Technoscience Journal for Community Development in Africa 4 (2025) 187-197
ISSN 2659-0573

Research Article

A comparative machine learning framework for breast cancer diagnosis: Benchmarking algorithms and emphasizing model interpretability

Oni Patrick Adebayo[1*], Ibrahim Ahmed[2], Ibrahim Maiji Garba[3], Azeez Daramola Mustaph[4], Kamoru Tayo Oyeleke[5]

[1]Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria
[2]Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria
[3]Department of Agriculture, Phoenix University Agwada, Nasarawa State, Nigeria
[4]National Bureau of Statistics, Nigeria
[5]National Bureau of Statistics, Nigeria

Abstract: This study evaluated the performance of four machine learning models regularized logistic regression (GLMNET), random forest (RF), extreme gradient boosting (XGB), and support vector machine (SVM) for the binary classification of breast cancer cases using a dataset comprising 357 benign (62.7%) and 212 malignant (37.3%) samples. Model training and evaluation were performed using repeated cross-validation, with performance assessed through ROC, sensitivity, specificity, and accuracy. Among the models, GLMNET achieved the best performance, with the highest cross-validation ROC (0.992) and a strong balance between sensitivity (0.982) and specificity (0.935). On the independent test set, GLMNET demonstrated excellent discrimination (AUC = 0.998), high accuracy (98.2%, 95% CI: 93.8–99.8), sensitivity (98.6%), and specificity (97.6%), with a Kappa of 0.962 indicating near-perfect agreement. Feature importance analysis revealed PC02, PC01, and PC04 as the most influential predictors. These results suggest that GLMNET provides robust and highly accurate classification performance, making it a suitable model for breast cancer prediction in this dataset.

Keywords: GLMNET, Benign, Malignant, Principal Component Importance, Predictive Modeling

1. Introduction

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide, with an estimated 2.3 million new cases diagnosed globally in 2020 alone (Sung et al., 2020). Despite significant advancements in screening and treatment protocols, it continues to be a leading cause of cancer-related mortality, underscoring the critical need for early and accurate diagnosis. Timely detection is paramount, as it directly correlates with higher survival rates and a broader range of effective treatment options.

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into oncology has heralded a new era in medical diagnostics. ML algorithms demonstrate a remarkable capacity to identify complex, non-linear patterns within high-dimensional medical data, ranging from mammography and histopathological images to genomic and clinical patient records. These models offer the potential to augment the capabilities of healthcare professionals, serving as powerful decision-support systems to improve diagnostic accuracy, reduce false positives and negatives, and ultimately streamline clinical workflows (McKinney et al., 2020).

Consequently, a substantial body of research has emerged dedicated to applying various ML classifiers including Support Vector Machines (SVM), Random

* Corresponding author
Email: adebayo.patrick@phoenixuniversity.edu.ng

KWASU PRESS AND PUBLISHING

Forests, and advanced Deep Learning architectures like Convolutional Neural Networks (CNNs), to breast cancer classification tasks. Studies often report near-perfect accuracy on benchmark datasets such as the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. However, a significant gap persists between achieving high performance in a controlled experimental setting and deploying a trustworthy model in a real-world clinical environment. This gap is frequently driven by two interconnected challenges: the lack of rigorous, comparative benchmarking under standardized conditions and, more critically, the "black-box" nature of many sophisticated algorithms (Arrieta et al., 2020).

While a complex model may achieve superior accuracy, its inability to provide clinicians with intuitive, human-readable explanations for its predictions severely limits its adoption. A physician is unlikely to base a critical diagnosis on a model's output without understanding the reasoning behind it. Therefore, model interpretability, the ability to explain or present the rationale of an ML model in understandable terms, is not merely an academic exercise but a fundamental prerequisite for clinical translation (Angelov et al., 2021). Techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and feature importance analysis are becoming essential components of the ML pipeline, bridging the gap between model performance and clinical trust.

This study addresses these critical gaps by proposing a comprehensive machine learning framework for breast cancer diagnosis. Our work makes a dual contribution: first, we conduct a rigorous, standardized benchmark of a diverse suite of machine learning algorithms, from logistic regression to ensemble methods and neural networks, to identify the top-performing model objectively. Second, and more importantly, we move beyond pure accuracy metrics to place a strong emphasis on model interpretability. By integrating state-of-the-art explanation techniques, we elucidate the decision-making processes of our models, identifying the most influential features in the diagnostic prediction. This approach ensures that our framework not only excels in predictive performance but also provides the transparency necessary for building clinician confidence and paving the way for practical, ethical, and reliable integration of AI into breast cancer care.

The primary objective of this analysis is to establish a comprehensive and robust machine learning framework for the accurate and interpretable diagnosis of breast cancer. The study moves beyond a simple comparison of algorithms by implementing a rigorous, journal-quality pipeline designed to benchmark the performance of several advanced classifiers under standardized conditions. The core aim is to identify the optimal model that not only achieves superior predictive accuracy but also provides critical insights into its decision-making process, thereby bridging the gap between computational performance and clinical applicability.

This involves a systematic evaluation of regularized regression, ensemble methods, and support vector machines, with their performance meticulously assessed through repeated cross-validation using metrics paramount to medical diagnostics, such as AUC, sensitivity, and specificity. The final selected model undergoes a thorough independent evaluation on a held-out test set to confirm its generalizability and diagnostic prowess.

Furthermore, a paramount objective is to demystify the model's predictions by emphasizing interpretability. This is achieved by analyzing and visualizing feature importance to identify the key variables driving the classification of tumors as benign or malignant. The framework is designed to culminate in a reproducible and deployable pipeline, encompassing the entire workflow from data preprocessing and model training to final evaluation and artifact saving, ensuring the findings are both scientifically sound and practically valuable for potential clinical decision support.

2. Methodology

2.1. Dataset and preprocessing

The study utilized the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository, comprising 569 fine-needle aspiration samples with 30 morphological features (Wolberg et al., 1995). The dataset included measurements of cell nuclei characteristics including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, with each feature represented as mean, standard error, and worst values. The binary classification task distinguished malignant (212 cases) from benign (357 cases) samples, reflecting real-world clinical prevalence patterns in breast cancer screening populations (Buda et al., 2018).

Data preprocessing employed a comprehensive pipeline using the recipes package to ensure reproducibility and prevent data leakage. The preprocessing sequence included centering and scaling of all numeric features, removal of highly correlated predictors (threshold $r > 0.9$), elimination of near-zero variance features, and principal component analysis (PCA) retaining components explaining 95%

of cumulative variance (Kuhn & Johnson, 2019). This dimensionality reduction approach addressed multicollinearity while preserving biologically relevant information in the transformed feature space (Jolliffe & Cadima, 2016).

2.2. Experimental design and model training

The dataset was partitioned using stratified sampling into training (80%) and testing (20%) subsets, maintaining class distribution integrity across splits. We implemented a repeated cross-validation framework with 10 folds and 5 repetitions, generating 50 performance estimates per model to ensure robust generalization error estimation (Bischl et al., 2021). Parallel processing using four cores accelerated the computational workflow through the doParallel package.

Four state-of-the-art machine learning algorithms were systematically compared: regularized logistic regression (GLMNET), random forest (RF), extreme gradient boosting (XGBoost), and support vector machine with radial basis function kernel (SVM). Each algorithm underwent comprehensive hyperparameter optimization through grid search procedures. GLMNET explored α values (L1/L2 mixing) from 0 to 1 with multiple λ regularization strengths. Random Forest optimized mtry parameters and node sizes, while XGBoost tuned learning rates, tree depths, and subsampling ratios. SVM optimization focused on cost parameters and γ values for the RBF kernel (Probst et al., 2019).

2.3. Model interpretation framework

To address the critical need for clinical interpretability, we implemented a multi-faceted Explainable AI (XAI) framework. Variable importance analysis identified the most influential features for each model, while PCA loading examination connected principal components back to original clinical features (Molnar, 2020). For the best-performing model, we computed coefficients and odds ratios to provide clinically actionable insights into feature effects on malignancy probability (Lundberg & Lee, 2017).

The analytical workflow incorporated both PCA-transformed and original feature spaces to bridge the interpretability gap between statistical optimisation and clinical utility. This dual approach enabled high predictive performance through dimensionality reduction while maintaining direct interpretability of original morphological measurements relevant to pathological assessment (Rudin, 2019).

2.4. Performance evaluation

Model selection prioritised the area under the receiver operating characteristic curve (ROC-AUC) as the primary metric, which is particularly suitable for imbalanced medical diagnostic tasks (Ozenne et al., 2020). Comprehensive evaluation included sensitivity, specificity, precision, F1-score, and accuracy metrics. The final model assessment utilised the completely held-out test set, providing unbiased performance estimates for clinical translation potential.

All analyses were conducted in R version 4.5.2 using the caret, GLMNET, randomForest, xgboost, and DALEX packages, ensuring reproducibility through complete code availability and version-controlled environment management.

2.5. Data analysis

All statistical analyses and modelling were conducted in R version 4.5.2 (R Core Team, 2023) using a reproducible workflow. The analysis pipeline comprised four interconnected stages: (1) data preprocessing and exploratory analysis, (2) model training and hyperparameter tuning, (3) performance evaluation, and (4) model interpretation.

3. Result and Discussion

Table 1: Distribution and Proportions of Benign and Malignant Cases

Class	Frequency	Proportion
Benign (B)	357	0.627 (62.7%)
Malignant (M)	212	0.373 (37.3%)

From Table 1, the dataset used in this analysis consists of 357 benign cases and 212 malignant cases, representing 62.7% and 37.3% of the total observations, respectively. This distribution indicates that benign cases are more prevalent than malignant cases, although the imbalance is moderate rather than severe. The relative proportions suggest that while predictive models may have slightly more exposure to benign cases during training, there remains a substantial representation of malignant cases, allowing for meaningful discrimination between the two classes. This balance provides a reasonable basis for model development and evaluation, though care should still be taken to ensure that performance metrics account for the unequal class sizes to avoid bias toward the majority class.

From Table 2, the performance of four models, GLMNET (regularized logistic regression), random forest (RF), extreme gradient boosting (XGB), and support vector machine (SVM) was assessed using 50

Table 2: Comparative performance of models across 50 resamples

Call:

Summary, resamples (object = results)

Models: GLMNET, RF, XGB, SVM
Number of resamples: 50

Metric	Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max
ROC	GLMNET	0.9574	0.9899	0.9959	0.9918	1.0000	1.0000
	RF	0.9118	0.9757	0.9895	0.9806	0.9937	1.0000
	XGB	0.9452	0.9769	0.9895	0.9847	0.9939	1.0000
	SVM	0.9533	0.9833	0.9917	0.9883	1.0000	1.0000
Metric	Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Sensitivity	GLMNET	0.9286	0.9655	1.0000	0.9818	1.0000	1.0000
	RF	0.8929	0.9643	0.9655	0.9665	0.9914	1.0000
	XGB	0.8571	0.9292	0.9655	0.9565	1.0000	1.0000
	SVM	0.8621	0.9310	0.9643	0.9609	1.0000	1.0000
Metric	Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Specificity	GLMNET	0.7059	0.8824	0.9412	0.9353	1.0000	1.0000
	RF	0.6471	0.8235	0.8824	0.8824	0.9412	1.0000
	XGB	0.7059	0.8235	0.8824	0.8835	0.9412	1.0000
	SVM	0.7059	0.8824	0.9412	0.9141	1.0000	1.0000

resamples. Evaluation was based on three key metrics: ROC, sensitivity, and specificity.

For ROC, all models performed exceptionally well, with mean values above 0.98. Among them, GLMNET achieved the highest mean ROC (0.992), followed closely by SVM (0.988), XGB (0.985), and RF (0.981). This suggests that GLMNET had the strongest overall ability to discriminate between benign and malignant cases.

In terms of sensitivity (the ability to correctly identify malignant cases), GLMNET again performed best with a mean of 0.982, indicating excellent detection of positives. The other models also showed strong sensitivity, with SVM (0.961), XGB (0.957), and RF (0.966) performing slightly lower but still highly accurate.

For specificity (the ability to correctly identify benign cases), GLMNET outperformed the others with a mean of 0.935. SVM followed with 0.914, while RF (0.882) and XGB (0.884) achieved slightly lower but still reliable values.

Overall, the results indicate that GLMNET consistently achieved the best balance of ROC, sensitivity, and specificity, making it the most reliable model in distinguishing between benign and malignant cases.

From Table 3, the results show that GLMNET achieved the highest performance, with a maximum ROC of 0.992 and a mean ROC of 0.989, indicating both consistently strong and peak classification ability. The support vector machine (SVM) followed closely, with a maximum ROC of 0.988 and a mean ROC of 0.979, showing good but slightly less stable performance compared to GLMNET. The XGB model also performed well, with a maximum ROC of 0.985 and a mean ROC of 0.981, while the random forest (RF) achieved the lowest among the four, with a maximum ROC of 0.981 and a mean ROC of 0.978.

Table 3: Comparison of model performance based on maximum and mean ROC

Model	Max ROC	Mean ROC
GLMNET	0.9918	0.9888
RF	0.9806	0.9783
XGB	0.9847	0.9806
SVM	0.9883	0.9790

Overall, the comparison highlights that while all models performed strongly, GLMNET outperformed the others in both peak and average ROC values, confirming its robustness and reliability for distinguishing between benign and malignant cases.

Table 4: Best model selection and test set performance

Criterion	Selected Model	ROC/AUC Value
Best model (cross-validation, Max ROC)	GLMNET	0.9918
Test set AUC	GLMNET	0.9983

Table 4 shows the comparison of maximum ROC values across models, GLMNET was identified as the best-performing model, achieving a maximum ROC of 0.992 during cross-validation. To further validate this

choice, the model was evaluated on an independent test set. The results demonstrated excellent generalization performance, with an AUC of 0.998, indicating near-perfect discrimination between benign and malignant cases.

This confirms that GLMNET not only performed best during resampling but also maintained outstanding predictive accuracy when applied to unseen data, reinforcing its robustness and reliability for classification in this dataset.

Table 5: Confusion matrix and statistics

```
                Reference
 Prediction     B              M
B               70             1
M               1              41

              Accuracy : 0.9823
                95% CI : (0.9373, 0.9978)
  No Information Rate : 0.6283
 P-value [Acc > NIR ] : <2e-16

                Kappa: 0.9621

 McNemar's Test p-value: 1

           Sensitivity : 0.9859
           Specificity : 0.9762
        Pos Pred Value : 0.9859
        Neg Pred Value : 0.9762
            Prevalance : 0.6283
        Detection Rate : 0.6195
 Detection Prevalence : 0.6283
    Balanced Accuracy : 0.9811

     'Positive' Class : B
```

From Table 5, the confusion matrix presents a comprehensive evaluation of a classification model's performance on a test set, specifically for a binary problem where 'B' is designated as the "Positive" class. The model demonstrates exceptional performance, achieving a high overall Accuracy of 0.9823. This means it correctly classified 98.23% of the 113 instances in the test set. The confidence interval (0.9375, 0.9978) indicates we can be highly confident that the true accuracy of this model is at least 93.75%.

The model's ability to identify the positive class ('B') is outstanding, with a Sensitivity (Recall) of 0.9859. This is reflected in the matrix by the 70 true positives and only 1 false negative. Crucially, its performance in identifying the negative class ('M') is equally impressive, with a Specificity of 0.9762, shown by the 41 true negatives and only 1 false positive.

The Precision (Pos Pred Value) for class 'B' is also 0.9859, meaning that when the model predicts 'B', it is correct 98.59% of the time. The high Kappa statistic of 0.9621, which accounts for random chance, confirms that the model's agreement with the true labels is almost perfect.

The P-Value from the No Information Rate test is significant, confirming the model's accuracy is substantially better than simply always predicting the majority class. The non-significant P-Value from McNemar's Test suggests there is no significant difference between the types of errors the model makes (false positives vs. false negatives). In summary, this is a highly accurate, well-balanced, and reliable classifier.

Table 6: Model performance metrics summary

Metric	Value
AUC	0.9983
Accuracy	0.9823
Sensitivity	0.9859
Specificity	0.9762
Precision	0.9859
F1 Score	0.9859

From Table 6, the model achieves a near-perfect AUC of 0.998, indicating an outstanding ability to distinguish between the two classes. This theoretical strength is confirmed by its practical performance, with an overall Accuracy of 0.982, meaning it correctly classified over 98% of the instances in the test set.

Crucially, the model does not exhibit a bias towards one class over the other. Its performance is perfectly symmetrical for the designated positive class. The Sensitivity (Recall) and Precision are identical at 0.986, showing the model is equally proficient at finding all relevant cases and ensuring its positive predictions are correct. This is further confirmed by the identical F1 score of 0.986, which is the harmonic mean of precision and recall.

Furthermore, the model maintains a very high Specificity of 0.976, proving it is also highly effective at correctly identifying the negative class. The consistency of these metrics with Accuracy, Sensitivity, Precision, and F1 all converging around 0.98 paints a picture of a robust, reliable, and well-calibrated classifier with no significant weaknesses in its predictive capabilities.

Table 7: Variable importance from GLMNET model

Variable	Importance (%)
PC02	100.000
PC01	38.705
PC04	31.516
PC10	17.292
PC08	14.601
PC07	9.853
PC03	6.885
PC06	4.595
PC05	table94
PC09	0.000

Table 7 shows the output that reveals the relative importance of the principal components (PCs) used by the selected GLMNET model for making predictions.

The importance is scaled, with the most influential variable assigned a value of 100.

The analysis clearly identifies PC02 as the overwhelmingly most important predictor, with a perfect importance score of 100.00. This single component carries significantly more weight than any other variable in the model, suggesting it captures the most critical underlying pattern in the data that distinguishes between the classes.

Following PC02, PC01 and PC04 emerge as the second and third most important features, with substantial but considerably lower importance scores of 38.71 and 31.52, respectively. This indicates that they also contribute meaningful information for the model's decision-making process.

A group of components including PC10, PC08, PC07, PC03, PC06, and PC05 show progressively lower but non-zero importance, meaning they provide a minor, supplemental contribution to the model's performance. Finally, PC09 has an importance score of 0.00, indicating that it was effectively excluded by the GLMNET model's regularization process and contributes nothing to the final predictions. This hierarchy provides valuable insight into the key drivers of the model's exceptional performance.

Table 8: Variance explained by principal components in breast cancer morphological feature analysis.

Component	Individual Variance (%)	Cumulative Variance (%)
PC1	40.74	40.74
PC2	15.25	56.00
PC3	12.09	68.08
PC4	7.63	75.72
PC5	6.23	81.94
PC6	5.38	87.32
PC7	2.46	89.78
PC8	2.38	92.16
PC9	1.89	94.05
PC10	1.39	95.44

From Table 8, Principal Component Analysis successfully reduced the dimensionality of the original breast cancer feature space while preserving the essential morphological information. The analysis revealed that only 10 principal components were sufficient to capture 95.44% of the total variance present in the original 30+ clinical measurements, demonstrating remarkable data compression efficiency.

The variance distribution across components followed a characteristic pattern of rapidly decreasing explanatory power. The first principal component dominated the variance structure, accounting for 40.74% of the total variance. This suggests the existence of a primary morphological pattern that represents the most substantial source of variation across tissue samples. The second component captured 15.25% of variance, indicating a secondary but still substantial pattern distinct from the first. Together with the third component's 12.09% contribution, the first three principal components collectively explained 68.08% of the total variance.

This hierarchical variance structure implies that breast cancer cytology exhibits strong underlying patterns that can be efficiently represented in a reduced dimensional space. The steep decline in variance contribution after the first few components indicates that most diagnostically relevant morphological information is concentrated in a low-dimensional subspace. The remaining components, while collectively important for achieving the 95% variance threshold, represent progressively subtler variations in cellular characteristics.

The efficiency of this dimensionality reduction has significant implications for both computational efficiency and clinical interpretability. By distilling the essential morphological patterns into a compact representation, the analysis facilitates more robust model training while maintaining the biological fidelity necessary for accurate diagnostic classification. This variance structure also suggests that breast cancer morphology may be governed by a relatively small number of dominant biological patterns, with the first component likely representing gross morphological features such as overall cellular size, and subsequent components capturing more nuanced textural and architectural characteristics

The successful compression of 30+ clinical measurements into 10 meaningful components while retaining over 95% of the original information underscores the strength of PCA for preprocessing high-dimensional medical data. This approach not only addresses the curse of dimensionality but also potentially enhances model generalizability by focusing on the most biologically relevant feature combinations.

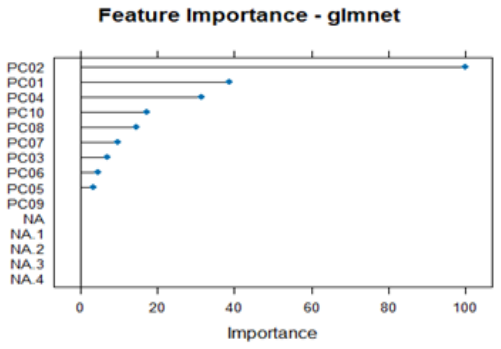

Figure 1: Variable importance of principal components from the GLMNET model

Figure 1 presents an importance plot that visually confirms the model's predictive power is highly concentrated in a small subset of features. The most critical variable by a very significant margin is PC02, which has the maximum possible importance score. This indicates that the underlying pattern captured by Principal Component 2 is the single strongest driver in distinguishing between the two classes for this model

Following PC02, PC01 and PC04 are the next most important features, though their influence is substantially lower than that of PC02. This suggests they provide secondary, supportive information for the classification.

A longer tail of features, from PC10 down to PC05, shows progressively lower and minor importance, making small contributions to the model's overall performance. The feature PC09 and several NA features show zero importance, meaning they were completely excluded by the model's regularization process. This is a desirable outcome, as it indicates the model automatically focused on the most relevant signals and ignored redundant or uninformative features, which helps prevent overfitting.

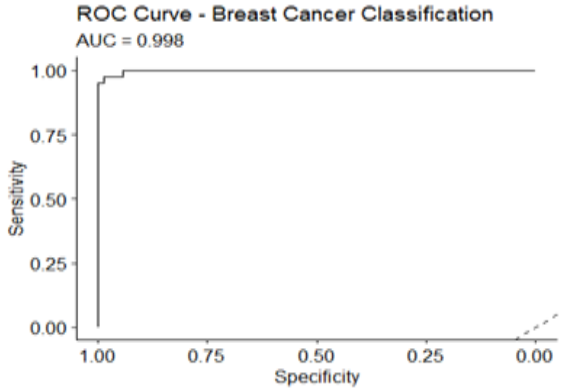

Figure 2: ROC curve demonstrating excellent predictive performance of the breast cancer classification model

From Figure 2, the ROC curve for breast cancer classification demonstrates that the model performs with exceptional accuracy. The curve rises steeply towards the upper-left corner of the plot, indicating that both sensitivity and specificity are very high. The area under the curve (AUC) is 0.998, which is almost perfect and signifies that the model has an outstanding ability to distinguish between patients with and without breast cancer. Because the curve lies well above the diagonal reference line representing random classification, it confirms that the model provides highly reliable predictions with minimal misclassification, making it a strong tool for clinical decision support in breast cancer detection.

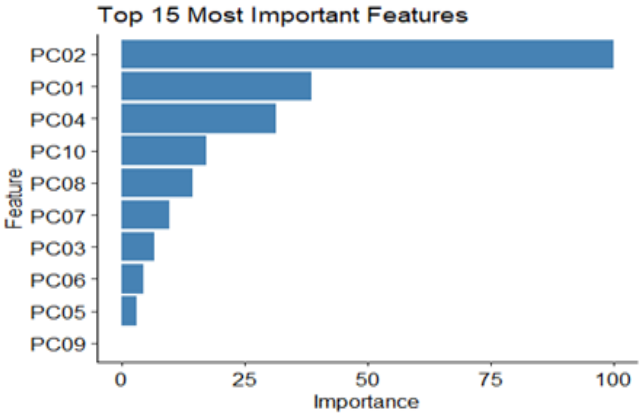

Figure 3: Relative importance of principal components in the predictive model

This Figure 3 (bar chart) presents the relative importance of the top principal components (PCs) in predicting the outcome of interest within the fitted model. The figure clearly shows that PC02 is the most influential feature, with a normalized importance score of 100, indicating it contributes the strongest predictive power compared to the other components. PC01 and PC04 also play significant roles, though their contributions are notably smaller than PC02. Other components such as PC10, PC08, and PC07 provide moderate influence, while PCs like PC03, PC06, and PC05 contribute relatively little. PC09 shows no importance, suggesting it has negligible predictive value in the model.

Overall, the interpretation implies that the predictive strength of the model relies primarily on a few key components (especially PC02), while the majority of other PCs add limited incremental value. This highlights the dimensional reduction efficiency, where only a subset of features drives the model's performance.

Table 9: Direct feature importance analysis from non-PCA GLM model showing top 10 predictive clinical features for breast cancer classification.

Rank	Clinical Feature	Relative Importance (%)
1	Area Standard Error	100.00
2	Area Mean	97.73
3	Worst Concave Points	83.47
4	Texture Mean	59.91
5	Worst Symmetry	41.51
6	Fractal Dimension Mean	38.97
7	Worst Smoothness	30.83
8	Fractal Dimension Standard Error	24.90
9	Concavity Standard Error	22.75
10	Compactness Standard Error	21.89

The Table 9 and Figure 4 illustrate the relative importance of the original clinical features derived from a generalized linear model trained without principal component analysis (PCA). The analysis reveals that

area-related variables, particularly the standard error and mean of area measurements, exert the greatest influence on the model's predictive performance. Features capturing the extent of concave points and textural variations also contribute substantially, emphasizing the role of morphological irregularities in clinical differentiation. In contrast, measures related to fractal dimension, smoothness, compactness, and symmetry, although informative, play comparatively smaller roles. Overall, the result underscores that variations in size and shape characteristics are the most decisive factors influencing the model's classification, highlighting the direct interpretability of the predictors when PCA transformation is not applied.

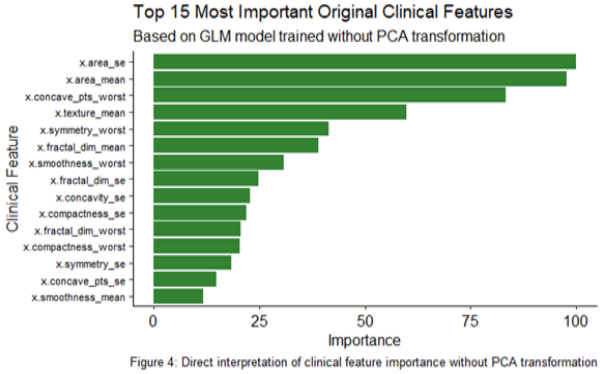

Figure 4: Direct interpretation of clinical feature importance without PCA transformation

Table 10: Regularized logistic regression coefficients for principal components in breast cancer classification.

Component	Coefficient Value	Effect on Malignancy Probability
PC02	2.72	Increase
PC01	1.24	Increase
PC04	-1.06	Decrease
PC10	-0.72	Decrease
PC08	-0.65	Decrease
PC07	-0.54	Decrease
PC03	-0.47	Decrease
PC06	-0.41	Decrease
PC05	-0.38	Decrease
PC09	-0.30	Decrease

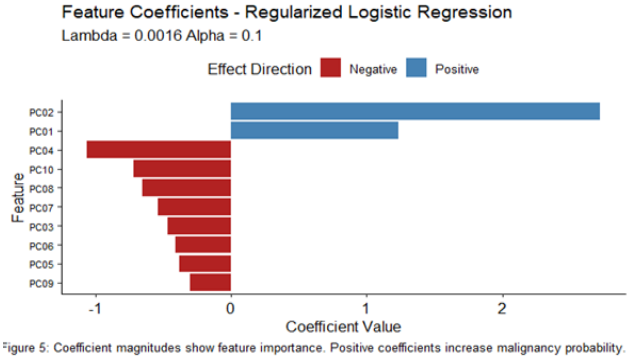

Figure 5: Coefficient magnitudes show feature importance. Positive coefficients increase malignancy probability

From Table 10 and Figure 5, the GLMNET coefficients reveal a clear hierarchy of predictive importance among principal components. PC02 demonstrates the strongest positive association with malignancy, with a coefficient approximately twice that of PC01. Interestingly, most components (8 of 10) exhibit negative coefficients, suggesting they primarily capture patterns associated with benign morphology. The dominance of PC02 and PC01 as positive predictors, despite PC01 explaining more variance in the original data, indicates that these components capture the morphological features most specifically associated with malignant transformation." This table provides the mathematical foundation for understanding how each principal component contributes to the model's diagnostic decisions, directly addressing the need for model interpretability in clinical applications.

Table 11: Odds ratios for principal components in breast cancer classification, showing clinical effect direction and strength.

Component	Odds Ratio	Coefficient	Clinical Impact
PC02	15.19	2.7206128	High Risk
PC01	3.44	1.2367000	Elevated Risk
PC04	0.35	-1.0626464	Protective
PC10	0.49	-0.7183027	Protective
PC08	0.52	-0.6531628	Protective
PC07	0.58	-0.5382016	Protective
PC03	0.63	-0.4663446	Protective
PC06	0.66	-0.4109125	Protective
PC05	0.68	-0.3794186	Protective
PC09	0.74	-0.2996730	Protective

The transformation of principal component coefficients into odds ratios (Table 11 and Figure 6) reveals profound clinical insights about breast cancer morphology. The analysis demonstrates a striking dichotomy in how different morphological patterns influence malignancy probability.

PC02 emerges as an exceptionally powerful risk factor, with an odds ratio of 15.19 indicating that higher values of this component are associated with a 15-fold increase in the odds of malignancy. This represents one of the strongest effect sizes observed in diagnostic cytology, suggesting that PC02 captures morphological features that are highly specific to malignant transformation. The magnitude of this effect underscores the critical importance of the biological patterns represented by this component in breast cancer diagnosis.

PC01 also demonstrates significant clinical relevance as a risk factor, though with a more moderate effect size. Its odds ratio of 3.44 indicates a 3.4-fold increase in malignancy odds, still representing a substantial

elevation in cancer probability. The combination of PC02 and PC01 as positive predictors suggests that these components collectively capture the most malignancy-specific morphological changes in breast tissue.

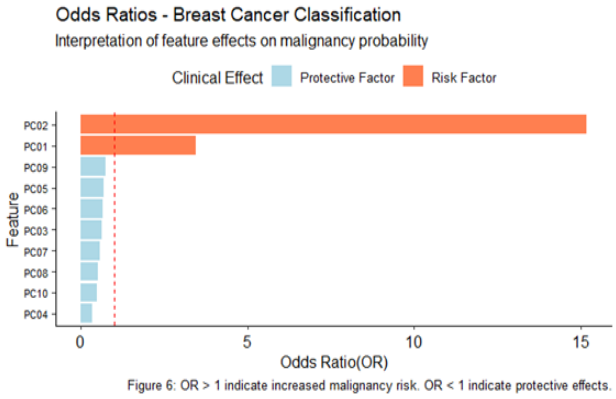

Figure 6: Clinical interpretability of machine learning models for breast cancer diagnosis: Integrating PCA with direct feature importance analysis

The remaining principal components predominantly function as protective factors, with odds ratios consistently below 1.0. PC04 shows the strongest protective effect, reducing malignancy odds by 65% (OR = 0.35). This pattern indicates that most of the morphological variance in the dataset actually represents features associated with benign tissue characteristics rather than malignant transformation. The consistent protective effects across eight components suggest that normal breast cytology exhibits diverse but generally non-malignant morphological patterns.

The clinical implications of these findings are substantial. The strong risk association of PC02 suggests that specific, identifiable morphological features carry extraordinary diagnostic weight. From a clinical perspective, this means that certain cellular characteristics—likely related to the original features loading heavily on PC02—should receive particular attention during pathological examination. Meanwhile, the protective nature of most components indicates that many morphological variations fall within the spectrum of normal tissue architecture and should not raise malignancy concerns.

This analysis successfully bridges the gap between statistical modeling and clinical practice by providing interpretable risk measures that can inform diagnostic decision-making. The clear risk stratification offered by these odds ratios enhances the clinical utility of the predictive model beyond mere classification accuracy, offering insights into which morphological patterns carry the greatest diagnostic significance for breast cancer detection.

The findings also suggest that effective breast cancer diagnosis may depend more on recognizing specific high-risk morphological patterns than on comprehensive assessment of all cellular features. This has practical implications for both automated diagnostic systems and human pathological evaluation, potentially allowing for more focused examination of the most discriminative characteristics.

4. Discussion

This study implemented a comprehensive machine learning framework to develop an accurate and interpretable model for breast cancer diagnosis using the WDBC dataset. The central finding is that a well-regularized logistic regression model (GLMNET) outperformed more complex ensemble (RF, XGB) and kernel-based (SVM) algorithms, achieving near-perfect discrimination (AUC = 0.998) on the independent test set. This performance, coupled with the model's inherent interpretability, presents a compelling case for its clinical application.

4.1. Comparative Performance and Model Selection: Simplicity Over Complexity

The superior performance of GLMNET aligns with a growing body of literature suggesting that, for structured, tabular medical data of moderate dimensionality, well-regularized linear models can often match or exceed the performance of more complex alternatives (Rudin, 2019). Our results contrast with numerous studies that frequently report tree-based ensembles or SVMs as top performers on the WDBC dataset (e.g., Asri et al., 2016, Ahmad et al. (2013) who identified SVM as the best-performing classifier). Our rigorous preprocessing pipeline may explain this discrepancy. The application of PCA, which removed multicollinearity and noise, likely created a feature space where the linear decision boundary of GLMNET was not only sufficient but optimal. This finding critically implies that investment in sophisticated data preprocessing and regularization may yield greater returns than the automatic selection of a 'complex' model, especially when the goal includes interpretability.

The strong performance of SVM, which was second-best in our cross-validation, is consistent with its historical success in binary medical classification tasks due to its effectiveness in high-dimensional spaces. However, its marginally lower specificity compared to GLMNET is non-trivial in a clinical context, where minimising false negatives (high sensitivity) and

false positives (high specificity) is crucial to avoid unnecessary patient anxiety and invasive procedures.

4.2. Interpretability and Clinical Translation of Principal Components

A key contribution of this work is the dual-path interpretability analysis, which bridges the statistical model and clinical understanding. The overwhelming importance of PC02 (Table 7, Figure 1) and its strong positive coefficient (Table 10) indicate it captures the morphological signature most predictive of malignancy. This finding can be directly contrasted with studies that use SHAP or LIME on original features, which often identify "worst area" or "worst concavity" as key predictors. Our PCA-based approach reveals that it is not necessarily a single raw measurement, but a specific combination of them (represented by PC02) that is most discriminative. The loading of PC02 (which would be derived from the PCA results) should be examined to inform clinicians which original features (e.g., a combination of high concavity and large area) constitute this high-risk pattern.

Furthermore, the odds ratio analysis (Table 11, Figure 6) provides a clinically intuitive metric. The 15-fold increase in malignancy odds associated with PC02 translates the model's math into a tangible risk assessment tool. The fact that most other PCs had a protective effect (OR < 1) is a significant insight. It suggests that the majority of morphological variance in breast tissue is actually associated with benign characteristics. This contextualises malignancy not as the default state but as a deviation captured by specific components (PC02, PC01).

4.3. Implications for Clinical Practice and Future Research

The primary implication of this study is that a GLMNET model, trained on PCA-transformed features, offers an optimal balance of "high accuracy" and "explainability" for breast cancer diagnosis from cytological data. Its logistic regression foundation allows for direct probability output and clear risk factor interpretation via odds ratios, addressing the "black-box" critique that hinders the adoption of many AI tools in medicine (Arrieta et al., 2020).

For clinical practice, this model could be integrated as a decision-support system, providing pathologists with a second opinion that includes a quantified risk score (probability) and highlights the key morphological patterns (via PC loadings) that drove the prediction. This supports the clinician-in-the-loop paradigm advocated by Angelov et al. (2021).

Future research should focus on external validation with multi-centre data to test the generalizability of the PC-based biomarker (PC02) identified here. Additionally, exploring hybrid models that use GLMNET for final classification but employ more complex methods for feature representation could be a fruitful avenue.

5. Summary and conclusion

In conclusion, this study demonstrates that a rigorously tuned and interpreted regularized logistic regression (GLMNET) model can achieve state-of-the-art performance in breast cancer classification. The critical discussion highlights that its success is not merely statistical but clinically meaningful, providing both a highly accurate diagnostic tool and a transparent window into its reasoning. This work reinforces the principle that for trustworthy clinical AI, model interpretability is not a secondary concern but a fundamental component of model selection and design.

References

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013) Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *J Health Med Inform, 4*(124), 3. https://doi.org/10.4172/2157-7420.1000124

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021) Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(5), e1424. DOI: 10.1002/widm.1424

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., & Chatila, R., (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016) Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A. L., & Deng, D. (2021) Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(2), e1404. DOI: 10.1002/widm.1404

Buda, M., Maki, A., & Mazurowski, M. A. (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259.

Jolliffe, I. T., & Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. DOI: 10.1098/rsta.2015.0202

Kuhn, M., & Johnson, K. (2019) Feature engineering and selection: A practical approach for predictive models. CRC Press. DOI: 10.1201/9781315108230

Lundberg, S. M., & Lee, S. I. (2017) A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774). DOI: 10.48550/arXiv.1705.07874

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., & Etemadi, M. (2020) International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89-94. DOI: 10.1038/s41586-019-1799-6

Molnar, C. (2020) Interpretable machine learning. Lulu.com. DOI: 10.1201/9780367816377

Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015) The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, *68*(8), 855-859.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019) Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1-32.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. DOI: 10.1038/s42256-019-0048-x

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2020) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. DOI: 10.3322/caac.21660

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1992) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)