A Statistically Validated and Interpretable Machine Learning Framework for Breast Cancer Classification Using Shape Descriptors

K. A. Sotonwa[1*], J. O. Akintayo[1], Alex Pearson[2], Adenike Sofoluwe[3], Samson Arekete[4], Steffen Sammet[5], Funmilayo Olopade[2,5], Benjamin Aribisala[1,2]

[1]Department of Computer Science, Lagos State University, Nigeria
[2]Department of Medicine, University of Chicago, Chicago, USA
[3]Department of Radiology, University of Ibadan College of Medicine, Ibadan, Nigeria
[4]Department of Computer Science, Redeemer's University, Ede, Nigeria
[5]Center for Clinical Cancer Genetics and Global Health, University of Chicago; Chicago, USA

ABSTRACT: Breast cancer (BC) remains a major global health priority, with nearly half a million invasive cases reported in 2024. Early diagnosis is critical, and Machine Learning (ML) offers significant potential for automated detection. This study evaluated ML models using tumour shape descriptors from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (569 cases: 357 benign, 212 malignant). Ten numerical descriptors, including radius, texture, and concave points, were extracted from digitized fine needle aspirates (FNA). The dataset was split 70:30 into training and test sets, with Chi-square feature selection, mutual information (MI), and ANOVA feature selection applied to reduce dimensionality and prevent data leakage. Three models: Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Logistic Regression (LR) were trained using stratified cross-validation and evaluated on the independent test set. LR achieved the highest ROC (97.31%), significantly outperforming ANN ($p = 0.009$) and CART ($p = 0.001$) by DeLong's test, while McNemar's test revealed no significant differences in misclassification rates ($p > 0.05$). CART achieved the best F1-score, reflecting balanced sensitivity and specificity. These findings demonstrated that statistically validated and interpretable ML models provide robust diagnostic performance for BC classification

KEYWORDS: Breast Cancer, Chi-Square, Feature Selection, Machine Learning Algorithms, Shape Descriptor

[Received Dec. 06, 2025; Revised Feb. 07, 2026; Accepted Mar. 17, 2026] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

Breast cancer (BC) remains one of the leading causes of cancer-related mortality among women worldwide. Early detection is critical, as timely diagnosis significantly improves survival rates and reduces treatment costs. The global burden of BC underscores the need for reliable diagnostic frameworks that can support clinicians in identifying malignant cases at the earliest possible stage (Wilkinson & Gathani, 2022; Arnold *et al.,* 2022).

Fine-needle aspiration (FNA) cytology is widely used as an initial diagnostic tool due to its minimally invasive nature and cost-effectiveness. However, its interpretation is often subjective and dependent on clinical expertise, which can lead to variability in diagnostic accuracy. These limitations highlight the importance of developing automated, reproducible approaches that can complement clinical judgment and reduce diagnostic uncertainty (Kim *et al.,* 2025; Arnold *et al.,* 2022; Wilkinson & Gathani, 2022). Machine learning (ML) techniques have been increasingly applied to medical diagnosis, including BC classification. Although numerous studies have utilized the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, most focus primarily on predictive performance benchmarking using diverse classifiers rather than systematic evaluation of feature selection strategies, rigorous statistical comparison of models and leakage prevention during training (Huang *et al.,* 2008; Hou, 2020; Eserner *et al.,* 2023). Prior research has shown that simpler linear models often perform comparably to, or better than, more complex architectures when applied to structured clinical datasets, due to reduced overfitting and enhanced interpretability (Arshad *et al.,* 2023; Christodoulou *et al.,* 2019; Rashmi *et al.,* 2022). Furthermore, feature optimization has been shown to enhance the performance of tree-based classifiers by reducing the influence of non-informative variables (Modi & Ghanchi, 2016; Camattari *et al.,* 2024).

While the WDBC dataset has been widely used ML research, this study focuses specifically on the diagnostic relevance of shape descriptors derived from FNA cytology and the development of an interpretable ML pipeline suitable for clinical decision support. In context, shape-based morphological descriptors derived from digitized FNA samples provide structured, interpretable representations of tumour characteristics. Unlike high-dimension imaging pipelines that rely on deep learning feature extraction (Ekpo *et al.,* 2018; Park *et al.,* 2025; Ahsan *et al.,* 2022; Arkoudis & Papadakos, 2025; Esteva *et al.,* 2017; Litjens *et al.,* 2017; McKinney *et al.,* 2020; Zhou *et.al.,* 2024), structured

*Corresponding author: kehindesotonwa8@gmail.com ***doi:*** http://doi.org/10.63746/njtd.v23i1.4248

geometric features such as radius, perimeter, concavity and compactness capture biologically meaningful tumour morphology in a transpraent and computaionally efficient manner.

Therefore, this study aims to develop and statistically validate an interpretable ML framework for BC classification using morphological shape descriptors derived from cytological analysis. Artificial neural networks (ANN), classification and regression trees (CART), and logistic regression (LR) are systematically compared, with Chi-square feature selection applied to enhance model efficiency while preventing data leakage. Unlike many prior WDBC studies that emphasize maximizing predictive accuracy, this work prioritizes statistical validation of model differences using DeLong and McNemar tests, strict leakage prevention, and reporting of 95% Confidence Intervals (CI) through bootstrapping. This methodological rigor strengthens the reproducibility and clinical reliability of the proposed framework and aligns with global efforts to improve early BC detection. In addition to Chi-square, we also evaluate Mutual Information (MI) and Analysis of Variance (ANOVA) feature selection methods to compare their diagnostic utility.

II. MATERIALS AND METHODS

A. Dataset and Preprocessing

The dataset used in this study was obtained from the WDBC Repository and comprises numerical shape descriptors derived from digitized FNA images of 569 subjects (357 benign and 212 malignant cases). The dataset contains 30 numerical features derived from ten fundamental morphological descriptors include radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension – each computed as mean, standard error, and worst (largest) values. All numerical descriptors were standardized using z-score normalization to ensure comparable scales and variance across features. (Fogliatto *et al.,* 2019; Patro & Sahu, 2015). As the dataset contained no missing values, imputation was not required (Japkowicz & Shah, 2011; Joel *et al.,* 2024; Ramosaj *et al*., 2020).

B. Data Partitioning and Feature Selection

The dataset was first divided into 70% training (n = 398) and 30% independent testing (n = 171) sets using stratified sampling to preserve class distribution. Within the training set, hyperparameter tuning was conducted using stratified 5-fold cross-validation to ensure robust parameter optimization while maintaining class balance. Final model performance was subsequently evaluated on the independent test set to provide an unbiased measure of generalization.

To prevent data leakage, feature selection was applied exclusively to the training data. For the Chi-square test, continuous variables were discretized using equal width binning approximate categorical associations with class labels (Camattari *et al.,* 2024; Modi & Ghanchi, 2016). The Chi-square statistic was computed according to the workflow algorithm:

```
Input: Training dataset D with features F =
{f₁, f₂, ..., fₙ}, Labels Y
Output: Selected feature subset S

For each feature fi in F:
    Compute χ² statistic between fi and F
    Rank features by χ² score

Select top k features with highest χ² values
Return S
```

The algorithm begins with a training dataset D, consists of multiple input features F which represent the predictor variable, and a target label variable Y represent the class or output that the ML models aim to predict. The algorithm is to produce a subset S containing only the most relevant feature expected to contribute the most toward predicting the target variable while eliminating irrelevant or redundant features. The algorithm evaluates every feature f_1 in F individually and each feature is processed independently, so its relationship with the target variable is assessed without considering interaction with other features.

Chi-square (χ^2) measures the degree of association between the feature and the target variable which compare the observed frequencies with the expected frequencies under the assumption that the feature and the target are independent. The statistical is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

Where: O_{ij} = observed frequency of feature i with class j
E_{ij}= expected frequency under independence assumption.

Features with higher χ^2 values were ranked as more predictive of malignancy. Models were trained using both the full feature set and the Chi-square selected subset. Additional robustness checks, MI and ANOVA F-test feature selection methods were applied to the training set to identify statistically significant and informative descriptors. An alternative 80:20 stratified train-test split was further evaluated to confirm the stability of results (Sarwar *et al.,* 2018).

C. Model Configuration and Hyperparameters

To ensure reproducibility, all model architectures, *hyperparameters*, and tuning procedures are summarized in Table 1. Hyperparameter tuning was performed using grid search with stratified 5-fold cross-validation on the training set. The parameter ranges explored for each of the model are detailed below. Models were trained using both the full feature set and the Chi-square selected subset. Additional robustness was applied to the training set to identify statistically significant and informative descriptors. An alternative 80:20 stratified train-test split further evaluated to confirm the stability of results (Sarwar *et al.,* 2018.

D. Performance Evaluation and Statistical Testing

Performance metrics (sensitivity, specificity, accuracy, F1-score and AUC) were reported with 95% CI, estimated using stratified bootstrapping with 1,000 resamples of the test set predictions. To assess whether difference between models were statistically significant, DeLong's test was applied to compare AUCs, while McNemar's test was used to compare paired classification outcomes. In addition, confusion matrices were generated for each model under all feature selection methods to provide a detailed breakdown of True Negatives (TN), False Positives (FP), False Negatives

(FN), and True Positives (TP). All statistical analyses were conducted on predictions from the independence test set to ensure unbiased evaluation of model generalization.

Table 1 Model architectures, hyperparameters and tuning procedures for ANN, CART and LR

Model	Architecture	Regularization /Pruning	Optimizer	Hyper-parameter (values or ranges)	Tuning Methods
ANN	3 hidden layers (32-16-8 neurons), ReLU activations in the first two hidden layers and Sigmoid activation in the third hidden layer.	Dropout (0.2), Early stopping	Adam	Learning rate = {0.001, 0.01}, Batch size = {16, 32}, Epochs = up to 100	Grid search, 5-fold CV
CART	Gini index splitting	Cost-complexity pruning (α)	-	Max depth = {3, 5, 7}, Min samples split = {2, 5, 10}, Pruning α = {0.001, 0.01, 0.1}	Grid search, 5-fold CV
LR	-	L2 penalty	Solver = liblinear	Regularization strength C = {0.01, 0.1, 1, 10}	Grid search, 5-fold CV

E. Model Implementation

The ANN was implemented as a feed-forward neural network with three hidden layers containing 32, 16, and 8 neurons, respectively. ReLU activation was applied to the first two hidden layers, and sigmoid activation was applied to the third layer and output node. The network was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32 for a maximum of 100 epochs. Dropout (rate = 0.2) was applied after each hidden layer, and early stopping based on validation loss was used to mitigate overfitting. CART was implemented using the Gini index as the splitting criterion, with cost-complexity pruning applied after tree growth to prevent overfitting.

LR was implemented with L2 regularization using the liblinear solver to manage coefficient magnitude and reduce overfitting.

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}} \quad (2)$$

Where:
y = predicted class (malignant = 1, benign = 0),
$x = (x_1, x_2, \ldots, x_k)$ = selected shape descriptors,
β_0 = intercept, and
β_i = model coefficients estimated during training.

F. Cross Validation and Supplementary Models

Stratified 5-fold cross-validation was employed within the training set for hyperparameter tuning and model selection. This approach ensured robust parameter optimization while preserving class distribution. Final model performance metrics were computed exclusively on the independent 30% test set to provide an unbiased evaluation of model generalization. Ensemble models such as Random Forest (RF) and Gradient Booting (GB) were not included in the study, as the analysis focused exclusively on the selected baseline and comparative models. All experiments were implemented in Python using the Scikit-learn and TensorFlow libraries. A fixed random seed (42) was used across all experiments to ensure reproducibility.

III. RESULTS AND DISCUSSION

Chi-square feature selection, applied after discretization of continuous variables, identified seven key diagnostic features: perimeter, area, radius, texture, concavity, concave points, and compactness. To demonstrate the diagnostic relevance of these descriptors, we compared descriptive statistics for five of these features between benign and malignant tumours in the WDBC dataset (Table 2).

Malignant cases consistently showed higher values across radius, area, perimeter, concavity, and concave points, reflecting uncontrolled cellular growth and irregular nuclear margins. This analysis demonstrates that simple geometric descriptors capture biologically meaningful differences in tumour morphology. Their alignment with established pathological observations supports their utility in predictive modelling for early BC detection. Collectively, these features represent morphological characteristics most significantly associated with malignancy in digitized FNA images.

In addition to Chi-square, we also evaluated MI and ANOVA feature selection methods. Including these approached allows comparison of different statistical criteria for identifying informative descriptors and ensures that the observed model performance is robust across multiple feature selection strategies. The Chi-square statistics for these seven features showed statistically significant association with malignancy classification ($p < 0.01$ for all selected descriptors), supporting their inclusion in the reduced features model.

To further illustrate model performance, confusion matrices were generated for the three classifiers (ANN, CART, and LR) using the independent test set. These matrices provide a detailed breakdown of TP, FP, TN, and FN, thereby complementing the summary metrics reported in Table 1. The confusion matrices for all feature selection methods, including the full features set, Chi-square, MI and ANOVA are presented in Table 3, enabling direct

comparison of classifier error distributions across different feature subnets.

To further illustrate model performance, confusion matrices were generated for the three classifiers (ANN, CART, and LR) using the held-out test set. These matrices provide a detailed breakdown of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), thereby complementing the summary metrics reported in Table 1. The confusion matrices are shown in Table 3.

Table 2 Descriptive statistics of key shape features (benign vs. malignant, WDBC dataset)

Feature	Benign (Mean ± SD)	Malignant (Mean ± SD)	Clinical Interpretation
Radius	12.1 ± 1.8	17.5 ± 3.2	Malignant tumours tend to have larger nuclei sizes
Perimeter	84.0 ± 6.5	115.4 ± 13.0	Larger perimeter indicates irregular tumour boundary
Area	460 ± 134	978 ± 367	Larger tumour mass in malignant cases
Concavity	0.03 ± 0.02	0.18 ± 0.08	High concavity reflects irregular margins
Concave	0.03 ± 0.02	0.14 ± 0.05	Indicates speculated tumour shape

Table 3 Confusion matrices by feature selection methods

Feature selection	Model	TN	FP	FN	TP
All features	ANN	100	8	12	51
	CART	102	6	5	58
	LR	101	7	11	52
Chi-square	ANN	100	8	12	51
	CART	102	6	5	58
	LR	101	7	11	52
Mutual information	ANN	99	9	21	42
	CART	103	5	5	58
	LR	102	6	12	51
ANOVA	ANN	81	27	18	45
	CART	101	7	7	56
	LR	103	5	12	51

Table 4 presents performance metrics with 95% CI for ANN, CART, and LR models trained on both the full feature set and the Chi-square selected subset. Chi-square feature selection identified seven descriptors as most predictive: concave points, concavity, perimeter, area, radius, compactness, and texture as the most predictive descriptors. These features were consistently ranked highest by their Chi-square scores, indicating strong statistical association with malignancy classification. DeLong's test demonstrated that LR statistically outperformed ANN ($Z = 2.59$, $p = 0.009$) and CART ($Z = -3.33$, $p = 0.0009$) when all features were used. McNemar's test however, indicated no statistically significant differences in misclassification rates between the models ($p > 0.05$), suggesting comparable classification error patterns despite differences in ranking performance.

DeLong's test demonstrated that LR statistically outperformed ANN ($Z = 2.59$, $p = 0.009$) and CART ($Z = -3.33$, $p = 0.0009$) when all features were used. McNemar's test however, indicated no statistically significant differences in misclassification rates between the models ($p > 0.05$), suggesting comparable classification error patterns despite differences in ranking performance. Figure 1 illustrates the ROC curves for the three models, with panel (a) showing performance using all feature selection and panel (b) showing performance using Chi-square feature selection. The ROC curves visually confirm the strong discriminative ability of all models, consistent with the AUC values reported in Table 4. For completeness, ROC curves were also generated for models training using MI and ANOVA feature selection (Figure 2a and 2b). Under MI, CART and LR maintained strong discriminative ability, with LR achieving the highest AUC (95.71%) and CART showing balanced sensitivity and specificity.

ANN, however, exhibited reduced performance, with lower sensitivity and a higher rate of FN, consistent with its confusion matrix results. Under ANOVA, LR again achieved superior AUC (96.43%), while CART demonstrated competitive performance with balanced classification outcomes. ANN performed weakest under ANOVA, with increased FP and reduced specificity. These ROC curves confirm that while Chi-square provided the most consistent improvements across models, MI and ANOVA highlighted important trade-offs, particularly the vulnerability of ANN to feature selection variability.

The reported AUCs are consistent with prior WDBC studies, underscoring that our contribution lies in methodological rigor rather than surpassing predictive benchmarks. This study demonstrates how best practices in feature selection, leakage prevention, and model interpretability can support clinically relevant ML pipelines. Our reported AUCs (LR: 97.31%, ANN: 94.44%) are consistent with benchmark studies using SVMs and RF which typically achieve AUCs in the 94–98% range (Huang *et al.,* 2008; Hou, 2020). This reinforces that our contribution lies not in exceeding predictive benchmarks but in demonstrating a reproducible, clinically-aware ML pipeline emphasizing feature selection, leakage prevention, and interpretability.

Upon implementing Chi-square feature selection, the models maintained high diagnostic performance despite the reduction in dimensionality. MI and ANOVA provided complementary perspectives: MI highlighted descriptors with strong mutual independence on malignancy, while ANOVA identified features with statistically significant variance between benign and malignant groups.

Table 4 ML performance metrics for the three ML models

Features Set	Metrics	ANN (%)	CART (%)	LR (%)
All features	Sensitivity	77.78 (95 CI: 66.15-87.10)	88.89 (95 CI: 80.77-96.23)	82.54 (95 CI: 72.88-90.91)
	Specificity	92.59 (95 CI: 87.27-97.22)	92.59 (95 CI: 87.23-97.27)	93.52 (95 CI: 88.29-98.06)
	Accuracy	87.13 (95 CI: 81.29-91.81)	91.23 (95 CI: 86.55-95.32)	89.47 (95 CI: 84.80-94.15)
	F1 Score	81.67 (95 CI:72.53-88.24)	88.19 (95 CI:81.67-93.34)	85.25 (95 CI:77.06-91.47)
	AUC	94.44 (95 CI: 90.21-97.47)	90.74 (95 CI: 86.08-95.12)	97.31 (95 CI: 94.86-98.94)
Chi-square selected	Sensitivity	80.95 (95 CI: 69.64-89.71)	92.06 (95 CI: 85.00-98.11)	82.54 (95 CI: 72.88-90.91)
	Specificity	92.59 (95 CI: 87.07-97.22)	94.44 (95 CI: 89.66-98.25)	93.52 (95 CI: 88.29-98.06)
	Accuracy	88.30 (95 CI: 83.04-92.89)	93.57 (95 CI: 89.47-97.08)	89.74 (95 CI: 84.80-94.15)
	F1 Score	83.61 (95 CI: 74.76-90.00)	91.34 (95 CI: 85.47-95.58)	85.25 (95 CI: 77.06-91.47)
	AUC	94.49 (95 CI: 90.20-97.60)	93.25 (95 CI: 89.14-96.78)	97.21 (95 CI: 94.71-98.90)
Mutual information (MI)	Sensitivity	66.67 (95 CI: 53.85-77.94)	92.06 (95 CI: 84.85-98.08)	80.95 (95 CI: 70.58-89.55)
	Specificity	91.69 (95 CI: 86.27-96.43)	95.37 (95 CI: 90.99-99.07)	94.44 (95 CI: 89.52-98.18)
	Accuracy	82.46 (95 CI: 76.61-87.72)	94.15 (95 CI: 90.06-97.08)	89.47 (95 CI: 84.80-93.57)
	F1 Score	73.68 (95 CI: 63.55-81.89)	92.06 (95 CI: 86.54-96.12)	85.00 (95 CI: 76.78-90.91)
	AUC	85.82 (95 CI: 79.63-91.16)	93.72 (95 CI: 89.53-97.04)	95.71 (95 CI: 92.30-98.09)
ANOVA	Sensitivity	71.43 (95 CI: 59.09-81.67)	88.89 (95 CI: 80.77-96.23)	80.95 (95 CI: 70.37-90.00)
	Specificity	75.00 (95 CI: 66.99-83.17)	93.52 (95 CI: 88.29-98.00)	95.37 (95 CI: 91.15-99.07)
	Accuracy	73.68 (95 CI: 67.25-80.12)	91.81 (95 CI: 87.72-95.91)	90.06 (95 CI: 85.38-94.15)
	F1 Score	66.67 (95 CI: 56.66-75.00)	88.89 (95 CI:82.47-94.12)	85.71 (95 CI: 77.55-91.94)
	AUC	83.83 (95 CI: 77.30-89.66)	91.20 (95 CI: 86.58-95.56)	96.43 (95 CI: 93.52-98.52)

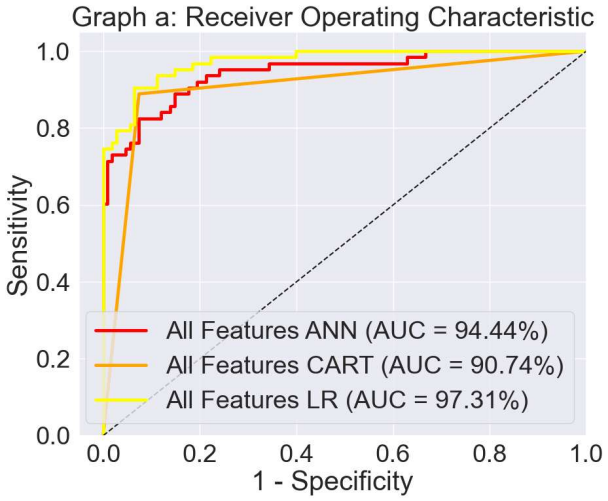

Figure 1(a): ROC curve using all feature selection

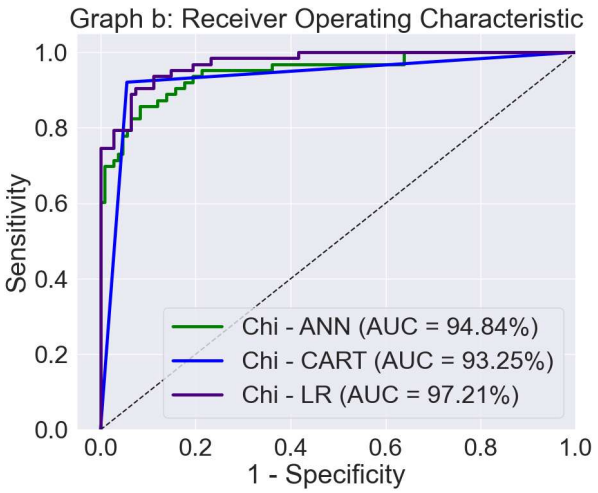

Figure 1(b): ROC curve using Chi-square feature selection

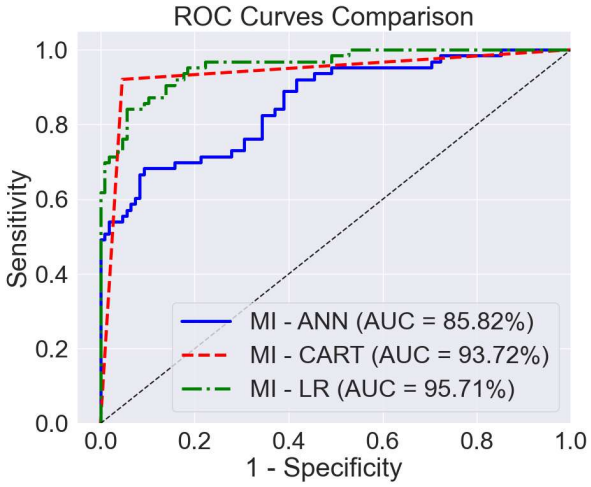

Figure 2(a): ROC curves using MI feature selection

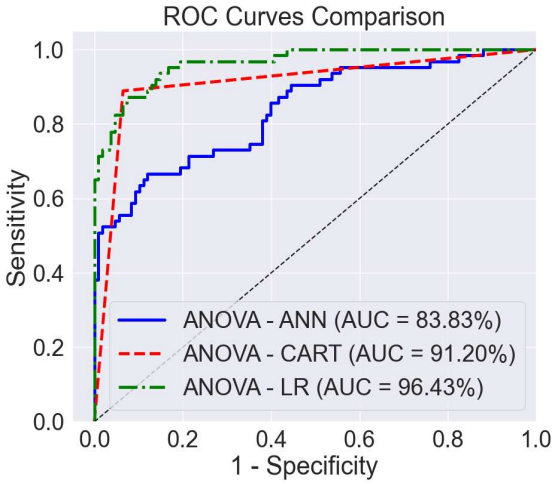

Figure 2(b): ROC curves using ANOVA feature selection

Together, these approaches confirmed the robustness of the selected descriptors and demonstrated that feature selection influences not only overall performance metrics but also distribution of classification errors across models. Specifically, the ANN achieved a sensitivity of 80.95%, specificity of 92.59%, accuracy of 88.30%, F1 score of 83.61%, and an AUC of 94.49%. The CART model showed the most significant improvement after feature selection, with metrics rising to 92.06% sensitivity, 94.44% specificity, 93.57% accuracy, F1 score of 91.34%, and an AUC of 93.25%. LR remained the top performer with an AUC of 97.21%, showing negligible performance loss while benefiting from improved computational efficiency.

Beyond numerical performance, the selected features align closely with established biomedical knowledge of breast tumour morphology. For instance, “concave points” and “concavity” capture irregular tumour margins, a hallmark of malignant lesions. Larger perimeter, area, and radius values correspond to increased tumour size, while compactness and symmetry reflect deviations from regular cellular architecture. These descriptors are consistent with pathological observations that malignant tumours often exhibit irregular, asymmetric forms with poorly defined boundaries. Thus, the feature selection process not only but improved model performance but also reinforced biological plausibility of the predictive variables.

IV. CONCLUSION

This study successfully developed and evaluated a BC prediction framework utilizing shape analysis and ML, grounded in morphological analysis. By extracting ten distinct descriptors from digitized FNA images, we confirmed that geometric tumour features provide a robust basis for automated diagnosis, supporting the broader literature on the efficacy of shape-based descriptors in oncology. Comparative analysis revealed that LR achieved the highest AUC, indicating superior ranking performance, while CART demonstrated competitive performance in terms of F1 score. This suggests that model choice may depend on whether ranking discrimination or balanced classification performance is prioritized.

These findings align with Arshad *et al.,* (2023) and Rashmi *et al.,* (2022), who argued that simpler linear models often exhibit strong robustness and interpretability when applied to structured clinical datasets. Furthermore, the application of Chi-square feature selection successfully reduced the feature space to seven key descriptors without compromising predictive power, reinforcing the importance of feature optimization in enhancing model efficiency. In conclusion, the integration of shape descriptors with optimized machine learning workflows specifically, LR provides a practical, computationally efficient, and clinical interpretable approach for BC classification. These findings

emphasize the importance of model selection and feature reduction in clinical settings where computational resources may be limited. This study contributes a reliable, reproducible, and highly accurate framework that aligns with existing global research efforts to improve early breast cancer detection through automated diagnostic tools.

V. LIMITATION AND FUTURE WORK

This study has several limitations that should be acknowledged, first, the WDBC dataset provides only pre-extracted numerical descriptors rather than original digitized FNA slide images. As a result, the framework could not be validated directly on raw image date, which may limit its applicability to real-world diagnostic clearance to evaluate the pipeline on full imaging data. Second, the dataset size is relatively modest, with limited sample diversity. While stratified sampling and cross-validation were employed to mitigate this constraint, medical models often require larger, multi-institutional datasets to ensure generalizability. Expanding the dataset and adopting alternative portioning strategies, such as 80:20 train test split, may provide more robust training while preserving sufficient test cases for evaluation.

Third, feature selection relied on Chi-square applied to discretized continuous variables. Although effective, discretization may reduce statistical precision. Alternative methods such as the ANOVA F-test, Mutual information, or embedded approached (e.g. Lasso regression) could provide more rigorous selection criteria for continuous descriptors. Future work should systematically compare these methods to assess whether they yield improve feature selection outcomes. Finally, feature augmentation represents a potential avenue for enhancing dataset size and diversity. Synthetic expansion of shape-based descriptors through perturbation could improve generalizability, but such techniques must be carefully validated to avoid introducing artificial bias. In this study, we prioritized reproducibility by using the original WDBC descriptors. Future research may explore augmentation strategies under expert supervision to determine whether they can strengthen model robustness without compromising clinical validity.

AUTHOR CONTRIBUTIONS

Dr. K. A. Sotonwa served as the corresponding author and was responsible for conceptualization**,** methodology**,** software development**,** validation**,** original draft preparation**,** writing**,** review and editing, and responses to reviewers' comments**. Mr. J. O. Akintayo** contributed to the methodology**,** software development, and validation of the study. He is a Ph.D. scholar under the DATICAN program. **Prof. B. S. Aribisala** provided supervision and contributed to writing, review, and editing of the manuscript from the initial draft through the submission stage. He also serves as the Program Director and Lead Principal Investigator of the Data Science Nigeria–Cancer Research Training Program (DATICAN). **Prof. O. I. Olopade** contributed through conceptual guidance**,** scientific oversight, and critical review and editing of the manuscript. She is a Principal Investigator in the DATICAN program.

ACKNOWLEDGEMENTS

We appreciate Data Science and Medical Image Analysis Training (DATICAN) and National Institute of Health (NIH) for their support. We also appreciate Mr. Mayowa Adeyemi of the Department of Computer Science, Lagos State University, Ojo, Lagos for helping during data collection and collation.

REFERENCES

Agarap, A. F. M. (2018). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. Paper presented at the Proceedings of the 2nd International Conference on Machine Learning and Soft Computing.

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. Paper presented at the Healthcare.

Arkoudis, N.-A., & Papadakos, S. P. (2025). Machine learning applications in healthcare clinical practice and research. World Journal of Clinical Cases, 13(1), 99744.

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., . . . Siesling, S. (2022). Current and future burden of breast cancer: Global Statistics for 2020 and 2040. The Breast, 66, 15-23.

Arshad, M. A., Shahriar, S., & Anjum, K. (2023). The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques--Case of Breast Cancer Diagnosis. arXiv preprint arXiv:2306.02449.

Awan, M. Z., Arif, M. S., Abideen, M. Z. U., & Abodayeh, K. (2024). Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach. Indonesian Journal of Electrical Engineering and Computer Science, 34(3), 2032-2044.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12-22.

Ekpo, E. U., Alakhras, M., & Brennan, P. (2018). Errors in mammography cannot be solved through technology alone. Asian Pacific Journal of Cancer Prevention: APJCP, 19(2), 291.

Esener, İ. I., Kara, Ş., Ergin, S., & Çalışır, C. A hybrid textural and geometrical feature extraction to reveal hidden information from suspicious regions on mammograms. Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering, 23(1), 70-86.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Fogliatto, F. S., Anzanello, M. J., Soares, F., & Brust-Renck, P. G. (2019). Decision support for breast cancer detection: classification improvement through feature selection. Cancer Control, 26(1), 1073274819876598.

Hou, Y. (2020). Breast cancer pathological image classification based on deep learning. Journal of X-ray Science and Technology, 28(4), 727-738.

Huang, Y. L., Chen, D. R., Jiang, Y. R., Kuo, S. J., Wu, H. K., & Moon, W. (2008). Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology, 32(4), 565-572.

Japkowicz, N., & Shah, M. (2011). Evaluating learning algorithms: a classification perspective: Cambridge University Press.

Joel, L. O., Doorsamy, W. & Paul, B. S. (2014). On the performance of imputation techniques for missing values on the healthcare datasets. arXiv preprint arXiv:2403.14687.

Kim, J., Harper, A., McCormack, V., Sung, H., Houssami, N., Morgan, E., Fidler-Benaoudia, M. M. (2025). Global patterns and trends in breast cancer incidence and mortality across 185 countries. Nature Medicine, 1-9.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Darzi, A. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

Park, J., Witowski, J., Xu, Y., Trivedi, H., Gichoya, J., Brown-Mulry, B., Lewin, A. (2025). A Multi-Modal AI System for Screening Mammography: Integrating 2D and 3D Imaging to Improve Breast Cancer Detection in a Prospective Clinical Study. arXiv preprint arXiv:2504.05636.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Petushi, S., Garcia, F. U., Haber, M. M., Katsinis, C., & Tozeren, A. (2006). Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Medical Imaging, 6, 1-11.

Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. Bioinformatics, 36(10), 3099-3106.

Rashmi, R., Prasad, K., & Udupa, C. B. K. (2022). Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review. Journal of Medical Systems, 46(1), 1-7.

Rezazadeh, A., Jafarian, Y., & Kord, A. (2022). Explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features. Forecasting, 4(1), 262-274.

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. Paper presented at the 2018 24th international conference on automation and computing (ICAC).

Tamrakar, D., & Ahuja, K. (2017). Density-wise two stage mammogram classification using texture exploiting descriptors. arXiv preprint arXiv:1701.04010.

Wilkinson, L., & Gathani, T. (2022). Understanding breast cancer as a global health concern. The British journal of radiology, 95(1130), 20211033.

Zhang, Y.-D., Wang, S.-H., Liu, G., & Yang, J. (2016). Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. Advances in Mechanical Engineering, 8(2), 1687814016634243.

Zhou, S., Hu, C., Wei, S., & Yan, X. (2024). Breast cancer prediction based on multiple machine learning algorithms. Technology in Cancer Research & Treatment, 23, 15330338241234791.