
[18] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.
[19] W. Y. Chen and G. A. Colditz, "Risk factors and hormone-receptor status: epidemiology, risk-prediction models and treatment implications for breast cancer," *Nat. Clin. Pract. Oncol.*, vol. 4, no. 7, pp. 415–423, Jul. 2007, doi: 10.1038/ncponc0851.
[20] "Breast cancer prediction modeling based on SHAP interpretability analysis and XGBoost algorithm," ResearchGate. Accessed: Jul. 30, 2026. [Online]. Available: https://www.researchgate.net/publication/390880576_Breast_cancer_prediction_modeling_based_on_SHAP_interpretability_analysis_and_XGBoost_algorithm