



Integrated Ensemble Strategy for Breast Cancer Detection Using Dimensionality Reduction Technique

Zulfikar Ali Ansari^a, Mohammad Arif^b, Nagendra Babu Rajaboina^c, Anwar Ahamed Shaikh^e, and Yaduvir Singh^d

^a Department of AI and ML, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra-412115, India

^b Department of Computer Science & Engineering, Parul University, Vadodra-391760, India

^c Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh-522302, India

^d Department of CSE-AI, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh-201306, India

^e School of Engineering and Technology, Sanjivani University Kopergaon Maharashtra
✉ zulf78692@gmail.com, arif_mohd2k@yahoo.com, rnagendrababu@kluniversity.in,
yaduyash@gmail.com, anwar.nizami@gmail.com

KEYWORDS

breast cancer;
ensemble learning;
machine learning;
dimensionality
reduction

ABSTRACT

Breast cancer remains a critical global health concern, requiring advanced and reliable diagnostic methods for early detection and effective intervention. This work introduces an integrated ensemble framework that combines multiple dimensionality reduction (DR) techniques, including Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), and Singular Value Decomposition (SVD), with robust machine learning (ML) classifiers for improved breast cancer detection. The publicly available Wisconsin Breast Cancer Dataset (WBCD) was utilized, with rigorous data preprocessing performed to address missing values, anomalies, and class imbalance through stratified sampling and median imputation. To mitigate overfitting and underfitting, dimensionality reduction was coupled with cross-validation and ensemble strategies. The predictive performance of Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Multi-Layer Perceptron (MLP) was systematically evaluated. Experimental results show that SVM consistently achieves a maximum accuracy of 97.9 % across all applied DR techniques, while MLP and LR also reach 97.9 %



accuracy with PCA and NMF, though MLP exhibits performance variability depending on the selected DR method. The findings provide practical guidance for healthcare practitioners and researchers, supporting the adoption of explainable and scalable AI-driven diagnostic tools. Limitations include the reliance on a single dataset and the need for further validation on larger and more diverse clinical cohorts. Future work will focus on enhancing model interpretability, external validation, and real-world deployment in resource-constrained settings.

1. Introduction

Millions of people worldwide suffer from breast cancer, which has a high death rate (Abdullah, 2023). Early and accurate prediction is fundamental for improving patient outcomes and survival rates, but timely identification remains a clinical challenge. In recent years, the integration of ML and data science into breast cancer diagnostics has enabled the development of novel, data-driven methodologies that enhance the precision and efficiency of disease detection (Ashokkumar et al., 2022; He et al., 2023)

In response, the field has witnessed a growing adoption of ML and data-driven methods for breast cancer diagnostics, aiming to support clinicians with enhanced predictive accuracy and efficient decision-making (Fatima et al., 2020). One key challenge in leveraging ML for medical data analysis is the high dimensionality inherent in clinical datasets, which often leads to increased computational complexity, the risk of overfitting, and reduced interpretability (Jaiswal et al., 2023;). DR techniques, such as PCA, NMF, and SVD, have become essential tools in addressing these challenges and transformed high-dimensional data into a more compact representation, preserving the most informative features while mitigating noise and redundancy. Applied ML models, including both classical and ensemble approaches, to achieve improved generalization, computational efficiency, and robustness.

Our primary objective is to get the optimal number of components for which these techniques give the best accuracy in breast cancer diagnosis while addressing the computational challenges associated with high-dimensional data. To achieve this goal, we explore a range of DR methods, including PCA, NMF and SVD. In Table 1, we have explored a list of abbreviations in this study.

1.1. Research Motivation

Breast cancer is considered to be one of the most frequent and dangerous diseases globally (Feng et al., 2018). Early and accurate detection is crucial for successful treatment and improved patient outcomes. DR techniques can potentially enhance the performance of breast cancer detection models, making them more effective in clinical practice. These challenges can be addressed by reducing the dimensionality of the data while preserving critical information.

1.2. Research Contribution

In our study, we have made several significant contributions to the field of breast cancer detection by employing DR techniques. Our research has aimed to address the following:

- Conduct an extensive comparative analysis: Various ML Algorithms using DR techniques have been studied, including PCA, NMF, and SVD.

Table 1. List of abbreviations

S/N	Keyword	Full form
1	SVM	Support Vector Machine
2	ML	Machine Learning
3	PCA	Principal Component Analysis
4	SVD	Singular Value Decomposition
5	NMF	Non-Negative Matrix Factorization
6	DR	Dimensionality Reduction
7	BC	Breast Cancer
8	CM	Confusion Matrix
9	RF	Random Forest
10	DT	Decision Tree
11	MLP	Multilayer Perceptron
12	ANN	Artificial Neural Networks
13	KNN	K-Nearest Neighbour
14	WBCD	Wisconsin Breast Cancer Database

- Apply DR techniques: We have effectively reduced the dimensionality of the breast cancer dataset.
- Connects ML to clinical practice: This research encourages healthcare providers to use DR approaches by showing their efficacy in breast cancer diagnosis.

2. Related Work

Recent advancements in ML have significantly enhanced breast cancer detection methodologies. A study by Li et al. (2023) identified a hybrid approach combining dimensionality reduction with ensemble classification, demonstrating improved diagnostic accuracy over traditional methods. Similarly, Paul et al. (2024) explored various dimensionality reduction techniques, emphasizing their role in refining diagnostic precision and model interpretability.

Regarding a dataset in this field, WBCD comes from a well-known ML database known as the UCI ML database (Islam et al., 2020). used the WBCD dataset with a variety of ML techniques to assess the efficacy and utility of these algorithms in diagnosing malignant and benign BC. (Bansal et al., 2022) provided a comparative evaluation of five ML algorithms to anticipate BC with models such SVM, K-NN, RF, ANN, and LR. In this study, Harinishree et al. (2021) compared three commonly used ML techniques for predicting breast cancer: NB, RF, and KNN. The WBCD data set was used as a training set to examine the effectiveness of several ML algorithms in terms of crucial characteristics, such as precision and accuracy. Naji et al. (2021) employed the ML algorithms only after collecting the data, and they compared and assessed how well these various classifiers performed.

In the field of deep learning, a novel convolutional neural network model was proposed to detect breast cancer across mammograms with varying densities, achieving notable accuracy improvements. Complementing this, Pérez-Núñez et al. (2024) developed a computer-aided diagnosis system utilizing

deep feature transfer learning, which enhanced the classification of breast tumours in histopathology images. Table 2 shows an examination of several strategies utilized for the diagnosis of BC. This examination included an analysis of the methodologies employed, the dataset utilized, and the corresponding accuracy information.

However, current studies still face notable gaps, including insufficient handling of class imbalance, limited data quality, and concerns regarding the interpretability and scalability of ML models for clinical deployment). Additionally, many previous studies have evaluated DR or classification models in isolation, rather than integrating both into a unified, comparative framework that aligns with the needs of contemporary healthcare.

Table 2. Existing methods and their corresponding accuracies

Author / Year	Methodology	Dataset		Accuracy
		WBCD	OTHERS	
(Sengar et al., 2020)	LR and DT	✓		95 %
(Ara et al., 2021)	SVM, LR, K-NN, DT, NB and RF classifiers	✓		96.5 %
(Kabiraj et al., 2020)	XGBoost, RF	✓		74.18 %
(Al-Fahaidy et al., 2022)	SVM		MIAS dataset of mammogram images	87.10 %
(Binsaif, 2022)	DT, RF, K-NN, ANN, SVM and LR		BC Database of Coimbra (UCI)	64 %
(Dwarakanath and Bhairagond, 2023)	SVM and CNN		Histopathological images	95 %
(Botlagunta et al., 2023)	XGBoost, LR, KNN, DT, RF, SVM,		BIACH and RI as a semi-structured	83 %
(Yu et al., 2024)	AI-CAD		398 consecutive patients with 414 breast cancers	91.5 %
(Esfahani et al., 2024)	deep neural network		4,621 breast cancer mammogram images.	84 %,
(Tai et al., 2024)	CNN		253 patients	87.7

This study mainly addresses these limitations by introducing an integrated ensemble framework that combines multiple DR techniques, PCA, NMF, and SVD, with a suite of advanced ML classifiers using the widely recognized WBCD. We employ comprehensive data preprocessing steps, including missing value imputation, anomaly detection, and stratified sampling to handle class imbalance. Cross-validation and ensemble learning approaches are incorporated to further mitigate overfitting and enhance model robustness.

3. Materials and Method

This section provides a comprehensive overview of the methodological framework employed in this study. The overall workflow is illustrated in Figure 1, which outlines the sequential processes involved

in our strategies. The proposed classification model integrates multiple machine learning algorithms to enhance feature representation, and DR techniques are utilized. The flowchart describes the key stages of the approach: dataset collection, data preprocessing, feature extraction, feature analysis, DR techniques, and ML Classifiers. Each of these components is detailed in the subsequent sections.

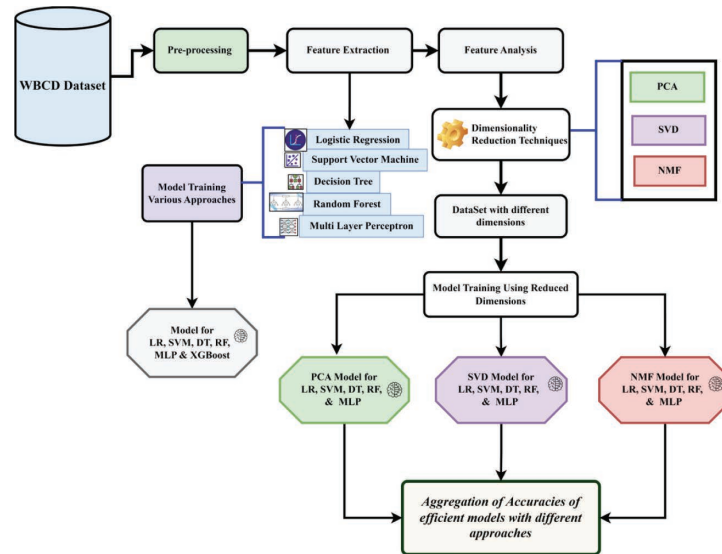


Figure 1. Proposed approach

3.1. Dataset Description

The dataset utilized in this work was found from the UCI ML Repository, which is affiliated with the University of California, Irvine (UCI). The UCI ML repository is a publicly accessible repository that hosts several datasets suitable for doing empirical examinations of ML algorithms. The dataset was initially provided by the University of Wisconsin and consists of 569 samples (212 malignant and 357 normal), each with 32 characteristics. These characteristics represent the basic properties of breast mass cell nuclei.

To ensure robust evaluation, the dataset was divided into training and testing subsets, with 80 % for training and 20 % for keeping as an unseen split, preserving the original ratio of malignant and benign cases. The split details are presented in Table 3.

Table 3. Training and testing split of the dataset

Split Type	Number of Instances	Malignant	Benign
Training Set (80 %)	455	170	285
Testing Set (20 %)	114	42	72

3.2. Preprocessing

Data preprocessing is an important step that helps to reduce noise, inconsistencies, and redundancy to provide high-quality data that increases performance (Maharana et al., 2022). Here, we first checked for missing values in this dataset, then removed any duplicates and handled the missing data by filling them in with the median of the respective values. Anomalies and outliers were identified using z-score analysis and visual inspections. Outlier handling was performed by either capping extreme values or removing them to ensure the reliability of the dataset.

3.3. Feature Extraction

Feature selection is a helpful stage in data modelling that aids in removing unnecessary and noisy data (Bhatia et al., 2019). The suggested study uses important factors to pick them. How the number of cancerous and benign cells connect to the traits may be seen graphically. This enables us to determine the features to consider when projecting the outcome.

3.4. Feature Analysis

Feature analysis, which is often referred to as feature engineering or feature selection, plays a vital role in the ML process. The task involves the procedure of choosing, modifying, or generating related characteristics (input variables or attributes) from the unprocessed data to enhance the efficacy of an ML model. It has been shown that the malignant cells exhibit greater mean values compared to the benign cells across these parameters, as well as the remaining 20 columns representing standard error and worst values for these 10 parameters.

3.4.1. DR Technique

DR: It plays a vital role in ML by allowing the transformation of data that is highly dimensional into a lower-dimensional environment, while at the same time retaining the fundamental characteristics of the data. It enhances the efficient use of data while concurrently alleviating the computational workload on automated operations (Gupta and Janghel, 2019).

PCA: It is a widely employed approach in the fields of ML and data analysis for the purpose of reducing the dimensionality of a dataset. Its primary objective is to minimize the number of features while retaining a significant portion of the initial variance. The algorithm is designed to process a data matrix X with dimensions $m \times n$, where m represents the number of observations and n represents the number of variables.

Covariance Matrix Calculation in the Context of Breast Cancer: Compute the covariance matrix for the centred breast cancer data. In this context, the covariance between two features, representing aspects of the breast cancer dataset, denoted as i and j , is determined using the following formula:

$$\text{cov}(\text{Feature}_i, \text{Feature}_j) = \frac{1}{n-1} \sum_{k=1}^n (\text{Feature}_{ki} - \mu_i)(\text{Feature}_{kj} - \mu_j) \quad (1)$$

This computation yields an $m \times m$ covariance matrix, which is significant for breast cancer analysis and is denoted as Σ .

SVD: The SVD is a fundamental approach for matrix factorization that is widely employed in the fields of linear algebra and numerical linear algebra. We will take three matrices, namely U , Σ , V , for

decomposing any matrix C of order $n \times d$. For any matrix $C \in R^{n \times d}$, U and V are the two orthogonal matrices of order $n \times n$ and $d \times d$. Σ is a nonnegative, “diagonal” matrix $\Sigma \in R^{n \times d}$. Mathematically, for a given matrix A , SVD factorizes it as follows:

Let A denote the initial matrix with dimensions $m \times n$.

The matrix U is a $m \times m$ orthogonal matrix, satisfying the condition $UU^T = I$, where I represents the identity matrix. The matrix Σ is a diagonal matrix of dimensions $m \times n$, where the diagonal elements consist of non-negative real numbers. These numbers are referred to as singular values and are arranged in decreasing order, denoted as $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$, where r is the rank of matrix A . The transpose of matrix V , denoted as V^T , is an orthogonal matrix of size $n \times n$, satisfying the property $VV^T = I$. The left singular vectors of matrix U are denoted as the columns and serve as representations of the directions that exhibit the highest variation among the rows of matrix A .

The right singular vectors, denoted as V , in matrix A are referred to as the columns that indicate the directions of highest variance in the columns of A . It starts with the original matrix A . $A^T * A$ and $A * A^T$ are computed. These are called the Gramian matrices. Then, the eigenvalues and eigenvectors of both $A^T * A$ and $A * A^T$ are computed.

$$\Sigma = \begin{bmatrix} \sqrt{\sigma_1} & & & \\ & \sqrt{\sigma_2} & & \\ & & \sqrt{\sigma_3} & \\ & & & \dots \end{bmatrix} \quad (2)$$

The square root of the eigenvalues of $A^T * A$ (or $A * A^T$) forms the singular values (Σ) in descending order. These singular values are non-negative and represent the importance of the corresponding singular vectors. The eigenvectors are normalized to form the columns of U and V . The left singular vectors (U) correspond to the eigenvectors of $A * A^T$, and the right singular vectors (V) correspond to the eigenvectors of $A^T * A$. Once U , Σ , and V have been obtained, the original matrix A can be reconstructed using the formula.

$$A = U\Sigma V^T \quad (3)$$

NMF: NMF is a matrix factorization technique used in linear algebra and data analysis. Unlike SVD, which factorizes a matrix into the product of three matrices, NMF factorizes a matrix into the product of two non-negative matrices, which makes it particularly useful for certain types of data analysis, including feature extraction, image processing, and text mining. Mathematically, for a given non-negative matrix X of size $m \times n$, NMF seeks two non-negative matrices, W ($m \times r$) and H ($r \times n$), with r usually less than m and n . As a factorization:

$$X = WH \quad (4)$$

Where X is the original non-negative matrix to be factorized. W is a non-negative matrix of size $m \times r$. Each column of W represents a basis vector, and these basis vectors are combined to approximate the data in X . H is a non-negative matrix of size $r \times n$.

Dimensionality reduction is crucial in medical data analysis to combat the “curse of dimensionality,” which can hinder model performance and increase computational demands. Techniques such

as PCA transform correlated variables into principal components that capture the majority of data variance, enabling more efficient and accurate learning by classifiers.

3.5. ML Classifiers

In this section, we have discussed classification models mainly using LR, SVM, DT, RF and MLP.

3.5.1. LR

LR is a statistical and ML model commonly employed for binary classification tasks. This model enables the prediction of one of two potential outcomes by utilizing a collection of input information. In binary classification problems, there are two possible outcomes, often labelled as 0 and 1, or “Benign” and “Malignant”. LR uses a set of input features (also called independent variables or predictors) to make predictions. These features can be numeric or categorical. LR models the relationship between the input features and the probability of the positive outcome. It does this by taking a linear combination of the input features, weighted by coefficients:

$$\text{Log}(\text{Odds}) = \text{Log}\left(\frac{P(Y = \text{Malignant} | X)}{P(Y = \text{Benign} | X)}\right) = \text{Logit}(P(Y = \text{Malignant} | X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (5)$$

$P(Y = \text{Malignant})$ represents the probability of the patient having cancer where $X_1, X_2, \dots, X_n$ are the input features, β_0 is the intercept, and $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients associated with each feature.

To ensure that the output is between 0 and 1 (representing probabilities), the linear combination is passed through a logistic function also known as the sigmoid function:

$$P(Y = \text{Malignant}) = \frac{1}{1 + e^{-z}} \quad (6)$$

Where z is the linear combination. The sigmoid function transforms the linear combination into a value between 0 and 1, which can be interpreted as the probability of the positive outcome. To make a binary prediction, you need to choose a decision threshold (usually 0.5). If the predicted probability is greater than or equal to the threshold, the model predicts the positive class (1); otherwise, it predicts the negative class (0).

3.5.2. SVM

SVM is widely acknowledged as a prominent approach in the domain of supervised learning. It is frequently utilized for both classification and regression tasks. In this study, SVM was utilized as one of the core classification algorithms for breast cancer detection. Outstanding in its robust handling of high-dimensional data and binary classification tasks, SVM is particularly well-suited for medical diagnostic applications such as distinguishing between cases of benign and malignant breast cancer.

Mathematically, it aims to classify an optimal hyperplane that separates two classes with the maximum margin. Given a labelled training dataset $\{x_i, y_i\}_{i=1}^n$, where $x_i \in R^d$

represents the input features and $y_i \in \{-1, +1\}$ denotes the class labels, the decision boundary is defined as:

$$w^T x + b = 0 \quad (7)$$

3.5.3. DT

The DT technique is widely utilized in supervised ML for the purposes of classification and regression. In this study, DT was utilized as one of the classification algorithms to assess its performance in conjunction with DR techniques for breast cancer prediction. A decision tree recursively partitions the input space based on feature values, forming a tree-like structure where each internal node represents a decision based on a feature, each branch corresponds to an outcome of the decision, and each leaf node denotes a class label. The model aims to learn a set of if-then rules that can accurately classify input samples.

The main purpose of this algorithm is to select the feature that provides the maximum information gain or reduction in impurity at each split. The most used metric is the Gini Index, defined as:

$$Gini(t) = 1 - \sum_{i=1}^C p_i^2 \quad (8)$$

where:

t is a node,

C is the number of classes,

p_i is the proportion of samples belonging to class i at node t .

3.5.4. RF

The RF algorithm is a popular ensemble learning approach in the field of ML, commonly employed for both classification and regression applications. In this work, a robust ensemble-based classification algorithm is used to assess its effectiveness in breast cancer detection, particularly when combined with DR techniques. RF works by constructing a collection of decision trees during training and aggregating their outputs to produce the final output. We have utilized classification tasks, which output the class with the majority vote among all trees in the ensemble. Each tree is trained on a bootstrapped sample of the dataset, and at each split, only a random subset of features is considered, thereby introducing diversity and reducing model variance.

Mathematically, let $\{T_1(x), T_2(x), \dots, T_k(x)\}$ be the predictions of k individual decision trees. The final prediction $\hat{y}$ for an input sample x is given by:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_k(x)\} \quad (9)$$

3.5.5. MLP

It is a type of feedforward artificial neural network, which was implemented in this study to evaluate its performance in classifying breast cancer cases when used in conjunction with various DR techniques. MLP is a powerful supervised learning model capable of capturing complex nonlinear relationships in data, making it a suitable choice for medical diagnostic tasks involving intricate feature patterns.

It mainly consists of an input layer, one or more hidden layers with non-linear activation functions, and an output layer. Each neuron in each layer is connected to all neurons in the subsequent layer, and the model learns the weights of these connections through backpropagation by minimizing a loss function, such as cross-entropy for classification.

3.6. Performance measures

Accuracy, Precision, Recall, Specificity, and F-measure are the assessment metrics used for classification. The components of CM, which provide information about projected and actual outcomes, are used to construct these measures. Real-world values of the following equations represent the performance metrics.

True Positive= True

Malignant = TM, True Negative = True Benign=TB

$$Accuracy = \frac{TM + TB}{TM + TB + FM + FB} \quad (10)$$

$$Recall = \frac{TM}{TM + FB} \quad (11)$$

$$Precision = \frac{TM + TB}{TM + TB + FM + FB} \quad (12)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

$$Specificity = \frac{TB}{FM + TB} \quad (14)$$

3.7. Handling Class Imbalance

Class Imbalance is a critical task in ML. Given the moderate class imbalance in the dataset, we have adopted stratified sampling to maintain class proportions in both training and testing sets. For certain classifiers, such as SVM and LR, class weights were adjusted to compensate for imbalance, helping the models to better learn from minority class samples.

4. Experimental Setup and Result Analysis

In this section, the experimental procedures are defined, the setup is described, and the tools used for conducting a comparative analysis of DR techniques in the context of BC detection are discussed. We have imported essential libraries such as Scikit-Learn, Matplotlib, Pandas, and Seaborn for data analysis, DR, and visualization. We have employed various models, including LR, SVM, DT, RF and MLP for classification tasks on both the original and reduced-dimensional datasets. Hyperparameters for all classifiers and dimensionality reduction techniques were optimized using grid search combined with cross-validation. This approach facilitated optimal model configuration and minimized the risk

of overfitting. We have assessed the performance of each model using metrics such as accuracy, precision, recall, and F1-score, as shown in Figure 4a. With NMF, we have achieved maximum accuracy from MLP, which is 97.9 %. After that, we applied PCA on the datasets and again trained all these models with all components. Finally, we have achieved accuracy in all models, and now we observe that some models' performance is improved, achieving higher accuracy. The application of SVM and PCA rendered an accuracy of 97.9 %, as shown in Figure 4b. Similarly, the combined use of LR and PCA also resulted in an accuracy of 97.9 %. Meanwhile, the implementation of MLP and PCA led to a slightly lower accuracy of 97.2 %. Then, we used another dimensionality technique, namely SVD, on the same dataset. We then evaluated the performance of different classifiers using the components produced by SVD, as shown in Figure 4c. Among these classifiers, SVM achieved the highest accuracy of 97.9 %. The next best accuracy came from MLP, at 97.2 %. In Figure 4d, we compared the performance of all the models, namely, LR, RF, DT, SVM, and and MLP combined with different DR techniques, specifically NMF, PCA and SVD. Here we can see that SVM consistently achieved an accuracy of 97.9 % across all the DR techniques. On the other hand, MLP achieved its best performance (97.9 %) with SVD, as also shown in Figure 4c, but its accuracy varied when using NMF and PCA. Similarly, LR showed strong results with NMF and PCA (both 97.9 %), but its accuracy declined when using SVD.

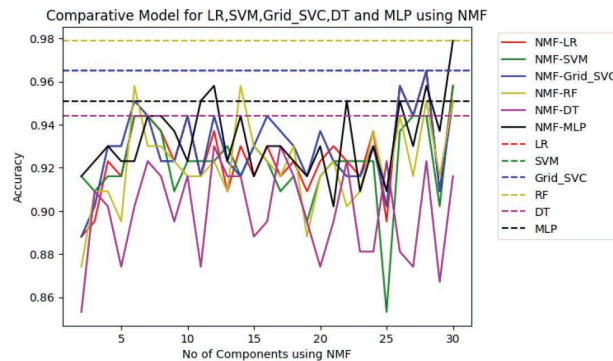


Figure 4a. Accuracy comparison for all models using NMF

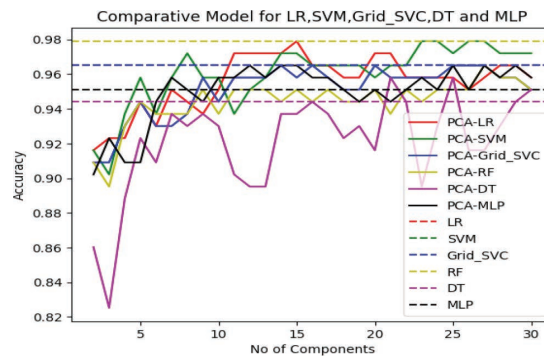


Figure 4b. Comparison of results with PCA

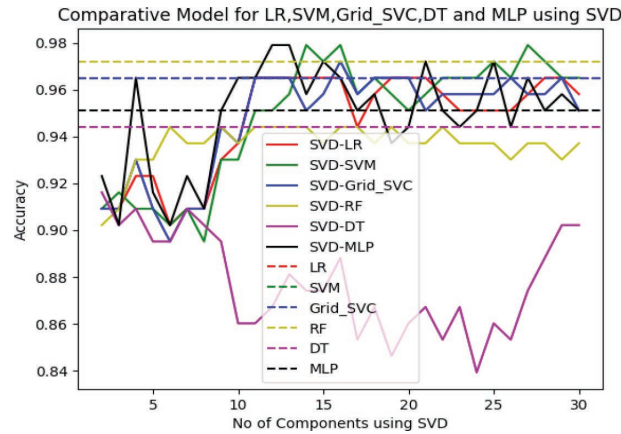


Figure 4c. Comparison of results with SVD

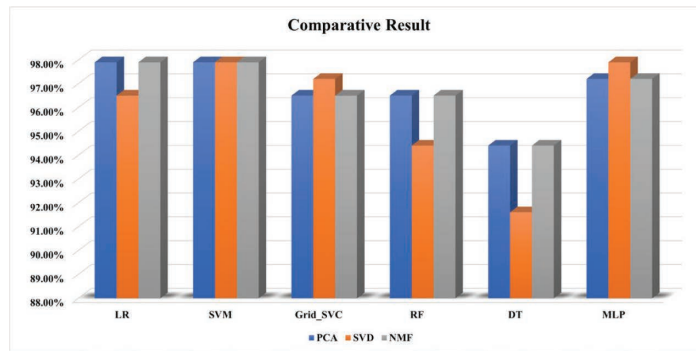


Figure 4d. Comparison of results of all models with DR Techniques

Finally, to address potential overfitting and underfitting, we have fully utilized dimensionality reduction techniques to eliminate redundant and noisy features. Here we can say the ensemble method, such as Random Forest, further reduces the risk of overfitting by aggregating predictions from multiple weak learners. Additionally, cross-validation was employed for hyperparameter selection, ensuring the generalizability of the models.

5. Discussion

Our study demonstrates that DR techniques can significantly enhance the performance of ML classifiers in BC detection. The reduction in feature dimensionality not only simplifies the data but also improves the classifiers' ability to capture essential patterns. PCA, NMF and SVD emerged as particularly effective techniques in our comparative analysis, and finally, we achieved

a maximum accuracy of 97.90 %. This finding underscores the importance of feature selection and DR in optimizing the BC detection model. Table 4 shows that our model has good accuracy as compared to the existing model. The consistent improvement in classifier performance across these datasets highlights the robustness and generalization capabilities of DR techniques. The observed variability in MLP performance with different DR techniques is attributed to its sensitivity to feature transformations and scaling. MLP, as a neural network, can exhibit non-linear interactions with features, making its performance more variable. In contrast, SVM tends to maintain consistent results across different DR approaches due to its reliance on support vectors and margin maximization.

Table 4. Comparison of DR technique models

Author	DR Techniques	Accuracy	Year
(Mishra and Rath, 2021)	PCA and Auto encoder	95.45	2021
(Rimi et al., 2022)	kernel-PCA	96.03	2022
(Mishra et al., 2023)	t-SNE, LLE and KPCA	89.49	2023
(Kabir et al., 2023)	PCA and Auto encoder	95.50	2023
(Karuppasamy et al., 2024)	NMF	90.00	2024
Our Proposed Model	PCA, SVD and NMF	97.90	2025

5.1. Comparison

Within the context of an ensemble strategy, every method for detecting breast cancer that makes use of a machine learning classifier and dimensionality reduction techniques has both scope and challenges. The properties of the dataset, the computational resources available, and the level of interpretability that is needed all play a role in the selection of the approach that is most appropriate. In Table 4, we have done comparisons only with DR techniques such as PCA, kernel PCA, and Autoencoder, with a Classifier using some existing model, and we compared with our proposed work, also, and achieved the best result as compared to previous research.

5.2. Limitations and Future Scope

It is essential to acknowledge certain limitations in our research. While we have demonstrated the effectiveness of DR techniques, the primary dataset is relatively small and may not fully reflect the heterogeneity of real-world clinical data. Additionally, our study primarily focused on established DR techniques. Consequently, external generalizability could be limited. While dimensionality reduction and ensemble methods help mitigate overfitting, further research is needed to enhance interpretability critical for clinical adoption. Future research can explore interpretable ML frameworks, novel DR techniques, and automated model deployment in diverse clinical scenarios. Models with high accuracy, such as MLP, may lack transparency. Clinicians frequently require interpretable models to understand, validate, and trust AI-powered decisions. Overdependence on non-interpretable models may result in unwillingness to adopt or in misdiagnosed concerns if the logic cannot be understood or disputed.

5.3. Clinical Implications

The integrated ensemble framework presented in this study can be leveraged by healthcare practitioners to improve the accuracy and efficiency of early breast cancer detection. By applying dimensionality reduction prior to classification, practitioners can efficiently handle high-dimensional clinical data and achieve robust diagnostic performance. Researchers are encouraged to validate and refine these methods on larger and more diverse patient cohorts for broader clinical adoption.

6. Conclusion

This study has mainly demonstrated the efficacy of integrating dimensionality reduction techniques with machine learning classifiers for the early detection of breast cancer. We have evaluated the performance of various widely used classifiers such as LR, SVM, DT, RF, and MLP in combination with three DR methods: PCA, NMF, and SVD. The DR techniques effectively reduced the high-dimensional feature space, thereby enhancing classification efficiency and reducing overfitting risk. Finally, the models were evaluated, with SVM consistently achieving the highest accuracy of 97.9 % across all three DR methods, highlighting its robustness and adaptability to transformed feature spaces. While PCA proved especially efficient for preprocessing large datasets, the variability observed with MLP and the differential performance of classifiers with different DR techniques underscore the importance of aligning the choice of preprocessing method with the specific characteristics of the classification algorithm. The findings prove a systematic comparative analysis and offer practical insights into the optimal combinations of dimensionality reduction and classification strategies for breast cancer detection. This research contributes to the growing field of AI-driven healthcare by presenting a scalable, generalizable, and interpretable framework that can guide clinicians and researchers in deploying robust diagnostic tools.

Acknowledgements

We would like to express our heartfelt gratitude to all the people and organizations whose assistance and efforts made this research project possible. We sincerely thank Dr. Rafeeq Ahmed for his essential advice, mentorship, and supervision. His experience and insights had a big impact on the planning and execution of this work. We would like to thank Symbiosis Institute of Technology, Symbiosis International (Deemed University), for providing the facilities, infrastructure, and resources that were needed to carry out this research.

Disclosure statement: The authors report there are no competing interests to declare.

Informed Consent Statement: Not applicable.

Data Availability Statement: Available on UCI Repository.

References

Abdullah, N. N. M. (2023). Coercive Biomedical Body Politics: Redefining Breast Cancer as a Gender-Marked Experience in the Case Study of Linda Park-Fuller's 'A Clean Breast of It.' *International Journal of Arabic-English Studies*, 23(1), 327–344. <https://doi.org/10.33806/IJAES2000.23.1.17>

- Al-Fahaidy, F. A. K., Al-Fuhaidi, B., AL-Darouby, I., AL-Abady, F., AL-Qadry, M., & AL-Gamal, A. (2022). A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques. *Applied Computational Intelligence and Soft Computing*, 2022, 1–17. <https://doi.org/10.1155/2022/3895976>
- Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *2021 International Conference on Artificial Intelligence (ICAI)*, 97–101. <https://doi.org/10.1109/ICAI52203.2021.9445249>
- Ashokkumar, N., Meera, S., Anandan, P., Murthy, M. Y. B., Kalaivani, K. S., Alahmadi, T. A., Alharbi, S. A., Raghavan, S. S., & Jayadhas, S. A. (2022). Deep Learning Mechanism for Predicting the Axillary Lymph Node Metastasis in Patients with Primary Breast Cancer. *BioMed Research International*, 2022. <https://doi.org/10.1155/2022/8616535>
- Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short-Term Memory algorithms in machine learning. *Decision Analytics Journal*, 3, e100071. <https://doi.org/10.1016/J.DAJOUR.2022.100071>
- Bhatia, V., Rawat, P., Kumar, A., & Shah, R. R. (2019). *End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT*. <http://arxiv.org/abs/1910.03089>
- Binsaif, N. (2022). Application of Machine Learning Models to the Detection of Breast Cancer. *Mobile Information Systems*, 2022. <https://doi.org/10.1155/2022/7340689>
- Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports 2023 13:1*, 13(1), 1–17. <https://doi.org/10.1038/s41598-023-27548-w>
- Esfahani, P. R., Maalouf, M. M., Reddy, A. J., & Chawla, P. (2024). Abstract A088: Utilizing Machine Learning Techniques to Investigate Mammograms for Breast Cancer Detection. *Cancer Research*, 84(3_Supplement_1), A088–A088. <https://doi.org/10.1158/1538-7445.ADVBC23-A088>
- Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, 8, 150360–150376. <https://doi.org/10.1109/ACCESS.2020.3016715>
- Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppapapati, A., Luu, H. H., Haydon, R. C., He, T. C., & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes and Diseases*, 5(2), 77–106. <https://doi.org/10.1016/J.GENDIS.2018.05.001>
- Gupta, K., & Janghel, R. R. (2019). Dimensionality reduction-based breast cancer classification using machine learning. *Advances in Intelligent Systems and Computing*, 798, 133–146. https://doi.org/10.1007/978-981-13-1132-1_11/COVER
- Harinishree, M. S., Aditya, C. R., & Sachin, D. N. (2021). Detection of Breast Cancer using Machine Learning Algorithms - A Survey. *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, 1598–1601. <https://doi.org/10.1109/ICCMC51019.2021.9418488>
- He, X., Liu, X., Zuo, F., Shi, H., & Jing, J. (2023). Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. *Seminars in Cancer Biology*, 88, 187–200. <https://doi.org/10.1016/J.SEMCANCER.2022.12.009>

- Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, 1(5), 1–14. <https://doi.org/10.1007/S42979-020-00305-W/METRICS>
- Jaiswal, G., Rani, R., Mangotra, H., & Sharma, A. (2023). Integration of hyperspectral imaging and autoencoders: Benefits, applications, hyperparameter tuning and challenges. *Computer Science Review*, 50, e100584. <https://doi.org/10.1016/J.COSREV.2023.100584>
- Kabir, M. F., Chen, T., & Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, 3, e100125. <https://doi.org/10.1016/J.HEALTH.2022.100125>
- Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020). Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. <https://doi.org/10.1109/ICCCNT49239.2020.9225451>
- Karuppasamy, A. D., Abdesselam, A., zidoum, H., Hedjam, R., & Al-Bahri, M. (2024). Combining a forward supervised filter learning with a sparse NMF for breast cancer histopathological image classification. *Intelligence-Based Medicine*, 10, e100174. <https://doi.org/10.1016/J.IBMED.2024.100174>
- Li, X., Chen, X., & Rezaeiapanah, A. (2023). Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. *Journal of Cancer Research and Clinical Oncology*, 149(10), 7609–7627. <https://doi.org/10.1007/S00432-023-04699-X>
- Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1), 91–99. <https://doi.org/10.1016/J.GLTP.2022.04.020>
- Mishra, V., & Rath, S. K. (2021). Detection of breast cancer tumours based on feature reduction and classification of thermograms. *Quantitative InfraRed Thermography Journal*, 18(5), 300–313. <https://doi.org/10.1080/17686733.2020.1768497>
- Mishra, V., Rath, S., & Rath, S. K. (2023). *Feature Analysis for Detection of Breast Cancer Thermograms Using Dimensionality Reduction Techniques* (pp. 311–321). https://doi.org/10.1007/978-981-19-9090-8_27
- Naji, M. A., Filali, S. El, Aarika, K., Benlahmar, E. H., Abdelouahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*, 191, 487–492. <https://doi.org/10.1016/j.procs.2021.07.062>
- Paul, A., Joe, M., Brian, M., Hussein, M., & Rosette, K. (2024). Exploring Dimensionality Reduction Techniques for Improved Breast Cancer Diagnosis. *International Journal of Research and Scientific Innovation*, 11(5), 808–824. <https://ideas.repec.org/a/bjc/journal/v11y2024i5p808-824.html>
- Pérez-Núñez, J. R., Rodríguez, C., Vásquez-Serpa, L. J., & Navarro, C. (2024). The Challenge of Deep Learning for the Prevention and Automatic Diagnosis of Breast Cancer: A Systematic Review. *Diagnostics*, 14(24), e2896. <https://doi.org/10.3390/DIAGNOSTICS14242896/S1>
- Rimi, I., Mondal, Md. N. I., & Oishy, J. (2022). Prediction Approach of Breast Cancer using Dimensionality Reduction and Outlier Detection. *2022 4th International Conference on Electrical, Computer and Telecommunication Engineering (ICECTE)*, 1–4. <https://doi.org/10.1109/ICECTE57896.2022.10114477>

- Sengar, P. P., Gaikwad, M. J., & Nagdive, A. S. (2020). Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 796–801. <https://doi.org/10.1109/ICSSIT48917.2020.9214267>
- Tai, C. en A., Gunraj, H., Hodzic, N., Flanagan, N., Sabri, A., & Wong, A. (2024). Enhancing Clinical Support for Breast Cancer with Deep Learning Models Using Synthetic Correlated Diffusion Imaging. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14313 LNCS, 83–93. https://doi.org/10.1007/978-3-031-47076-9_9/COVER
- Dwarakanath, G. V., & Bhairagond, M. (2023). Breast Cancer Detection and Prediction using Image Processing and ML. *International Journal for Science Technology and Engineering*, 11(8), 1902–1911. <https://doi.org/10.22214/IJRASET.2023.55503>
- Yu, S. M., Young, C. Y. M., Chan, Y. H., Chan, Y. S., Tsoi, C., Choi, M. N. Y., Chan, T. H., Leung, J., Chu, W. C. W., Hung, E. H. Y., & Chau, H. H. L. (2024). Artificial intelligence–based computer-aided diagnosis for breast cancer detection on digital mammography in Hong Kong. *Hong Kong Medical Journal*, 30(6), 468–477. <https://doi.org/10.12809/HKMJ2310920>