Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 3, Month 2026, 000–000

Principal Component Analysis for Dimensionality Reduction of the Breast Cancer Dataset

Tran Anh Vu[1], Le Dieu Huyen[1], Dong Thi Diu[1], Hoang Quang Huy[1], Pham Thi Viet Huong[2*]
[1] *Hanoi University of Science and Technology, Hanoi, Vietnam*
[2] *International School, Vietnam National University, Hanoi, Vietnam*
[*] *Corresponding author email: huongpv@vnu.edu.vn*

Abstract

Breast cancer is a prevalent global health concern among women. This study systematically investigates the application of Principal Component Analysis for dimensionality reduction on the Wisconsin Breast Cancer Dataset. We evaluate the impact of varying PCA dimensions on the performance of several machine learning and deep learning models, including Support Vector Machine, K-Nearest Neighbors, Random Forest, Multilayer Perceptron, Fully Connected Neural Network, and Dropout models. Our findings demonstrate that PCA can enhance model performance and accuracy when dimensions are reduced from high to moderate levels. Conversely, overly aggressive dimensionality reduction leads to a significant degradation in performance (k<5). ML models exhibited optimal performance at different values of k. In which SVM achieves its best results at k=25, KNN shows stability in the range of k= 10-15, and Random Forest performs effectively at k=5. For DL models, accuracy remained stable for k ≥ 10 and saturated between k = 12-18, consistently achieving over 97% accuracy. These results underscore the critical importance of selecting an appropriate number of PCA dimensions to balance accuracy and computational efficiency, thereby improving the efficacy of breast cancer diagnostic support systems.

Keywords: Breast cancer, component analysis, feature extraction, feature selection, principal dimensionality reduction.

1. Introduction

Breast cancer continues to be the most common cancer among women globally, with approximately 2.3 million new cases diagnosed in 2020 [1]. In Vietnam, as of 2022, it is the leading cancer among women, accounting for 28.9% of all new female cancer diagnoses [2]. While men can also be affected, the incidence rate is significantly lower [1]. The disease originates from the uncontrolled, abnormal growth of cells within breast tissue. Key risk factors include age, family history, genetic mutations, and lifestyle choices. Early detection through regular screening is paramount for successful treatment [3]. The World Health Organization statistics for 2025 highlight that 30% of cancer cases are curable if detected early, emphasizing the importance of preventive methods [4].

The development of computer-aided diagnostic - CAD systems has advanced significantly, offering valuable assistance to physicians in breast cancer diagnosis, enhancing both accuracy and consistency of results [5]. Machine learning algorithms, such as Support Vector Machines and Artificial Neural Networks, have been effectively applied in breast cancer classification. Research [6] indicates that both SVM and ANN models achieve high diagnostic accuracy (98%–99%), with hybrid SVM-ANN models potentially further improving performance over individual models. Beyond traditional ML, deep learning techniques, including Convolutional Neural Networks, are also employed for breast cancer detection, with their accuracy largely dependent on model architecture and data quality. However, in high-dimensional biomedical data analysis, an increasing number of features can lead to the "curse of dimensionality," negatively impacting the performance and accuracy of ML and DL models [7]. Consequently, dimensionality reduction, particularly through techniques like Principal Component Analysis, is recognized as a crucial preprocessing step to enhance model efficiency, stability, and interpretability [8].

Dimensionality Reduction (DR) is a fundamental approach in machine learning aimed at reducing the number of feature variables in a dataset while preserving its most salient characteristics [9]. This is particularly beneficial for large and complex datasets where high dimensionality can impede analysis and visualization. DR methods are broadly categorized into feature extraction [10] and feature selection [11]. Feature extraction methods create new features by combining original ones, with PCA and Linear Discriminant Analysis (LDA) being prime examples [10]. Feature selection methods, conversely, identify and select a subset of the most relevant original features without alteration, encompassing filter, wrapper, and embedded techniques [11]. The application of DR not only reduces computational load but also improves ML model

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.192.ssad.2026.36.3.xx
Received: Jan 16, 2026; Revised: Apr 3, 2026;
Accepted: May 18, 2026; Online: Jun 22. 2026

performance, mitigates overfitting risks, and enhances model interpretability [11].

Among linear dimensionality reduction techniques, PCA stands out as one of the simplest and most widely used methods for reducing a large set of variables into a smaller set while retaining most of the original dataset's information [12]. PCA has been successfully applied across various domains, notably in biomedical applications and cancer diagnosis. For instance, PCA improved the performance of classification algorithms like SVM and KNN on the Iris dataset [13]. In medical imaging, PCA has been utilized to analyze large fMRI datasets, reducing memory requirements and accelerating processing while preserving critical information [14]. Essentially, PCA effectively reduces data dimensionality, helps eliminate noise, concentrates on the most important features, accelerates computational processes during ML model training, and concurrently enhances model accuracy.

However, most existing studies employ PCA using a fixed number of principal components or based solely on cumulative variance thresholds, without systematically analyzing the sensitivity of different learning models to varying dimensional settings.

In this paper, we present a systematic analysis of the impact of PCA dimensions (k = 1–30) on the performance of popular ML and DL models using the Wisconsin Breast Cancer Dataset.

Unlike prior works that typically select a predefined dimensional threshold, this study investigates the entire dimensional spectrum (k = 1–30), thereby enabling a complete empirical characterization of model behavior under progressive information compression.

The main contributions of this research are:

• A full-spectrum dimensional sensitivity analysis (k = 1–30), allowing identification of performance transition regions, degradation zones (k < 5), and stability plateaus.

• A unified experimental framework in which both machine learning and deep learning models are evaluated under identical preprocessing, PCA transformation, and data-splitting protocols, ensuring methodological consistency and fair comparison across heterogeneous learning paradigms.

• Identification of a performance saturation threshold (k ≈ 12–18), where additional principal components yield negligible accuracy gains despite increased dimensional complexity.

• A comparative sensitivity analysis revealing distinct responses to dimensionality reduction among models — with SVM benefiting from moderate compression, KNN demonstrating mid-range stability, Random Forest showing robustness at low dimensions, and deep learning models exhibiting plateau behavior once sufficient variance is retained.

• An extended evaluation using clinically relevant performance metrics, including Precision, Recall, F1-score, and Area Under the ROC Curve (AUC), in addition to overall accuracy, thereby providing a more comprehensive assessment of diagnostic reliability in biomedical classification tasks.

While PCA has been extensively applied to the Wisconsin Breast Cancer Dataset to enhance classification performance [15], typical studies often rely on a fixed number of PCA dimensions or cumulative variance thresholds [16]. Furthermore, these studies predominantly focus on individual models, rarely evaluating the simultaneous impact of PCA dimensions on the accuracy, stability, and training time of both ML and DL models. Therefore, a systematic analysis is essential to determine the optimal PCA dimensions for each model type within a consistent experimental approach.

Therefore, this work moves beyond the conventional "apply-PCA-and-report-accuracy" paradigm and proposes a systematic empirical framework for dimensional sensitivity analysis across heterogeneous learning paradigms, offering deeper insight into how dimensional compression influences different model architectures in biomedical diagnostics.

2. Materials and Methods

2.1. Dataset Description

The Wisconsin Breast Cancer Diagnostic dataset, obtained from the UCI Machine Learning Repository, was utilized in this study [17]. This dataset comprises 569 patient samples, each characterized by 30 numerical features extracted from digitized images of fine needle aspirates of breast masses.

The dataset exhibits an imbalance, with 212 malignant samples (approximately 37%) and 357 benign samples (63%), reflecting the inherent complexity often encountered in biomedical diagnostic problems.

Data Structure:

Each sample in the dataset includes:

• **ID**: Patient identification number.

• **Diagnosis Label**: Categorical label (M for malignant, B for benign).

• **30 morphological features**: These features describe various characteristics of cell nuclei, including mean, standard error, and "worst" (largest) values for attributes such as radius, perimeter, area, concavity, and symmetry.

Data Preprocessing

Prior to PCA application, the data underwent the following steps:

• Irrelevant columns, such as patient ID, were removed as they do not contribute to classification. The diagnosis label column was excluded during the PCA process.

• Diagnosis labels were binary encoded, assigning a value of 1 for malignant (M) samples and 0 for benign (B) samples. The feature matrix $X \in \mathbb{R}^{569\times30}$and label vector y were then separated.

• Given PCA's sensitivity to data scaling, all input features were standardized using z-score normalization via scikit-learn's StandardScaler library. Data inspection confirmed the absence of missing values, duplicate entries, or severe outliers, leading to the retention of all samples for the experiment

To avoid data leakage, all preprocessing steps, including standardization and PCA transformation, were performed within each training set during the repeated train-test split process. Specifically, for each split, the StandardScaler and PCA models were fitted exclusively on the training data and subsequently applied to the test data using the learned parameters. This ensures that no information from the test set was used during model training.

2.2. Principal Component Analysis

Principal Component Analysis (PCA) serves as a linear technique for dimensionality reduction, remapping data into a novel coordinate framework. Within this space, components are organized according to their diminishing levels of variance [12]. Unlike arbitrary retention of original data features, PCA preserves the majority of information through a few significant components. Fig. 1 shows the steps for performing PCA.

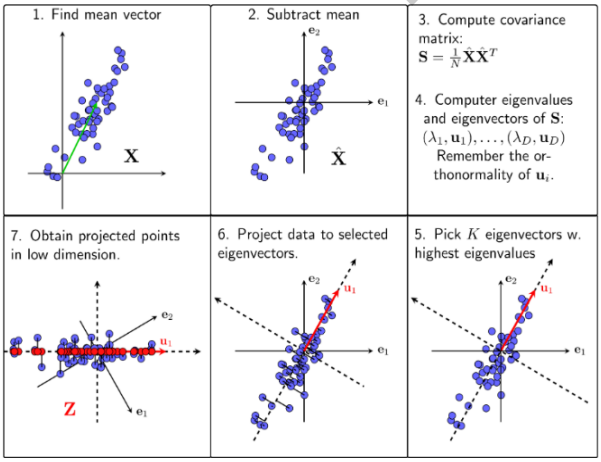

Fig. 1. Steps for performing PCA

Directly retaining K original features may not be optimal due to uneven information contributions, potentially leading to significant information loss. PCA addresses this by constructing new orthogonal axes; the first principal component captures the largest variance, and subsequent components capture progressively less information. This allows for the discarding of less important components while maintaining model performance.

In this study, PCA was applied subsequent to data standardization and prior to model training. The number of dimensions k was varied from 1 to 30 to investigate its impact on the accuracy and training time of both machine learning and deep learning models.

2.3. Model Configuration

To evaluate the impact of dimensionality reduction on classification performance, three classical machine learning algorithms were implemented: Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF). All models were implemented using the scikit-learn library in Python.

To ensure a fair dimensional sensitivity analysis, all hyperparameters were intentionally fixed across all PCA settings, and no additional tuning was performed. This experimental design ensures that observed performance variations are attributed solely to changes in the number of principal components rather than model re-optimization effects.

For each PCA dimension $k \in [1,30]$, models were trained and evaluated using repeated random stratified train-test splits (80% training, 20% testing) over 30 independent repetitions. This repeated evaluation strategy reduces variance due to data partitioning and improves result stability.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm primarily used for classification, which works by finding the optimal hyperplane that maximizes the margin between different data classes. By utilizing kernel functions, SVM can effectively handle non-linear data by projecting it into higher-dimensional spaces where a linear separation becomes possible.

The model configuration was as follows:

• Kernel: RBF

• Regularization parameter C: default value (1.0)

• Gamma: default (“scale”)

• Probability estimation: enabled (probability=True)

• Random state: 42

The RBF kernel was selected to allow nonlinear decision boundaries in the reduced PCA feature space. Probability estimation was enabled to compute the Area Under the ROC Curve (AUC).

All hyperparameters were kept constant across different PCA dimensional settings. No hyperparameter optimization was conducted in order to isolate the effect of dimensionality reduction on classification performance.

K-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm - a non-parametric, distance-centric approach - was employed to perform classification tasks.

The configuration was:

- Number of neighbors k: 5
- Distance metric: Euclidean distance (default in scikit-learn)
- Weighting scheme: uniform weights

The choice of $k = 5$provides a balance between bias and variance and is commonly used as a baseline configuration.

Since KNN relies on distance computations, feature standardization was applied prior to PCA transformation to ensure all features contributed equally to the distance metric.

No hyperparameter tuning was performed. The same configuration was applied across all PCA dimensions to maintain experimental consistency.

Random Forest (RF)

The Random Forest (RF) classifier, an ensemble learning method based on bootstrap aggregation and random feature selection, was implemented to provide a robust nonlinear benchmark.

The model parameters were:

- Number of trees: 100
- Maximum tree depth: 5
- Random state: 42
- Other parameters: default values in scikit-learn

Limiting the maximum depth to 5 controls model complexity and reduces overfitting, particularly in low-dimensional PCA spaces.

The number of trees (100) was chosen to ensure stable ensemble performance while maintaining computational efficiency.

As with other models, hyperparameters were fixed across all PCA dimensional settings. No hyperparameter search or adaptive tuning was conducted.

Deep Learning Architecture

In addition to classical machine learning algorithms, fully connected neural networks were implemented using the TensorFlow/Keras framework to evaluate dimensional sensitivity under PCA transformation.

Three architectures were considered:

MLP:

- Input: k principal components
- One hidden layer (32 neurons, ReLU)
- Output layer (1 neuron, Sigmoid)

FCNN (Fully Connected Neural Network):

- Two hidden layers (64 and 32 neurons, ReLU)
- Output layer (Sigmoid)

Dropout Network:

- Two hidden layers (64 and 32 neurons, ReLU)
- Dropout layer (rate = 0.3) between hidden layers
- Output layer (Sigmoid)

All models were trained using:

- Optimizer: Adam (Kingma & Ba, 2014)
- Learning rate: 0.001
- Loss: Binary cross-entropy
- Batch size: 16
- Epochs: 10
- Classification threshold: 0.5

For each PCA dimension ($k = 1$ to 30), models were evaluated using repeated stratified 80/20 train-test splits (10 repetitions). Performance metrics were averaged across repetitions.

To ensure controlled comparison, all architectural and training hyperparameters were fixed across PCA dimensions. No hyperparameter tuning or adaptive re-optimization was performed, allowing observed performance differences to reflect solely the impact of dimensionality reduction.

2.4. Evaluation Metrics

In addition to classification accuracy, clinically relevant metrics including precision, recall (sensitivity), F1-score, and area under the Receiver Operating Characteristic curve (AUC) were evaluated to provide a comprehensive assessment of model performance in a medical diagnostic context.

Accuracy measures the overall proportion of correctly classified samples. However, in medical diagnosis, relying solely on accuracy may be misleading due to potential class imbalance.

For both machine learning and deep learning models, metrics were computed on the test set for each repetition and averaged across repeated stratified splits to ensure robust performance estimation.

2.5. Experimental Protocol

For each PCA dimension (k=1–30), models were evaluated using repeated stratified 80/20 train-test splits. Machine learning models were repeated 30 times, while deep learning models were repeated 10 times. Performance was reported as mean accuracy ± standard deviation and log10(training + prediction time).

For DL models, evaluations were repeated 10 times, specifically measuring:

- Accuracy at iteration i.

$$acc_i = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ test\ samples}$$

- Average accuracy after n iterations.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} acc_i$$

- Standard deviation of n iterations.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(acc_i - \mu)^2}$$

- The average training and prediction time was calculated as.

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

For ML models, performance was visualized using Accuracy ± standard deviation and log10 (training + prediction time). The same evaluation protocol was applied to DL models.

3. Results

3.1. PCA Variance Analysis

The PCA yielded:

- **Individual Explained Variance Ratio**: The percentage of information attributed to each principal component.
- **Cumulative Explained Variance Ratio**: The total percentage of information retained by summing the principal components.

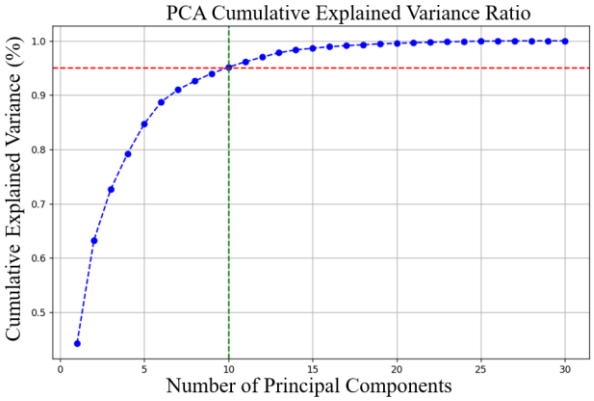

Fig. 2. The cumulative variance plot of PCA

Fig. 2 illustrates the cumulative explained variance obtained from PCA applied to the standardized dataset.

The curve exhibits a characteristic elbow pattern. The first ten principal components retain approximately 90% of the total variance. When the number of components exceeds 15, more than 95% of the variance is preserved. Beyond 25 components, the incremental variance gain becomes marginal, approaching the original 30-dimensional feature space.

These findings suggest that a moderate dimensionality range ($k \approx$ 10–20) is sufficient to preserve most of the dataset's variance while significantly reducing complexity.

3.2. Overall Classification Performance

ML Models Results

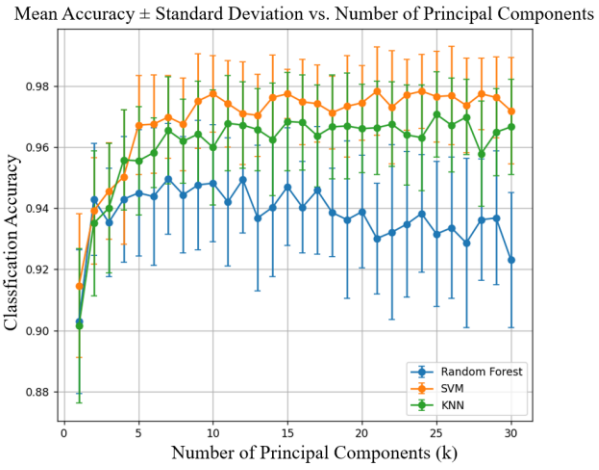

Fig. 3. Average accuracy of ML models with k different data points

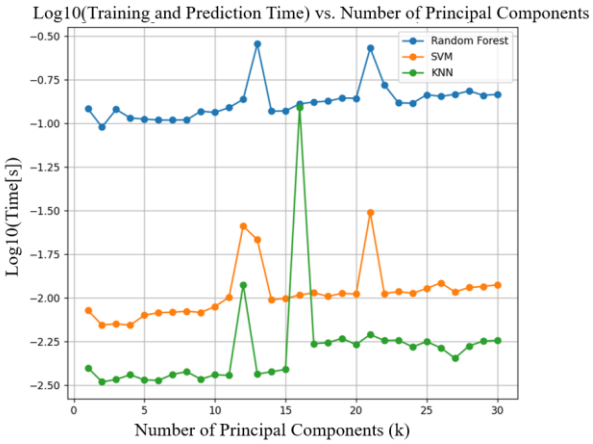

Fig. 4. Training and prediction time of ML models with k different data points

Fig. 3 and Fig. 4 illustrate the changes in accuracy and training time of ML models as a function of the number of PCA dimensions. The results indicate that PCA can improve classification performance when dimensionality is reduced from the original space to a moderate level ($k \approx 10$–25). However, if the reduction is excessively aggressive ($k < 5$), the accuracy of all models significantly declines due to the loss of critical characteristic information.

Table 2. Comparison of Model Accuracy (Acc.) by PCA Dimensions

Model	Acc. at k = 30	Highest Acc.	Optimal PCA Dimension	Remarks
SVM	0.9602	0.9795	25	Benefits strongly from PCA
KNN	0.9507	0.9711	15	Stable, good balance
RF	0.9257	0.9518	5	Robust, less dependent on dimensionality

Accuracy increased when PCA dimensions were reduced from the original 30 features to a moderate range. When $k < 5$, accuracy decreased for all machine learning models.

Compared to $k = 30$, SVM achieved its highest accuracy at $k = 25$ (0.9795). KNN achieved its highest accuracy at $k = 15$ (0.9711). Random Forest reached its highest accuracy at $k = 5$ (0.9518). Table 2 summarizes the comparison of model accuracy across PCA dimensions.

Training and prediction time varied across PCA dimensions. SVM processing time remained relatively stable across different k values. KNN processing time was lower at small k and increased as dimensionality increased. Random Forest processing time showed minor variation across PCA dimensions.

Across models, the range $k = 10$–15 maintained accuracy above 0.96 for SVM and KNN. Random Forest maintained accuracy above 0.94 at small dimensional settings ($k = 5$–10).

DL Models Results

Fig. 5 and Fig. 6 present the accuracy and training time of deep learning models across different PCA dimensions. Classification accuracy increased as the number of principal components increased. At small k values ($k < 5$), accuracy was lower for all deep learning models. Accuracy increased rapidly up to approximately $k = 10$–15 and then plateaued as k continued to increase. When $k \geq 10$–15, accuracy remained stable and close to the maximum values observed across all dimensional settings.

The cumulative explained variance exceeded approximately 95%–98% when k was within the range of 10–15 principal components. Training time varied across PCA dimensions. Reducing dimensionality resulted in limited changes in training time. At certain k values, training time increased compared to higher-dimensional settings. Across models, the range $k = 12$–18 maintained near-maximum accuracy while preserving most of the total variance.

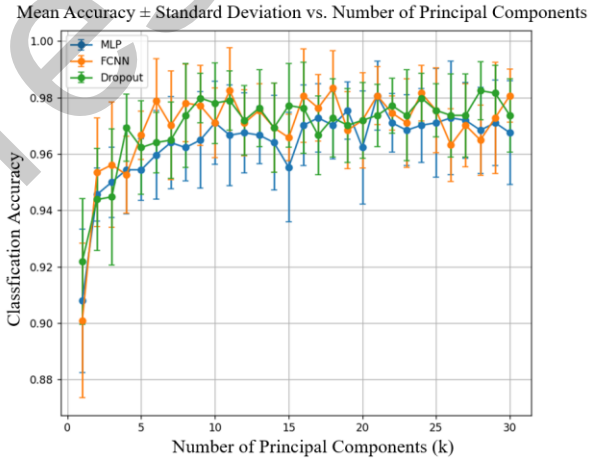

Fig. 5. Average accuracy of DL models with k different data points

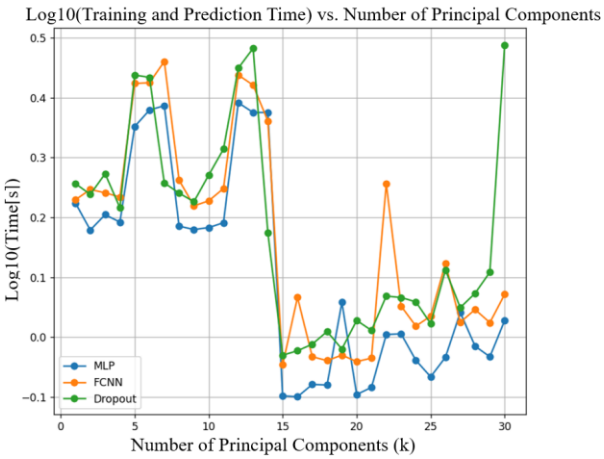

Fig. 6. Training and prediction time of DL models with k different data points

3.3. Clinical Performance Evaluation

Table 3. Clinical performance metrics at optimal PCA dimensions

Model	k	Precision	Recall	AUC	F1
SVM	10	0.9795	0.9595	0.9958	0.9689
KNN	15	0.9781	0.9357	0.9907	0.9558
RF	7	0.9418	0.9230	0.9896	0.9311
FCNN	13	0.9719	0.9619	0.9963	0.9665
Dropout	11	0.9813	0.9619	0.9958	0.9711

Table 3 presents precision, recall, F1-score, and AUC for each model at its optimal PCA dimension. It should be noted that the optimal PCA dimensions reported in Table 2 are based on overall accuracy, whereas Table 3 presents models selected based on clinically relevant metrics such as recall and AUC. Therefore, slight differences in optimal k values may occur depending on the evaluation criterion. Although k=25 yielded the highest overall accuracy for SVM, k=10 was selected for clinical evaluation as it achieved an optimal balance between high AUC (0.9958) and model parsimony (reducing features by 66%).

Among classical machine learning models, SVM (k=10) achieved high recall (0.9595) and AUC (0.9958), indicating strong discriminative capability. Among deep learning models, the Dropout network (k=11) obtained the highest precision (0.9813) and F1-score (0.9711), while FCNN (k=13) achieved the highest AUC (0.9963).

Overall, models with higher recall and AUC are more desirable for breast cancer diagnosis, as minimizing false negatives is clinically critical. In this context, the Dropout network and FCNN demonstrated the most balanced and clinically favorable performance. Although Random Forest achieved competitive performance, its recall (0.9230) was lower compared to SVM and neural network models, indicating a relatively higher risk of false negatives. All reported values have been carefully rechecked to ensure full consistency between the tables and the corresponding discussion.

4. Discussion

4.1. Impact of PCA on Model Performance

This study systematically evaluated the effect of PCA dimensionality on both classical machine learning and deep learning models for breast cancer classification. The results demonstrate that moderate dimensionality reduction ($k \approx 10$–20) preserves nearly all discriminative information while maintaining or improving classification performance.

For classical machine learning models, dimensionality reduction improved performance compared to the original 30-dimensional space. For example, SVM accuracy increased from 0.9602 ($k = 30$) to 0.9795 at $k = 25$. Similarly, KNN improved from 0.9507 to 0.9711 at $k = 15$. Random Forest achieved its highest accuracy of 0.9518 at $k = 5$, compared to 0.9257 in the original feature space. Across models, accuracy declined substantially when $k < 5$, indicating insufficient retained variance.

The PCA variance analysis showed that the first 10 principal components retained approximately 90% of total variance, while 15 components preserved more than 95%. Notably, performance saturation in both ML and DL models occurred when cumulative explained variance exceeded approximately 95–98% ($k \approx 10$–15), suggesting that additional components contributed limited incremental benefit.

For deep learning models, accuracy increased rapidly from very low dimensions to $k \approx 10$–15 and then plateaued. Within the range $k = 12$–18, models maintained near-maximum accuracy while preserving most of the dataset variance. These findings indicate that even neural architectures do not require the full 30-dimensional space to achieve optimal performance on this dataset.

Compared with [15], important methodological differences emerge. It is reported that "the cumulative explained variance reaches approximately 90% after five principal components" and selected the top five components accordingly. Their study fixed $k = 5$ (≈90% cumulative variance) for all models and did not evaluate multiple dimensional settings. Furthermore, they did not conduct a sensitivity analysis across $k = 1$–30, nor did they include deep learning architectures.

In contrast, the present study systematically evaluated performance across the full dimensional range ($k = 1$–30), enabling identification of a performance plateau region rather than relying on a single variance threshold. The results show that although 90% variance is reached at approximately 10 components in our analysis, optimal and stable performance was more consistently observed when cumulative variance exceeded 95%, typically at $k \geq 10$–15. This suggests that selecting PCA dimensions solely based on a 90% variance threshold may not always capture the optimal classification region.

Additionally, while [15] evaluated Decision Trees, Logistic Regression, SVM, and LDA, the present study extends the analysis to deep learning models (FCNN and Dropout network), providing broader methodological coverage. Additionally, compared to other published studies, our results demonstrate comparable or improved performance while providing a more comprehensive dimensional sensitivity analysis.

4.2. Clinical Implications

From a clinical perspective, recall and AUC are particularly critical due to the consequences of false negatives in breast cancer diagnosis.

At optimal PCA dimensions, SVM achieved recall = 0.9595 and AUC = 0.9958. The FCNN achieved the highest AUC (0.9963) with recall = 0.9619. The Dropout network also maintained recall = 0.9619 and achieved the highest precision (0.9813) and F1-score (0.9711). In comparison, Random Forest recall was lower (0.9230), despite competitive overall accuracy. The differences in AUC across models were relatively small (0.9896–0.9963), indicating strong class separability after PCA transformation. However, the approximately 3%–4% recall gap between Random Forest and neural network models may be clinically relevant in scenarios where minimizing false negatives is prioritized.

These findings reinforce that moderate PCA dimensionality ($k \approx 10$–15) supports high recall and near-perfect AUC across multiple model families, providing flexibility in model selection depending on deployment constraints.

4.3. Computational Considerations

While PCA reduced feature dimensionality substantially, the impact on computational time was not strictly linear. For machine learning models, modest reductions in training time were observed in certain dimensional ranges; however, further dimensional reduction did not always yield proportional gains.

For deep learning models, training time showed only limited sensitivity to PCA dimensionality. This phenomenon may be attributed to the relatively small dataset size (569 samples), where computational cost is dominated more by network operations and optimization steps than by input dimensionality. Consequently, PCA's primary advantage in this context appears to lie in stability and noise reduction rather than dramatic computational acceleration. In larger biomedical datasets with higher feature counts, the computational benefits of PCA may become more pronounced.

5. Conclusion

This study systematically evaluated the impact of PCA dimensionality on machine learning and deep learning models for breast cancer classification. The results show that moderate dimensionality ($k \approx 10$–15), corresponding to approximately 95%–98% cumulative variance, is sufficient to achieve near-optimal performance, while overly aggressive reduction ($k < 5$) degrades accuracy. Among the evaluated models, SVM and deep learning architectures demonstrated strong performance, with neural networks achieving the highest AUC (up to 0.9963) and recall. These findings highlight the importance of selecting an appropriate PCA dimension to balance accuracy and computational efficiency. Compared to prior studies using fixed PCA thresholds, this work provides a comprehensive dimensional sensitivity analysis across $k = 1$–30 and extends evaluation to both ML and DL models.

References

[1] World Health Organization, WHO launches new roadmap on breast cancer, World Health Organization, Geneva, 2023.

[2] World Health Organization, Vietnam fact sheets, 2025. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/populations/704-viet-nam-fact-sheet.pdf

[3] A. Tata, M. Woolman, M. Ventura, N. Bernards, M. Ganguly, A. Gribble, B. Shrestha, E. Bluemke, H. J. Ginsberg, A. Vitkin, J. Zheng, anf A. Zarrine-Afsar, Rapid detection of necrosis in breast cancer with desorption electrospray ionization mass spectrometry, Scientific reports, vol. 6, iss. 1, Oct. 2016, Art. no. 35374. https://doi.org/10.1038/srep35374

[4] World Health Organization, "Cancer," World Health Organization, 3 February 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[5] E. M. F. E. Houby, Framework of computer aided diagnosis systems for cancer classification based on medical images, Journal of Medical Systems, vol. 42, iss. 8, pp. 1–12, Jul. 2018. https://doi.org/10.1007/s10916-018-1010-x

[6] T. S. Lim, K. G. Tay, A. Huong and X. Y. Lim, Breast cancer diagnosis system using hybrid support vector machine-artificial neural network, International Journal of Electrical and Computer Engineering., vol. 11, no. 4, pp. 3059–3069, 2021. http://doi.org/10.11591/ijece.v11i4.pp3059-3069

[7] D. L. Banks and S. E. Fienberg, Curse of dimensionality, in Encyclopedia of Physical Science and Technology (Third Edition), 2003.

[8] A. Paul, S. Paul, E. Gamukama and K. Margaret, Exploring dimensionality reduction techniques for improved performance, Journal of Applied Science and Information Science, vol. 5, iss. 1, p. 10, Jun. 2024.

[9] L. V. D. Maaten, E. Postma, and J. V. D. Herik, Dimensionality reduction: A comparative review, Journal of Machine Learning Research, vol. 20, no. 1, p. 6, 2007.

[10] I. de-la-Bandera, D. Palacios, J. Mendoza, and R. Barco, Feature extraction for dimensionality reduction in cellular networks performance analysis, Sensors, vol. 20, iss. 23, p. 6944, Dec. 2020. https://doi.org/10.3390/s20236944

[11] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Joseph, A Review of dimensionality reduction techniques for efficient computation, Procedia Computer Science, vol. 165, pp. 104-111, Nov. 2019. https://doi.org/10.1016/j.procs.2020.01.079

[12] S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems, vol. 2, iss. 1-3, pp. 37-52, Aug. 1987. https://doi.org/10.1016/0169-7439(87)80084-9

[13] N. Panahi, M. G. Shayesteh, S. Mihandoost, and B. Z. Varghahan, Recognition of different datasets using PCA, LDA, and various classifiers, in 5th International

Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, Oct. 2011. https://doi.org/10.1109/ICAICT.2011.6110912

[14] S. M. Smith, A. Hyvarinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann, Group-PCA for very large fMRI datasets, NeuroImage, vol. 101, pp. 738-749, Nov. 2014. https://doi.org/10.1016/j.neuroimage.2014.07.051

[15] G. Esen, A. Altaibek, J. Amankulov, B. Matkerim, and M. Nurtas, Enhancing breast cancer detection with dimensionality reduction techniques: a study using pca and lda on wisconsin breast cancer data, Procedia Computer Science, vol. 251, pp. 414-421, 2024. https://doi.org/10.1016/j.procs.2024.11.128

[16] K. Luo, Application of principal component analysis in the diagnostic classification of breast cancer, in Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023), Nov. 2023. https://doi.org/10.2991/978-94-6463-300-9_72

[17] W. Wolberg, O. Mangasarian, N. Street and W. Street, Breast Cancer Wisconsin (Diagnostic), 1993. [Online]. Available: http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. [Accessed 31 12 2025]