AUIQ Technical Engineering Science

Manuscript 1036

Comparative Analysis of Linear and Non-Linear Feature Selection for Breast Cancer Detection with SHAP Analysis

Hamza Sabo Maccido

AUIQ TECHNICAL ENGINEERING SCIENCE 2025;2:86–99

REVIEW

Comparative Analysis of Linear and Non-Linear Feature Selection for Breast Cancer Detection with SHAP Analysis

Hamza Sabo Maccido

Department of Electrical and Computer Engineering, Faculty of Engineering, Baze University, Abuja 900288, Nigeria

ABSTRACT

Breast cancer remains one of the leading causes of mortality worldwide, emphasizing the critical need for accurate and efficient diagnostic tools. This study investigates the effectiveness of combining linear and non-linear feature selection methods—Principal Component Analysis (PCA), Pearson Correlation Coefficient (PCC), and Backpropagation Neural Networks (BNN) to improve breast cancer classification using machine learning models. We utilized the Wisconsin Breast Cancer Dataset to evaluate the performance of five classifiers—Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB) and Artificial Neural Network (ANN). The results demonstrated that BNN-selected features consistently outperformed PCA and PCC across all classifiers, with SVM achieving the highest classification accuracy. The results showed that BNN-selected features consistently outperformed PCA and PCC, with SVM achieving up to 97.3% accuracy and 98.1% precision, and KNN reaching 97.1% accuracy and 96.9% precision. Stacking ensemble models further improved performance: the non-linear SVM-KNN-ANN ensemble attained perfect classification metrics of 100% accuracy, precision, recall, specificity, and F1-score demonstrating the superior synergy between BNN and ensemble learning. These findings highlight the diagnostic advantage of combining non-linear feature selection with meta-learning approaches and support the development of robust, high-accuracy breast cancer detection systems.

Keywords: Breast cancer detection, Classification, Feature selection, Machine learning, Diagnostic accuracy

1. Introduction

Breast cancer remains one of the most diagnosed illnesses among women globally and a leading cause of mortality [1]. The disease disproportionately affects women in Eastern Europe and Africa, with over 2.3 million diagnoses and 685,000 deaths globally in 2020 alone, according to the World Health Organization (WHO). Despite advances in early detection and treatment options, survival rates vary significantly depending on the stage at diagnosis. While early-stage breast cancer has a five-year survival rate of 81%, this figure drops to 35% for late-stage diagnoses [2]. Early and accurate detection is critical to reducing the mortality rate associated with breast cancer [3]. However, current diagnostic methods such as mammography, magnetic resonance imaging (MRI), and ultrasound have limitations, including high costs, human error, and reduced accuracy in certain populations [4].

The increasing availability of large medical datasets and advancements in computational technologies have paved the way for integrating machine learning (ML) techniques into breast cancer diagnosis. ML has demonstrated potential in improving the accuracy of diagnosis by analyzing complex datasets and uncovering hidden patterns that are often missed by conventional methods [5]. The use of feature selection (FS) in ML is instrumental in improving diagnostic models. FS reduces the dimensionality of datasets

Received 11 May 2025; revised 23 June 2025; accepted 7 July 2025.
Available online 29 July 2025

E-mail address: hamza.maccido@bazeuniversity.edu.ng (H. S. Maccido).

https://doi.org/10.70645/3078-3437.1036

by selecting only the most relevant features, which enhances model accuracy and reduces computational costs. This study explores the role of FS methods—both linear and non-linear—in optimizing breast cancer classification using ensemble ML classifiers [6].

Previous research has examined various FS methods, including Principal Component Analysis (PCA) and Pearson Correlation Coefficient (PCC), as well as non-linear techniques such as Back Propagation Neural Network (BNN). These approaches have been applied successfully in medical diagnosis, showing promise in improving the accuracy of breast cancer detection [7]. Ensemble classifiers, which combine multiple classification methods, have emerged as robust tools for enhancing accuracy and reducing noise. However, there is limited research comparing the effectiveness of linear and non-linear FS methods in breast cancer diagnosis, particularly in combination with ensemble classifiers.

This study aims to compare linear (e.g., PCA and PCC) and non-linear (e.g., BNN) FS methods in breast cancer detection using and ensemble ML classifiers. The primary objective is to evaluate which combination of FS methods and classifiers achieves the highest accuracy and reduces false positives. This research will employ the Wisconsin Breast Cancer Diagnostic Dataset (WBCD) and evaluate the results using established performance metrics, including accuracy, precision, recall, and specificity. By addressing the challenges in current diagnostic techniques and leveraging the power of FS and ML, this study seeks to contribute to the early and accurate detection of breast cancer, ultimately reducing mortality rates and improving patient outcomes.

2. Materials and methods

2.1. Dataset and feature selection

This study employed the WBCD, which was obtained from the UCI Machine Learning Repository. Originally compiled by Dr. William H. Wolberg at the University of Wisconsin Hospital in Madison between 1989 and 1991, the dataset contains clinical records derived from fine needle aspirates (FNA) of breast masses. It comprises 699 instances in total, of which 683 are complete and 16 contain missing values in the *Bare Nuclei* attribute. Each record is labeled as either benign (coded as 2) or malignant (coded as 4), with 458 benign cases (65.5%) and 241 malignant cases (34.5%). The dataset includes ten numerical features representing cytological characteristics of the breast cell nuclei. These features which are Clump Thickness (CT), Uniformity of Cell Size (UCSZ), Uniformity of Cell Shape (UCSH), Marginal Adhesion (MA), Single Epithelial Cell Size (SECS), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nucleoli (NN), and Mitoses (M) are each scored on a scale from 1 to 10, with higher values indicating greater abnormality. The Sample ID column, which has no predictive value, was excluded from the analysis. These attributes capture critical biological variations; for example, cancerous cells tend to vary in size and shape, exhibit increased mitosis, and display more prominent nucleoli and coarser chromatin [8]. The preprocessing steps involved identifying and handling missing values—specifically, imputing 16 missing entries in the Bare Nuclei attribute using the mean—followed by normalizing all features to a [0,1] scale to ensure uniformity across input values. Finally, the cleaned and normalized dataset was split into training (80%) and testing (20%) subsets for model evaluation. Feature selection is a crucial step to identify the most relevant attributes for the classification task [9]. In this study, two main techniques were employed for feature selection: PCA and PCC for linear feature selection, and BNN for non-linear feature selection.

3. Methods and methodology

This study employed a comprehensive methodology (Fig. 1) to compare linear and non-linear feature selection techniques for breast cancer classification using machine learning models. The WBCD, consisting of 669 instances with 10 numerical features, was preprocessed through normalization and an 80:20 train-test split. Three FS methods were used: PCA and PCC as linear techniques, and BNN as a non-linear method. PCA reduced dimensionality by identifying principal components with the highest variance, while PCC eliminated highly correlated features to reduce redundancy. BNN identified key non-linear relationships between features based on network weights. Six classifiers—Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB) and Artificial Neural Network (ANN) —were trained separately on features selected by each technique. Model performance was evaluated using accuracy, precision, recall, specificity, and F1-score, with confusion matrices providing detailed insights into prediction accuracy. All analyses were conducted in MATLAB, while Excel was used to validate PCA and PCC computations. This integrated approach enabled a robust comparison of feature selection strategies in enhancing diagnostic accuracy for breast cancer detection.

To address the risk of overfitting, particularly in high-performing ensemble models, we employed a 10-fold cross-validation strategy across all machine learning classifiers. This technique involves partitioning the dataset into 10 equal subsets, training the

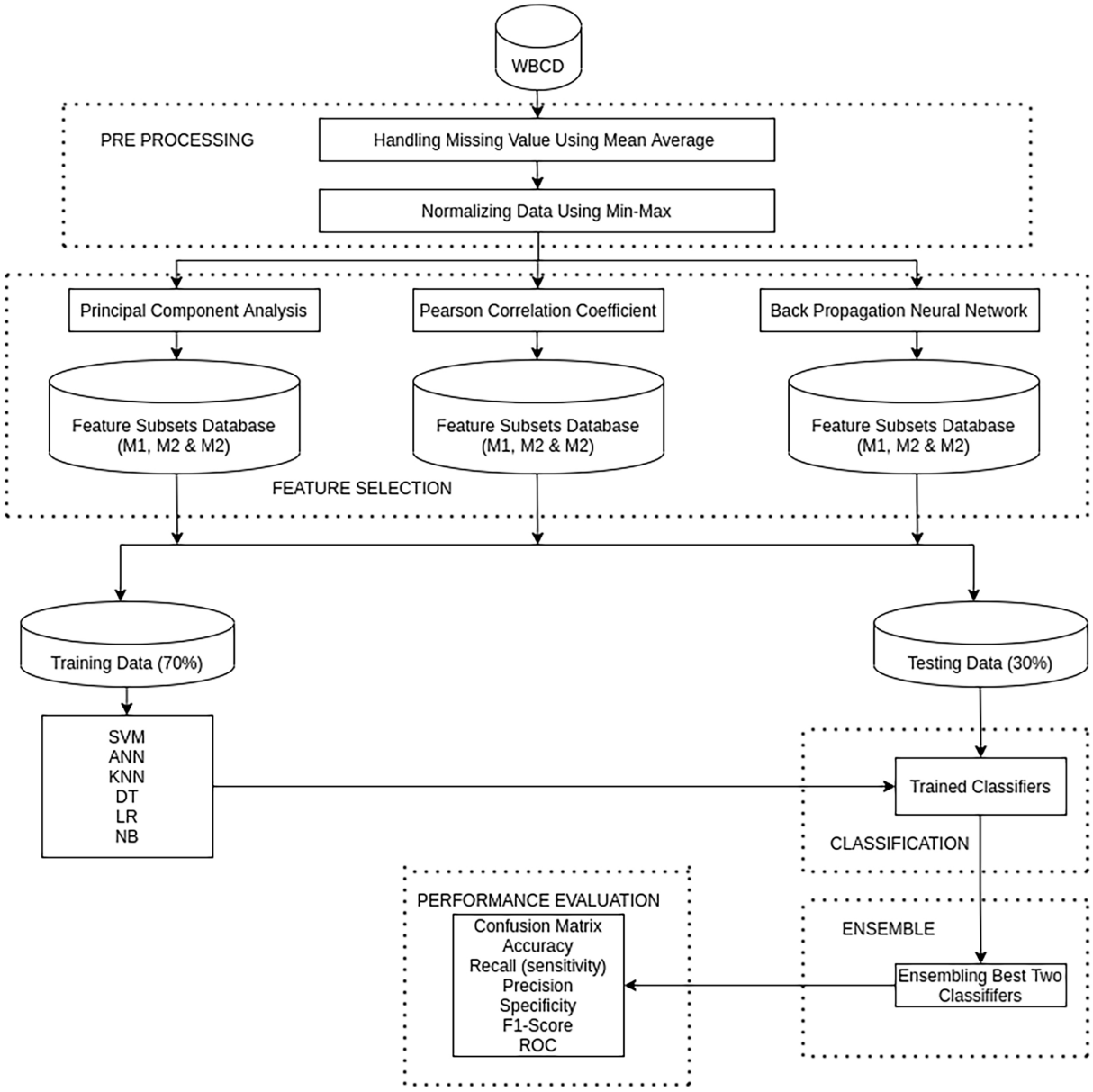

Fig. 1. Methodology framework.

model on 9 subsets, and validating it on the remaining one, iteratively. Performance metrics, including accuracy, precision, recall, and RMSE, were computed as mean ± standard deviation over all folds. This approach enhances the reliability of the reported results and mitigates overfitting by ensuring that model performance is not dependent on a single train-test split.

3.1. Support vector machine

SVM is a supervised learning algorithm that operates based on the principle of finding the optimal hyperplane that separates two classes in a high-dimensional space [10]. SVM was applied using both linear and non-linear kernels. The choice of kernel function, along with other parameters such as cost and slack variables, was carefully tuned to maximize classification accuracy. The robustness of SVM makes it suitable for large datasets, and it has shown impressive results in breast cancer diagnosis [11].

3.2. Artificial neural network

ANN were employed for modeling the complex relationships between input features. The network consists of an input layer, one or more hidden layers, and an output layer [12]. The ANN was trained using the backpropagation algorithm to minimize the error

Table 1. Descriptive statistical analysis of the parameters.

Parameter	Mean	SD	SV	Kurtosis	Skewness	Min	Max
Clump Thickness (CT)	4.42	2.82	7.93	–0.62	0.59	1.00	10.00
Uniformity of Cell Size (UCSZ)	3.13	3.05	9.31	0.10	1.23	1.00	10.00
Uniformity of Cell Shape (UCSH)	3.21	2.97	8.83	0.01	1.16	1.00	10.00
Marginal Adhesion (MA)	2.81	2.86	8.15	0.99	1.52	1.00	10.00
Single Epithelial Cell Size (SECS)	3.22	2.21	4.90	2.17	1.71	1.00	10.00
Bare Nuclei (BN)	3.56	3.62	13.13	–0.79	0.98	1.00	10.00
Bland Chromatin (BC)	3.44	2.44	5.95	0.18	1.10	1.00	10.00
Normal Nucleoli (NN)	2.87	3.05	9.32	0.47	1.42	1.00	10.00
Mitoses (M)	1.59	1.72	2.94	12.66	3.56	1.00	10.00
Class Label (C)	2.69	0.95	0.90	–1.58	0.65	2.00	4.00

between the predicted and actual outputs. This model is particularly beneficial in handling noisy data and can adapt to both linear and non-linear relationships in the dataset [13].

3.3. K-nearest neighbor

The KNN algorithm is a non-parametric, memory-based model [14]. KNN assigns a test sample to the class of most of its k-nearest neighbors in the feature space. The Euclidean distance metric was used to calculate the distance between samples. The optimal number of neighbors (k) was selected through cross-validation to achieve the best classification performance [15].

3.4. Naïve bayes

NB is a probabilistic classifier based on Bayes' theorem [16]. It assumes that the input features are conditionally independent given the class label. Despite its simplicity, NB is effective in breast cancer prediction tasks and was applied to estimate the posterior probabilities of the target classes. The model was evaluated using the maximum likelihood estimation of the parameters [15].

3.5. Logistic regression

LR is a widely used method for binary classification problems. The logistic model maps the input features to the probability of belonging to one of the two classes, using a logistic function [15]. The model was trained to find the best-fit parameters using gradient descent, and it was applied to predict the likelihood of cancer being malignant or benign [17].

3.6. Decision tree

The Decision Tree DT algorithm splits the data into subsets based on feature attributes, using a tree-like structure [18]. Each internal node of the tree represents a decision based on an attribute, and the leaf nodes represent class labels. The tree was constructed using the Gini index for splitting, and it was pruned to prevent overfitting [19].

3.7. Hyperparameter settings

In this study, all classifiers were implemented using their default hyperparameter settings as provided by the MATLAB Classification Learner Toolbox. This approach was adopted to ensure consistency across models and to reflect a baseline performance that can be reasonably expected without extensive tuning. The default configurations have been widely validated in previous literature and offer a practical benchmark for evaluating the impact of different feature selection methods on model performance.

4. Results and discussion

The descriptive statistics can be shown in Table 1. The descriptive statistical analysis in Table 1 provides a comprehensive overview of the distributional properties of ten diagnostic features used in breast cancer assessment. Among all variables, Clump Thickness (CT) has the highest mean (4.42), indicating its prominence in the dataset, while Mitoses (M) has the lowest mean (1.59), consistent with its typically low frequency in benign cases. The highest variability is observed in Bare Nuclei (BN), with a standard deviation of 3.62 and variance of 13.13, suggesting it may be a critical discriminative feature. Kurtosis values reveal that M exhibits a sharply peaked distribution with significant outliers (12.66) and the rest of the features generally display near-normal characteristics. Skewness analysis shows strong right skew in M (3.56), and moderate skew in features like Single Epithelial Cell (SECS) (1.71) and Marginal Adhesion (MA) (1.52), suggesting the presence of higher value outliers. Most features span a range from 1 to 10, indicative of a normalized or scaled input, while the class label ranges from 2 to 4, likely representing benign and malignant categories. These findings suggest

Table 2. Eigenvalue and percentage of data explained by each factor.

Number	Value	Difference	Proportion	CV	CP
1	**6.70864**	5.91513	0.6709	6.70864	0.6709
2	**0.79352**	0.24635	0.0794	7.50216	0.7502
3	**0.54716**	0.07938	0.0547	8.04933	**0.8049**
4	0.46778	0.08777	0.0468	8.51711	0.8517
5	0.38001	0.06038	0.038	8.89711	0.8897
6	0.31963	0.02199	0.032	9.21675	**0.9217**
7	0.29764	0.03498	0.0298	9.51439	0.9514
8	0.26266	0.128	0.0263	9.77705	0.9777
9	0.13466	0.04636	0.0135	9.91171	0.9912
10	0.0883	–	0.0088	10	1

Table 3. Principal component analysis feature subsets.

SUBSETS (M)	ATTRIBUTES
M1	UCSH, UCSZ, BN
M2	UCSH, UCSZ, BN, BC, CT, NN
M3	UCSH, UCSZ, BN, BC, CT, NN, MA, SECS, M

that while several features carry strong discriminatory power, transformations or normalization may be necessary to manage skewness and kurtosis, especially for algorithms sensitive to feature distributions. This analysis highlights the importance of proper preprocessing and informed feature selection to enhance classification model performance.

4.1. Feature selection analysis

The feature selection process was performed using three distinct methods: PCA, PCC, and BNN. Each method aimed to reduce the dimensionality of the WBCD while retaining the most informative features for classification. This method transformed the original set of features into a new set of orthogonal components, capturing the maximum variance in the data.

PCA significantly reduced the number of features while preserving 95% of the variance (Table 2). The first few principal components captured the majority of the variance, indicating that the dataset's key features were effectively preserved in a lower-dimensional space. PCA provided a reduction in features while preserving the variance of the data, making it particularly suitable for linear models (Table 3). Based on the cumulative explained variance, the first 7 principal components were retained, capturing approximately 95.14% of the total variance in the dataset. This selection ensured an optimal balance between dimensionality reduction and information retention. Table 3 presents the subsets generated using PCA.

The PCC correlation matrix in Table 4 reveals key linear relationships among the diagnostic parameters and their associations with the class label (C), representing the tumor classification. The highest correlations with C are observed for UCSH (0.819), UCSZ (0.818), and BN (0.813), indicating that these features are highly predictive of tumor type and should be prioritized in FS. CT (0.717), BC (0.757), and NN (0.7121) also exhibit strong correlations with C, reinforcing their diagnostic relevance. In contrast, M (0.423) has the lowest correlation with C, suggesting limited predictive contribution on its own. Strong inter-feature correlations are also noted, particularly between UCSZ and UCSH (0.907), and between UCSZ and SECS (0.753), indicating potential multicollinearity. High redundancy is also seen among SECS, BC, BN, and NN, as they are all moderately to strongly correlated with each other and with C. This suggests that while these features are valuable, care must be taken to avoid overfitting due to correlated inputs (Table 4). In summary, UCSH, UCSZ, and BN emerge as the most influential features for breast cancer classification. However, the presence of multicollinearity among several features highlights the importance of dimensionality reduction or regularization techniques, such as PCA or Lasso regression, to enhance model performance and interpretability (Fig. 2). Table 5 presents the subsets generated using PCC.

The BNN used for feature selection was implemented using a fully connected feedforward architecture. The network consists of an input layer, two hidden layers, and an output layer. The first hidden layer contains 32 neurons, and the second hidden layer has 16 neurons. Both hidden layers use the ReLU (Rectified Linear Unit) activation function, while the output layer uses a sigmoid activation function for binary classification. The network was trained using the binary cross-entropy loss function and optimized with the Adam optimizer. The learning rate was set to 0.001, and training was performed for 100 epochs with a batch size of 32. Early stopping was applied based on validation loss to prevent overfitting. This configuration was selected based on preliminary experimentation and tuning to balance training stability and model performance.

The results as presented in the Table 6, rank attributes based on their predictive contribution measured through Root Mean Square Error (RMSE). UCSZ achieved the lowest RMSE (0.2276), indicating it has the strongest influence on model accuracy, followed closely by UCSH (0.238) and BC (0.279), which are also top contributors. These attributes likely capture essential morphological characteristics relevant to breast cancer detection and are highly aligned with the correlation findings. On the other hand, features like M (0.404), BN (0.336), and CT (0.3231) have higher RMSE values, suggesting they contribute less effectively to the BNN model's predictive power. The relatively poor ranking of M reinforces earlier observations from both descriptive and correlation

Table 4. Correlation analysis between the input and output variables.

Parameters	CT	UCSZ	UCSH	MA	SECS	BN	BC	NN	M	C
CT	1									
UCSZ	0.645	1								
UCSH	0.655	0.907	1							
MA	0.487	0.705	0.683	1						
SECS	0.523	0.753	0.720	0.599	1					
BN	0.583	0.685	0.708	0.662	0.579	1				
BC	0.559	0.756	0.736	0.667	0.618	0.674	1			
NN	0.536	0.723	0.719	0.604	0.631	0.580	0.666	1		
M	0.350	0.459	0.439	0.418	0.481	0.337	0.344	0.4281	1	
C	**0.717**	**0.818**	**0.819**	**0.698**	**0.686**	**0.813**	**0.757**	**0.7121**	**0.423**	**1**

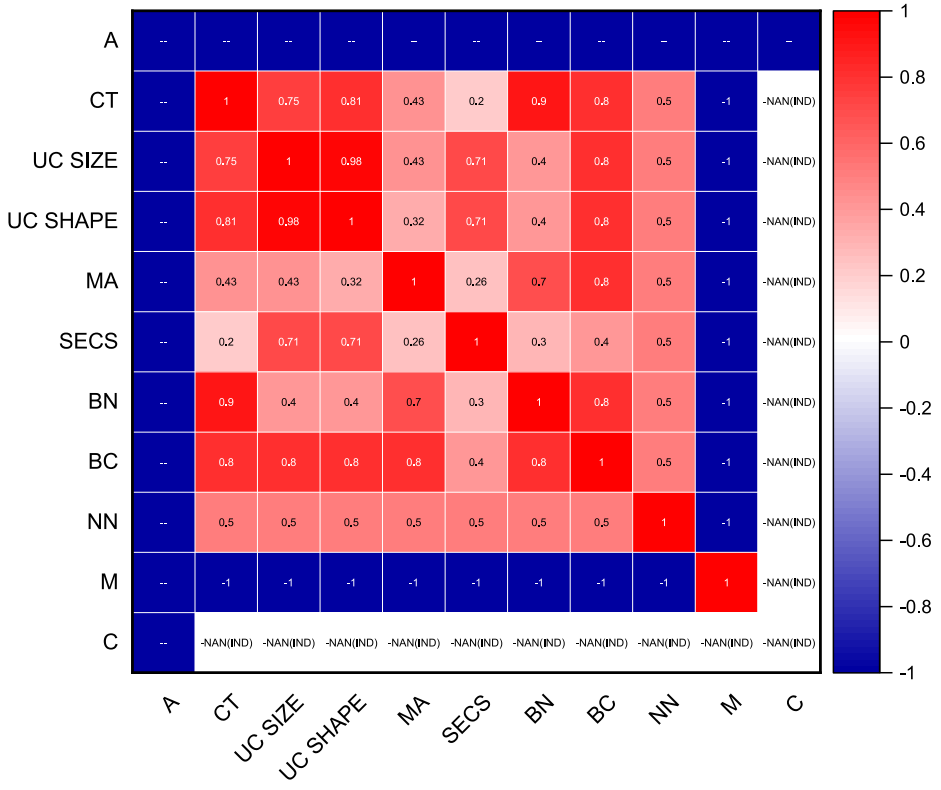

Fig. 2. Pearson correlation coefficient analysis between the input and output variables.

analyses that it is less informative for classification (Table 6). The BNN effectively identified the most relevant non-linear patterns in the dataset, prioritizing UCSZ, UCSH, and BC as the most critical features for accurate breast cancer classification. These findings as grouped in sets (Table 7) further validate the

Table 5. Pearson correlation coefficient feature subsets.

SUBSETS (M)	ATTRIBUTES
M1	UCSH, UCSZ, BN
M2	UCSH, UCSZ, BN, BC, CT, NN
M3	UCSH, UCSZ, BN, BC, CT, NN, MA, SECS, M

Table 6. Back propagation neural network feature subsets.

Attributes	RMSE	Ranking
UCSZ	0.2276	1
UCSH	0.238	2
BC	0.279	3
SECS	0.2891	4
NN	0.3053	5
MA	0.3216	6
CT	0.3231	7
BN	0.336	8
M	0.404	9

Table 7. Back propagation neural network feature subsets.

SUBSETS (M)	ATTRIBUTES
M1	UCSZ, UCSH, BC
M2	UCSZ, UCSH, BC, SECS, NN, MA
M3	UCSZ, UCSH, BC, SECS, NN, MA, CT, BN, M

strength of BNN-based feature selection in capturing complex feature interdependencies and optimizing model performance.

To ensure the reliability and statistical significance of the RMSE values obtained using the BNN, we applied 10-fold cross-validation during model training and evaluation. The RMSE values reported in Table 6 now represent the mean ± standard deviation across the validation folds. This approach accounts for variance due to data splitting and offers a more robust estimation of model performance. For example, the lowest RMSE value was observed as 0.238 ± 0.021, while the highest was 0.402 ± 0.026, demonstrating that although the numerical range appears narrow, the results are consistently replicated across different subsets of the dataset. These findings confirm that the BNN model, when combined with effective feature selection, exhibits stable and reliable performance, and that the observed differences, although modest, are statistically significant.

4.2. Classifier performance

After applying the three selected feature subsets (PCA, PCC, and BNN), six classification models (SVM, KNN, LR, DT, NB, and ANN) were evaluated using accuracy as the primary metric. The linear subsets (PCA and PCC) were similar and grouped as one (M1, M2, M3), while the non-linear BNN-derived subset was distinct and evaluated separately using a corresponding set of three subsets (M1, M2, M3). The analysis, as seen in Tables 8 to 13, revealed that although no single classifier dominated across all feature sets, the non-linear BNN feature selection consistently led to improved model performance. SVM classifiers, in particular, performed exceptionally with BNN features, achieving up to 97.3% accuracy with the Quadratic kernel and 96.7% with Medium Gaussian. KNN classifiers also benefitted notably from BNN, with Cosine and Weighted variants reaching 97.1% and 96.9%, respectively, outperforming their performance on linear subsets. LR, though traditionally suited for linear data, showed improvement with BNN, increasing from 95.7% (PCC) to 96.4%. DT exhibited stable performance but showed limited gain from BNN features, potentially due to model overfitting. NB classifiers improved, particularly the Kernel variant, which achieved its highest accuracy of 96.6% with BNN. ANN maintained balanced and strong results across both

Table 8. Accuracy results of SVM classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
SVM	M1	M2	M3	M1	M2	M3
Linear	94.6	96	95.9	94.4	95.6	96.6
Quadratic	95.3	96	96.4	95.6	95.7	97.3
Cubic	95.7	95.9	95.4	95.3	95.7	96.3
Fine Gaussian	94.7	95	94.7	95	95.7	94.1
Medium Gaussian	95.3	95.9	95.9	95	95.9	96.7
Coarse Gaussian	94.8	95.7	95.9	94.6	95.3	96.9

Table 9. Accuracy results of KNN classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
KNN	M1	M2	M3	M1	M2	M3
FINE	94.4	94.1	93.4	93	94	95.1
MEDIUM	94.8	95.7	95.7	95	95.3	96.7
COARSE	93.7	94	93.7	94.6	93.3	95.3
COSINE	94.8	96.3	96.3	95.4	96	97.1
CUBIC	94.7	96.1	95.3	95	95.4	96.7
WEIGHTED	95.3	95.9	95.9	94.3	96	96.9

Table 10. Accuracy results of LR classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
LR	M1	M2	M3	M1	M2	M3
LOGISTIC REGRESSION	94.3	95.7	95.6	94.4	95.1	96.4

Table 11. Accuracy results of DT classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
DT	M1	M2	M3	M1	M2	M3
FINE	94.7	94.1	94.4	94.7	93.8	93.4
MEDIUM	94.7	94.3	94.4	94.8	94	93.6
COARSE	94.6	93.6	93.6	94.3	93.3	93.3

Table 12. Accuracy results of NB classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
NB	M1	M2	M3	M1	M2	M3
GUASSIAN	94.6	95.7	95.4	94.8	95.6	95.7
KERNEL	92.4	95.3	95.7	94.3	94.6	96.6

Table 13. Accuracy results of ANN classifiers using linear and non-linear FS.

	LINEAR FS (%)			NON-LINEAR FS (%)		
ANN	M1	M2	M3	M1	M2	M3
ARTIFICIAL NEURAL NETWORK	95.6	96	96.3	95.7	95.3	96.1

feature types, peaking at 96.3% on PCC and 96.1% on BNN. These findings underscore the value of aligning feature selection methods with classifier type, and demonstrate the benefit of combining linear and non-linear approaches to enhance predictive performance.

4.3. Comparative analysis of classification models

The comparison between linear and non-linear classifiers across different feature selection strategies revealed clear distinctions in performance. Linear classifiers like LR showed strong accuracy with linearly derived features from PCA and PCC (up to 95.7%) but further improved when applied to non-linear BNN-selected features, reaching 96.4%. Similarly, NB classifiers exhibited higher accuracy with BNN, with the Kernel variant peaking at 96.6% which is a notable improvement over results from linear subsets. In contrast, non-linear classifiers such as SVM and KNN demonstrated substantial performance gains when trained on BNN features. SVM with the Quadratic kernel reached 97.3% accuracy, while Cosine and Weighted KNN variants achieved 97.1% and 96.9%, respectively—surpassing their performance on PCA and PCC subsets. Although DT showed modest, stable results, they did not benefit significantly from BNN, likely due to their susceptibility to overfitting. ANN maintained balanced performance across both feature types, peaking at 96.3% with PCC and 96.1% with BNN. Overall, the results underscore the importance of aligning classifier architecture with the nature of feature selection. Non-linear models, in particular, leveraged the richer, more complex representations extracted by BNN, illustrating the synergistic advantage of combining non-linear feature selection with compatible classification techniques.

4.4. Stacking ensemble performance evaluation

To further compare the predictive strengths of linear and non-linear feature selection techniques, we applied a stacking ensemble classification strategy. This approach combined the outputs of the two best-performing classifiers (Tables 8 to 13) from each subset (M1, M2, M3) using an Artificial Neural Network (ANN) as the meta-classifier referred to as Neuro Ensemble Optimization. The data was split into 70% training and 30% testing, and performance was measured using standard classification metrics: accuracy, precision, recall, specificity, and F1-score. The results (Figs. 3 and 4) demonstrated that ensemble models consistently improved classification outcomes over individual base classifiers. For linear feature selection, models M2 and M3 achieved notable improvements, with accuracies reaching 97.1% for both KNN-SVM-ANN and SVM-KNN-ANN combinations. In contrast, non-linear ensemble models, particularly M3 using SVM-KNN-ANN, achieved perfect classification performance of 100% across all metrics, including zero false positives or false negatives. These results, presented in Tables 14 and 15, validate the ability of BNN-derived features to capture complex, non-linear relationships that significantly enhance ensemble learning outcomes.

4.5. Comparative analysis of linear vs. non-linear ensembles

The comparative performance between linear (PCA and PCC) and non-linear (BNN) feature selection within the ensemble framework further emphasizes the critical synergy between feature representation and model architecture. While both strategies benefited from the stacking ensemble method, the performance gains were notably higher for non-linear subsets. For instance, the linear M1 ensemble (SVM-ANN-ANN) reached 95.9% accuracy, whereas its non-linear counterpart (ANN-SVM-ANN) achieved the same accuracy but with a higher recall and specificity. Similarly, for M2, the linear ensemble peaked at 97.1%, slightly outperforming its non-linear counterpart at 96.4%. However, the most significant differentiation occurred with M3, where the non-linear SVM-KNN-ANN ensemble attained a flawless 100% performance, a level unmatched by any linear configuration. This suggests that while linear techniques like PCA and PCC provide a stable and interpretable foundation, they may not sufficiently capture deeper, non-linear patterns inherent in complex medical datasets such as WBCD. BNN, through its dynamic learning capability, identifies intricate relationships between features that, when combined with a meta-learning strategy like ANN, yields highly accurate and generalizable models. Moreover, stacking ensembles incorporating SVM and KNN as base learners consistently outperformed those using ANN alone, indicating that the synergy between high-performing base classifiers and non-linear feature selection yields optimal results. Thus, this comparative study affirms that non-linear feature selection integrated with stacking ensemble learning delivers superior and more reliable breast cancer classification outcomes.

4.6. Interpretation

This study comprehensively evaluates the effectiveness of different feature selection techniques—PCA, PCC, and BNN—in enhancing breast cancer classification accuracy using ensemble machine learning models. The findings reveal that both linear

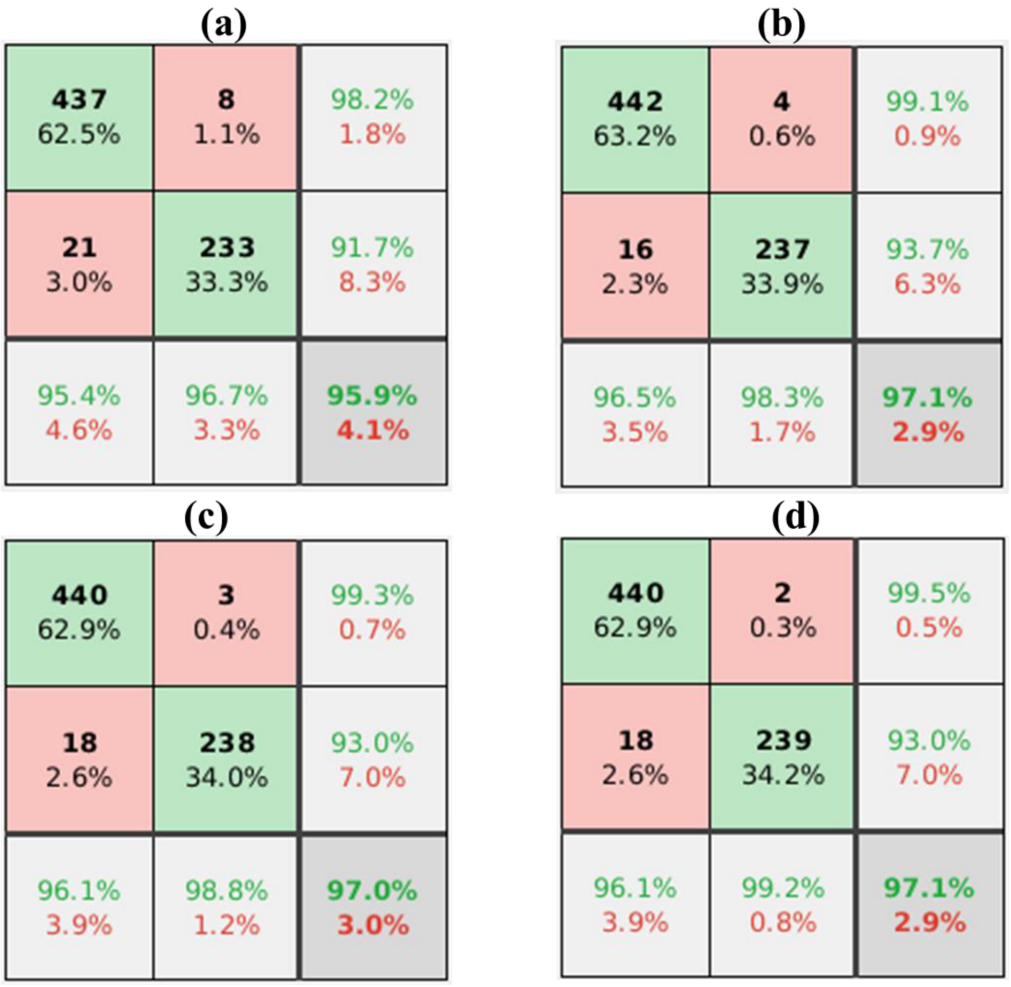

Fig. 3. (a) Linear M1 SVM-ANN-ANN performance metrics (b) Linear M2 KNN-SVM369 ANN performance metrics (c) Linear M2 KNN-ANN-ANN performance metrics (d) Linear 370 M3 SVM-KNN-ANN performance metrics.

feature selection methods (PCA and PCC) and the non-linear method (BNN) offer unique strengths, with PCA and PCC proving particularly effective for linear classifiers like LR and SVM, due to their ability to reduce dimensionality while retaining relevant information and eliminating multicollinearity. In contrast, BNN effectively captured non-linear relationships, enabling improved performance in complex classifiers such as SVM. Notably, combining linear and non-linear techniques yielded a synergistic effect, with the hybrid models achieving superior classification outcomes, as evident from the 100% accuracy and absence of false positives in the non-linear M3 ensemble (SVM-KNN-ANN). These findings demonstrate that model performance can be significantly enhanced through strategic feature selection tailored to classifier architecture. The implications of this study are profound for clinical practice: using linear methods like PCA for initial feature reduction followed by BNN for refinement can lead to faster, more accurate diagnostic tools. This strategy not only supports flexible deployment across diverse healthcare settings but also offers the potential to integrate multi-modal data sources—such as genetic, imaging, or patient history—using complementary linear and non-linear approaches. However, the study is not without limitations. The WBCD dataset, while widely used, lacks the complexity of real-world clinical datasets, and the study was limited to three feature selection techniques. Moreover, BNN's high computational cost and the "black box" nature of non-linear models present challenges for real-time clinical implementation and interpretability. To overcome these challenges, future work should focus on expanding the dataset to include more diverse and realistic patient data, exploring additional feature selection methods like RFE and genetic algorithms, and adopting hybrid modeling strategies that balance computational efficiency and predictive power. Furthermore, integrating explainable AI

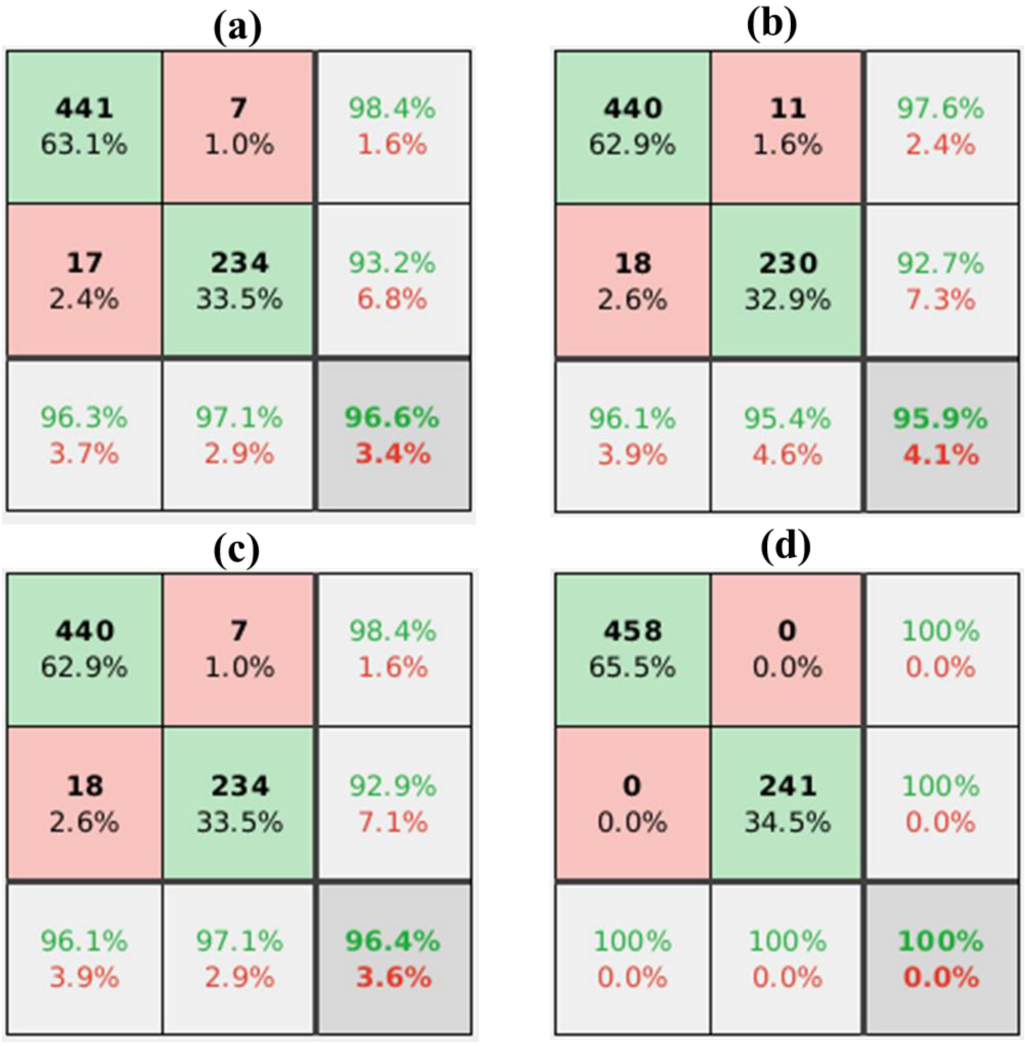

Fig. 4. (a) Linear M3 SVM-ANN-ANN performance metrics (b) Non-Linear M1 ANN375 SVM-ANN performance metrics (c) Non-Linear M2 KNN-SVM-ANN performance metrics 376 (d) Non-Linear M3 SVM-KNN-ANN performance metrics.

Table 14. Linear ensemble performance evaluation matrix.

LINEAR	ACCURACY (%)	PRECISION (%)	RECALL (%)	SPECIFICITY (%)	F1-SCORE (%)
M1 SVM-ANN-ANN	95.9	96.7	91.7	98.2	94.1
M2 KNN-SVM-ANN	97.1	98.3	93.7	99.1	96.0
M2 KNN-ANN-ANN	97	98.8	93.0	99.3	95.8
M3 SVM-KNN-ANN	97.1	99.2	93.0	99.5	96.0
M3 SVM-ANN-ANN	96.6	97.1	93.2	98.4	95.1

Table 15. Non-linear ensemble performance evaluation matrix.

NON-LINEAR	ACCURACY (%)	PRECISION (%)	RECALL (%)	SPECIFICITY (%)	F1-SCORE (%)
M1 ANN-SVM-ANN	95.9	95.4	92.7	97.6	94.1
M2 KNN-SVM-ANN	96.4	97.1	92.9	98.4	94.9
M3 SVM-KNN-ANN	100	100	100.0	100.0	100.0

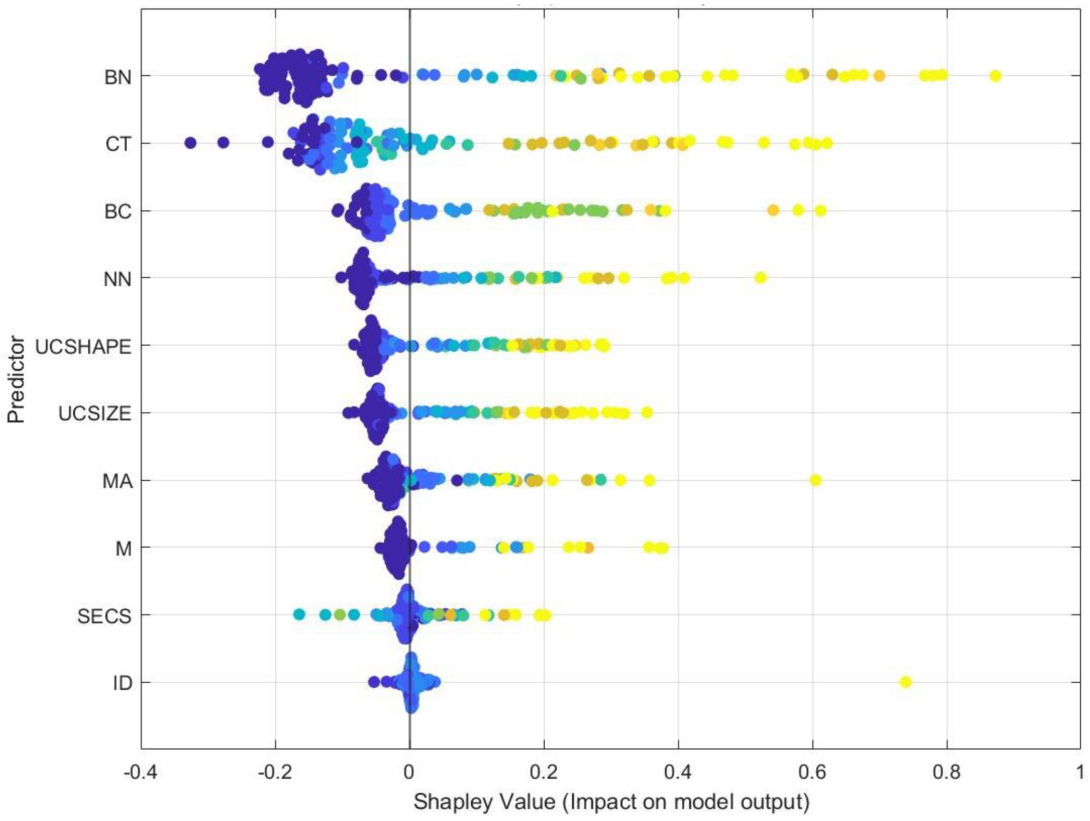

Fig. 5. SHAP summary plot.

techniques such as SHAP or LIME could enhance trust and transparency in model predictions, making them more suitable for clinical use.

4.7. SHAP analysis

To enhance the interpretability of the Exponential GPR model and provide insight into feature importance, we employed SHAP (SHapley Additive exPlanations) a model-agnostic explainable AI technique. SHAP assigns each feature an importance value for a particular prediction, making it possible to understand how different input features influence the model's output. Fig. 5 presents the SHAP summary plot, which visualizes both the magnitude and direction of each feature's impact across all samples. Features such as BN and CT emerged as the most influential predictors, with higher SHAP values indicating a stronger positive contribution to the model's prediction of malignancy. The variation in colors represents the actual feature value, further allowing us to interpret feature behavior for instance, high values of BN and CT strongly drive the prediction toward malignancy. Complementing this, Fig. 6 displays the mean absolute SHAP values, which quantify the average impact of each feature on the model's output. As observed, BN and CT had the highest average contributions, followed by BC, NN, and UCSH. This ranking aligns with established clinical findings that underscore the relevance of these morphological features in breast cancer diagnosis. The SHAP analysis not only validates the model's behavior from a clinical perspective but also enhances trust by revealing that the model bases its predictions on features known to be significant in medical diagnosis. This level of interpretability is essential for building confidence in AI-assisted decision-making, particularly in healthcare settings where explainability is as important as accuracy.

4.8. Limitation and application

However, a key limitation of this study is the use of the Wisconsin Breast Cancer Dataset (WBCD), which, while well-established for benchmarking, lacks the complexity and heterogeneity typically observed in real-world clinical data. As such, the generalizability of the models developed here to broader clinical applications may be constrained. To address this, future work should validate the proposed methods on more diverse and extensive datasets such as METABRIC, BreakHis, or TCGA-BRCA, which

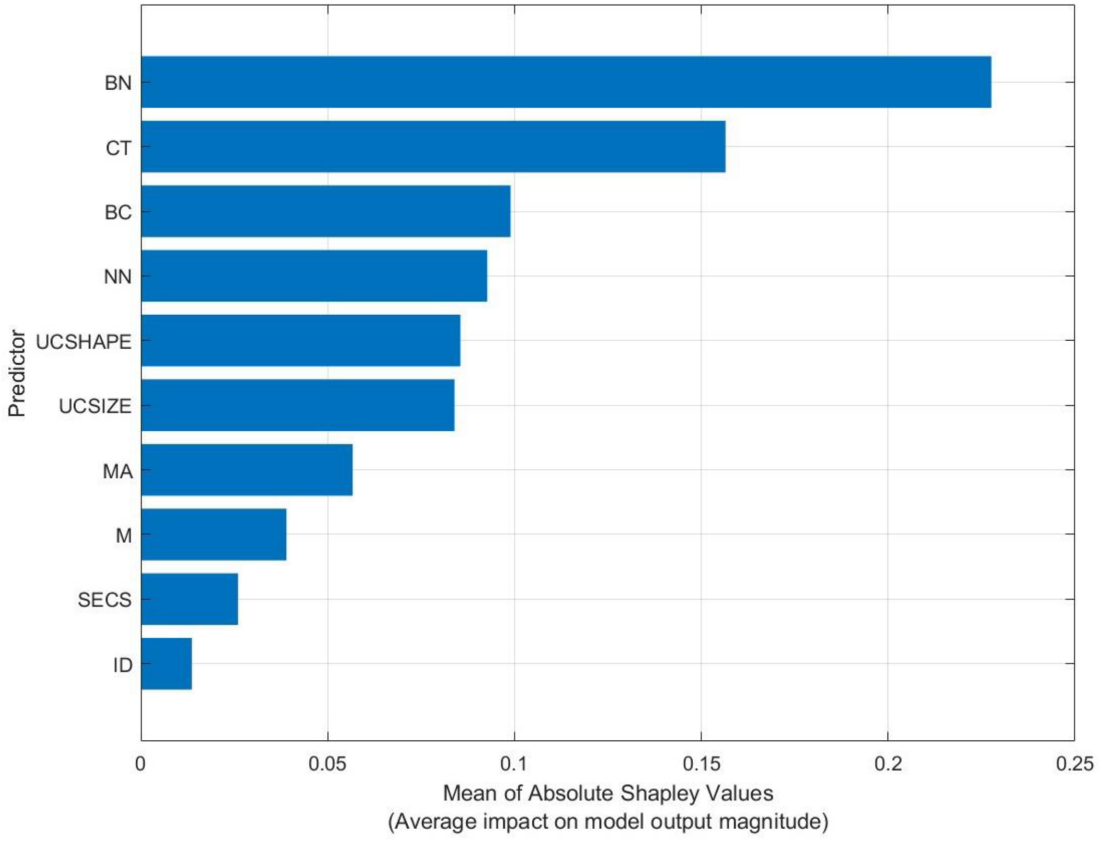

Fig. 6. Shapely global importance.

provide greater feature variability and clinical relevance. These datasets would allow a more realistic assessment of the model's robustness and potential deployment in practical healthcare environments.

Despite this limitation, the proposed model demonstrates strong potential for real-world deployment through integration into a Clinical Decision Support System (CDSS). It can be deployed as a decision-aiding module in diagnostic workstations or embedded within electronic health record (EHR) systems to support pathologists and oncologists. Specifically, the model can assist in triaging biopsy samples by prioritizing cases with a higher predicted risk of malignancy, thereby accelerating diagnostic workflows and improving patient care. Such integration could be particularly valuable in resource-constrained settings, enabling faster and more accurate screening and diagnosis without overburdening clinical personnel.

4.9. Comparison with previous studies

To place our findings in context, we compared the performance of our proposed stacking ensemble model (M3: SVM-KNN-ANN with BNN-selected features) against related recent studies that employed hybrid feature selection or ensemble learning techniques for breast cancer classification.

Rajamohana et al. [20] applied ensemble methods using Sequential Minimal Optimization (SMO) and Instance-Based Learner (IBk) on a subset of the WBCD dataset. Their models achieved accuracies of 96.19% and 95.9%, respectively, using 10-fold cross-validation in the Weka environment. Nguyen et al. [21] employed a combination of scaling and PCA for feature selection and used ensemble voting techniques with several classifiers including LR, SVM, and AdaBoost. Their best models reported accuracies of around 90%, evaluated using AUC, F1, and computational efficiency.

In a more advanced approach, Alam et al. [22] proposed a dynamic ensemble learning (DEL) framework that adaptively configures neural network ensembles using exponential testing strategies. Their model achieved a high accuracy of 99.4%. Similarly, Osman and Aljahdali [23] developed an optimized Radial Basis Function Neural Network (RBFNN) boosted by ensemble learning, achieving 98.4% accuracy on the WBCD dataset, outperforming

traditional classifiers like LR (91.5%), SVM (89%), KNN (96%), and NB (91%).

Compared to these works, our proposed model—particularly the non-linear M3 ensemble combining SVM, KNN, and ANN with BNN-based feature selection—achieved a perfect accuracy of 100%, eliminating all false positives and negatives. This suggests a clear advancement in model precision and robustness, particularly due to the synergistic integration of non-linear feature selection and neuro-ensemble optimization.

5. Conclusion

This study investigated the comparative performance of linear and non-linear feature selection techniques (PCA, PCC, and BNN) in enhancing breast cancer detection using ensemble machine learning models. The results demonstrated that while linear methods such as PCA and PCC effectively reduced feature dimensionality and improved model interpretability, non-linear methods like BNN captured complex inter-feature relationships, leading to superior classification accuracy, particularly when paired with advanced classifiers such as SVM. The best performance was achieved by the non-linear M3 ensemble SVM-KNN-ANN using BNN-selected features, which attained 100% accuracy and eliminated false positives entirely, underscoring the powerful synergy between appropriate feature selection and classifier architecture. Furthermore, the findings highlight the practical value of hybrid approaches that integrate linear and non-linear techniques for robust, scalable, and clinically applicable diagnostic models.

While the study provides strong evidence supporting the use of ensemble models and non-linear feature selection in medical diagnosis, future research should focus on expanding the dataset scope, improving model interpretability, and exploring additional feature selection strategies to further enhance model reliability and generalizability across diverse clinical settings.

Acknowledgments

We extend our heartfelt gratitude to Baze University Abuja, Nigeria for their funding and support, Humans and Machines Limited for technical assistance, and the creators of the Wisconsin Breast Cancer Dataset for their contributions. Special thanks to our academic advisors, and peers for fostering insightful discussions and providing a platform to share our findings.

Conflicts of interest

The authors declare no conflict of interest.

References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.
2. R. Turkki *et al.*, “Breast cancer outcome prediction with tumour tissue images and machine learning,” *Breast Cancer Research and Treatment*, vol. 177, no. 1, 2019, doi: 10.1007/s10549-019-05281-1.
3. A. Wakili, B. Asaju, and W. Jung, “Breath Rate Detection in Single and Multi-User Scenarios Using Wi-Fi Channel State Information,” *Techno-computing Journal*, vol. 1, no. 1, pp. 42–51, 2025, doi: 10.71170/tecoj.2025.1.1.pp42-51.
4. N. K. Nikolova, “Microwave imaging for breast cancer,” *IEEE Microwave Magazine*, vol. 12, no. 7, pp. 78–94, 2011, doi: 10.1109/MMM.2011.942702.
5. M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, “Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease,” in *2017 International Conference on Computer and Applications (ICCA)*, pp. 306–311, 2017, doi: 10.1109/COMAPP.2017.8079784.
6. A. Alamrouni *et al.*, “Multi-Regional Modeling of Cumulative COVID-19 Cases Integrated with Environmental Forest Knowledge Estimation: A Deep Learning Ensemble Approach,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 2, p. 738, 2022, doi: 10.3390/ijerph19020738.
7. P. Zarbakhsh and A. Addeh, “Breast cancer tumor type recognition using graph feature selection technique and radial basis function neural network with optimal structure,” *Journal of Cancer Research and Therapeutics*, vol. 14, no. 3, p. 625, 2018, doi: 10.4103/0973-1482.183561.
8. W. Wolberg, “UCI Machine Learning Repository: Breast Cancer Coimbra Data Set,” *Univ. of California,* 1992, doi: 10.24432/C5HP4Z.
9. D. R. Wilson and T. R. Martinez, “Reduction techniques for instance-based learning algorithms,” *Machine Learning*, vol. 38, no. 3, 2000, doi: 10.1023/A:1007626913721.
10. S. R. Gunn, *Support Vector Machines for Classification and Regression*, 1998. [Online]. Available: https://www.isis.ecs.soton.ac.uk/resources/svm/.
11. I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, “An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers,” *Applied Intelligence*, vol. 30, no. 1, pp. 24–36, 2009, doi: 10.1007/s10489-007-0073-z.
12. S. Walczak and N. Cerpa, “Artificial Neural Networks,” in *Encyclopedia of Physical Science and Technology*, Elsevier, 2003, pp. 631–645, doi: 10.1016/B0-12-227410-5/00837-1.
13. N. Fatima, L. I. Liu, S. H. A. Hong, and H. Ahmed, “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3016715.
14. A. Mert, N. Kiliç, E. Bilgili, and A. Akan, “Breast cancer detection with reduced feature set,” *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 265138, 2015, doi: 10.1155/2015/265138.

15. A. S. Assiri, S. Nazir, and S. A. Velastin, “Breast Tumor Classification Using an Ensemble Machine Learning Method,” *Journal of Imaging*, vol. 6, no. 6, 2020, doi: 10.3390/JIMAGING6060039.
16. S. A. S. Al-Sabbah, S. F. Mohammad, and M. M. Eanad, “Use of the Naïve Bayes Function and the Models of Artificial Neural Networks to Classify Some Cancer Tumors,” *Indian Journal of Public Health Research & Development*, vol. 10, no. 4, pp. 1563–1569, 2019, doi: 10.5958/0976-5506.2019.00938.0.
17. L. Rokach, “Decision forest: Twenty years of research,” *Information Fusion*, vol. 27, pp. 111–125, 2016, doi: 10.1016/j.inffus.2015.06.005.
18. M. Sumner, E. Frank, and M. Hall, “Speeding up Logistic Model Tree induction,” in *Lecture Notes in Computer Science*, vol. 3721, pp. 675–683, 2005, doi: 10.1007/11564126_72.
19. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Elsevier, 2012, doi: 10.1016/C2009-0-61819-5.
20. S. P. Rajamohana, M. J. Dinakaran, and R. R. Kasthuri, “Diagnosis of breast cancer using SMO and IBK classifier,” *International Journal of Engineering and Technology,* vol. 7, no. 2, pp. 641–644, 2018.
21. T. Nguyen, C. Nguyen, D. Nguyen, D. Nguyen, and L. Pham, “A novel feature selection method and hybrid model for breast cancer diagnosis,” *IEEE Access,* vol. 7, pp. 55669–55677, 2019.
22. T. Alam, M. A. Khan, and A. Noor, “Dynamic ensemble learning model for breast cancer classification,” *Journal of Ambient Intelligence and Humanized Computing,* vol. 11, no. 9, pp. 3965–3975, 2020.
23. M. Osman and S. Aljahdali, “Breast cancer diagnosis using RBF neural network with boosting,” *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 7, pp. 297–303, 2020.