Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 8 No. 2 (2026)
https://doi.org/10.35877/454RI.asci4755

Robust MRCD-PCA in Machine Learning for Breast Cancer Classification

Maharani Abu Bakar, Sharifah Sakinah Syed Abd Mutalib*, Danang Adi Pratama, & Muhamad Safiih Lola

Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

Abstract

Breast cancer classification plays a critical role in early diagnosis and clinical decision-making. However, medical datasets are often high-dimensional and contaminated with outliers, which can degrade classification performance. While Principal Component Analysis (PCA) is commonly used for dimensionality reduction, its sensitivity to outliers limits its effectiveness in medical data analysis. To address this limitation, this study proposes a robust PCA (RPCA) approach based on the Minimum Regularized Determinant (MRCD) estimator for dimensionality reduction and named the proposed method as Robust MRCD-PCA (RMPCA). The classification using proposed RMPCA is evaluated using Support Vector Machine (SVM), Artificial Neural Network (ANN), and K-means classifiers, and compared against baseline and PCA-based models. A total of nine classification models is examined using a breast cancer dataset, with performance assessed using accuracy, precision, recall, and F1-score. Experimental results demonstrate that RMPCA achieves more consistent and reliable classification performance, particularly when combined with supervised classifiers such as SVM, outperforming both baseline and PCA-based approaches. These findings highlight the importance of robust dimensionality reduction as an effective preprocessing strategy for improving machine learning-based breast cancer classification. The proposed method is also found to be promising and applicable method for classification. The findings support more reliable clinical decision-making, thereby contributing to improved healthcare outcomes and reduced socio-cultural burden associated with misdiagnosis.

Keywords: Type your keywords here; between 3 and 6; separated by semicolons.

Received: 15 March 2026 Revised: 5 June 2026 Published: 31 August 2026

1. Introduction

Over 2.1 million new cases of breast cancer are reported each year, making it a major global health concern (Rahman et al., 2025). The classification of breast cancer is a critical endeavour in medical data analysis, as it aids in facilitating early identification and informing therapeutic decision-making. Developments in machine learning have facilitated the creation of automated classification models that aid in differentiating between benign and malignant instances. However, breast cancer datasets frequently exhibit high dimensionality, feature redundancy, and outliers due to measurement mistakes and biological variability (Rahman et al., 2025; Yaqoob et al., 2025). These issues can negatively impact the performance and stability of machine learning classifiers, especially when implemented directly in the original feature space.

Dimensionality reduction techniques are frequently utilised to mitigate these challenges by converting high-dimensional data into a more concise and practical representation. Principal Component Analysis (PCA) is widely used for this purpose, due to its ability to retain maximal diversity and minimise information loss (Mohammed et al., 2023) . Studies by Adiwijaya et al. (2018) and Kartini et al. (2025) have demonstrated that PCA improves classification performance when employed in conjunction with machine learning models, and it has been successfully applied in a variety of biological classification applications, including cancer data processing. However, the sensitivity of classical PCA to outliers and deviations from normality restricts its usefulness in real-world medical datasets that frequently contain outliers.

* Corresponding author.
E-mail address: s.sakinah@umt.edu.my

Robust PCA has been proposed as a solution to this problem, utilizing robust estimators to lessen the impact of outliers during feature extraction (Midi et al., 2025). Robust Principal Component Analysis (RPCA) exhibits enhanced stability and reliability in managing abnormal data compared to classical PCA (He et al., 2023). Recent studies indicate that robust feature extraction can improve classification accuracy by yielding clearer and more distinctive feature representations, especially in high-dimensional biological contexts such as in Alessandrini et al. (2022) and He et al. (2023). The Minimum Regularized Covariance Determinant (MRCD) estimator, an established robust estimator, is incorporated into the PCA in this study to overcome the problem of high-dimensional data. Studies done by Boudt et al. (2020) and Zahariah and Midi (2023) have demonstrated the MRCD's efficacy as a reliable estimator for high-dimensional data.

Motivated by these findings, the objective of this study is to propose robust PCA for dimensionality reduction using the MRCD estimator, called Robust MRCD-PCA (RMPCA). Machine learning models are subsequently employed to classify the breast cancer dataset. Although Robust MRCD-PCA (RMPCA) has shown promise in handling high-dimensional and contaminated data, its integration within end-to-end breast cancer classification frameworks remains limited, and its impact on downstream machine learning classification performance has not been systematically evaluated. To address this gap, this study compares the proposed method with baseline models (K-means, Artificial Neural Network (ANN), and Support Vector Machine (SVM)) and classical PCA-based models to assess how robust dimensionality reduction enhances classification performance.

The remainder of this paper is organized as follows. Section 2 describes the dataset, the proposed Robust MRCD-PCA framework, and the machine learning models employed in this study. Section 3 presents and discusses the classification results obtained from the proposed method in comparison with baseline and classical PCA-based models. Finally, Section 4 concludes the paper by summarizing the main findings and outlining potential directions for future research.

2. Methods

This section presents the experimental results for one benchmark dataset, the Breast Cancer Wisconsin Diagnostic, obtained from the UC Irvine Machine Learning Repository. The use of a benchmark dataset is important in this study to ensure a fair and consistent evaluation of the proposed method, allowing comparison with existing approaches reported in the literature such as in Cakmak and Pacal (2025) and Li (2023). Details about the dataset used are shown in Table 1. The machine learning models used in this study are K-means, Artificial Neural Network (ANN), and Support Vector Machine (SVM). These classifiers are implemented using the original dimension as well as dimension obtained through classical PCA and the Robust MRCD-PCA (RMPCA). All computational experiments were conducted using R and Python.

Table 1. Details of the dataset.

Dataset	Sample size	Dimension	Group
Breast Cancer Wisconsin Diagnostic	569	30	Benign (357 observations) Malignant (212 observations)

Figure 1 illustrates the overall workflow of the proposed study. The process begins with obtaining the breast cancer dataset from the UC Irvine Machine Learning Repository. The two primary scenarios of the workflow are classification with dimensionality reduction and classification utilising the original dimensions. In the second scenario, two dimensionality reduction methods are used which are the suggested Robust MRCD-PCA (RMPCA) and classical PCA. Subsequently, the dataset is divided into training and testing data, with 80% of the data allocated for model training and 20% reserved for performance evaluation.

Three baseline classifiers (K-means, ANN, SVM) are used in the model's implementation stage. The same classifiers are applied to the reduced dimensions to create PCA-based and Robust MRCD-PCA (RMPCA) models in addition to the baseline models. This makes it possible to systematically assess classification performance across various dimensions. Finally, the performance of all models is evaluated and compared using accuracy, precision, recall, and F1-score. The overall workflow is illustrated in Figure 1 and summarized in Algorithm 1.

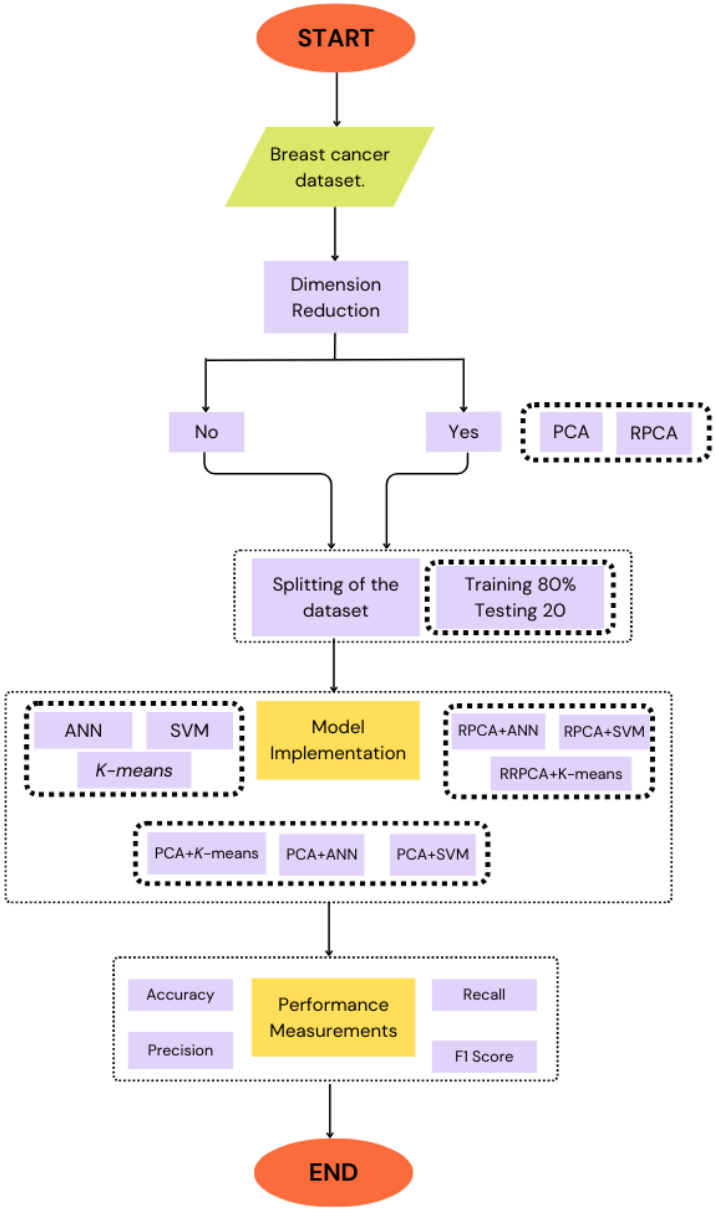


Figure 1. Flow chart of the study

Tabel 2. Breast Cancer Classification Framework

Algoritm 1. Breast Cancer Classification Framework

Require: Breast cancer dataset D from UC Irvine Machine Learning Repository
Ensure: Performance comparison across Baseline, PCA-based, and RMPCA-based models

1: Acquire dataset D
2: Perform data cleaning and preprocessing on D and get D_clean

3: Apply PCA to D_clean and get D_PCA
4: Apply RMPCA to D_clean and get D_RMPCA

5: Split D_clean, D_PCA, and D_RMPCA into Training (80%) and Testing (20%)

```
Define Models {Baseline, PCA-based, RMPCA-based}
Define Classifiers {K-means, ANN, SVM}

For each M_i in Models do
    If M_i = Baseline then
    For each C_i in Classifiers do
                Train C_i using Training D_clean
                Predict on Testing D_clean
                Compute Accuracy, Precision, Recall, F1-score
    End For

    Else if M_i = PCA-based then
    For each C_i in Classifiers do
                Train C using Training D_PCA
                Predict on Testing D_PCA
                Compute Accuracy, Precision, Recall, F1-score
    End For

    Else // RMPCA-based
    For each C_i in Classifiers do
                Train C_i using Training D_RMPCA
                Predict on Testing D_RMPCA
                Compute Accuracy, Precision, Recall, F1-score
    End For
    End If

Compare performance across all model categories
```

2.1. Dimensionality Reduction Techniques

In this section, two dimensionality reduction techniques are described, namely PCA and the proposed robust MRCD-PCA (RMPCA). These techniques are used to reduce the dimensionality of data and extract meaningful features before classification. While PCA is effective under uncontaminated data, robust PCA is incorporated to reduce the influence of outliers, as the robust method is insensitive to outliers.

2.1.1. Principal Component Analysis (PCA)

The procedure involved in PCA is explained in this section. PCA procedures involve data standardisation, calculation of covariance matrices, eigen decomposition, and projection onto selected principal components.

Step 1: Compute the sample mean, $\bar{x}_j$ and covariance matrix, $\mathbf{S}$ from the dataset, $\mathbf{X}$.

Step 2: Standardise the data using Eq. (1) so that each variable contributes equally to the analysis.

$$z_j = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad i = 1,2,\ldots,n \quad j = 1,2,\ldots,p. \tag{1}$$

Step 3: Obtain eigenvectors and eigenvalues of the sample covariance matrix $\mathbf{S}$ and let the eigenvalue eigenvector pairs $(\lambda_1,\mathbf{e}_1),(\lambda_2,\mathbf{e}_2),\ldots,(\lambda_p,\mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Step 4: The eigenvectors are arranged in descending order based on the corresponding eigenvalues, and the principal components are selected based on the proportion of variance explained, typically retaining components that account for 80% or 90% of the total variance (Hubert et al., 2005).

Step 5: Finally, the data are projected onto the selected principal component axes to obtain the reduced representation.

2.1.2. Proposed Robust MRCD-PCA (RMPCA)

The formulation of an enhanced robust PCA is achieved by incorporating principal component analysis with the Minimum Regularised Covariance Determinant (MRCD) estimator.

Step 1: Obtain the MRCD-based robust sample means, $\bar{x}_{j,MRCD}$ and covariance matrix, $\mathbf{S}_{MRCD}$ is obtained.

Step 2: Standardise the data using Eq. (2) so that each variable contributes equally to the analysis.

$$z_j = \frac{x_{ij} - \bar{x}_{j,MRCD}}{s_{j,MRCD}} \qquad i = 1,2,\ldots,n \quad j = 1,2,\ldots,p. \tag{2}$$

Step 3: Find the eigenvectors and eigenvalues of the sample covariance matrix, $\mathbf{S}_{MRCD}$ and let the eigenvalue-eigenvector pairs $(\lambda_1,\mathbf{e}_1),(\lambda_2,\mathbf{e}_2),\ldots,(\lambda_p,\mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

$$\left(\mathbf{S}_{MRCD} - \lambda_j \mathbf{I}\right)U_j = 0 \qquad j = 1,2,\ldots,p \tag{3}$$

$$\det\left(\mathbf{S}_{MRCD} - \lambda \mathbf{I}\right) = 0 \tag{4}$$

Step 4: The eigenvectors are ordered in descending order based on the corresponding eigenvalues, and the principal components are selected based on the proportion of variance explained, typically retaining those that account for 80% or 90% of the total (Hubert et al., 2005).

Step 5: In the final step, the data are projected onto the selected principal component axes to obtain the reduced representation.

2.2. Machine Learning Models

In this section, three machine learning models are explained, namely K-means, Artificial Neural Network (ANN), and Support Vector Machine (SVM). These models are used to classify the dataset into benign or malignant.

2.2.1. K-means

K-means is an unsupervised learning algorithm widely used for partitioning data into *K* distinct clusters based on similarity (Gholami et al., 2023; Zhu et al., 2021). The objective of the algorithm is to minimise the *within-cluster sum of squared distances* between data points and their corresponding cluster centroids. Let $\{x_1, x_2, \ldots, x_n\}$ represent a set of input data points in a multidimensional feature space.

The *K*-means algorithm begins by specifying the number of clusters *K* and initialising *K* centroids, typically selected randomly from the dataset. Each data point is then assigned to the cluster whose centroid is closest, based on the Euclidean distance measure. After the assignment step, new centroids are computed as the mean of all data points belonging to each cluster (Aggarwal, 2015). The resulting clusters are analysed to identify inherent patterns in the data. When true class labels are available, clustering results may be compared against ground truth using external evaluation measures.

Let $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$, be a dataset consisting of *n* data points in a *d*-dimensional feature space. The goal of *K*-means is to partition the dataset into *K* disjoint clusters $C = \{C_1, C_2, \ldots, C_K\}$, by minimising the objective function in Eq. (5).

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \tag{5}$$

where μ_k denotes the centroid of the cluster C_k, and $\|\cdot\|$ represents the Euclidean norm. The centroid μ_k is defined as the mean of all data points assigned to a cluster C_k, as in Eq. (6).

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i \in C_k} x_i \tag{6}$$

The *K*-means clustering process consists of the following sequential steps (Shrifan et al., 2022).

Step 1: Specify the number of clusters *K* before execution. The choice of *K* may be guided by domain knowledge or exploratory techniques such as the elbow method (Aggarwal, 2015).

Step 2: Initial centroids $\{\mu_1, \mu_2, \ldots, \mu_k\}$ are selected, commonly by randomly choosing *K* data points from the dataset. These centroids serve as the initial cluster centers.

Step 3: Each data point x_i is assigned to the cluster whose centroid is closest according to the Euclidean distance criterion as given in Eq. (7).

$$x_i \in C_k \text{ if } \|x_i - \mu_k\| \leq \|x_i - \mu_j\|, \forall j \neq k \tag{7}$$

This step partitions the dataset into *K* clusters based on proximity to centroids.

Step 4: For each cluster C_k, a new centroid is computed as the mean of the data points assigned to that cluster using Eq. (8).

$$\mu_k^{(new)} = \frac{1}{|C_k|}\sum_{x_i \in C_k} x_i \tag{8}$$

This update step relocates the centroid to the center of its assigned data points.

Step 5: Steps 3 and 4 are repeated iteratively until convergence is achieved, which occurs when cluster assignments remain unchanged, centroid updates fall below a predefined threshold, or a maximum number of iterations is reached. Upon convergence, the final cluster assignments and centroids are obtained.

2.2.2. Artificial Neural Network

An Artificial Neural Network (ANN) is a supervised machine learning approach designed to model complex and nonlinear relationships between input features and output targets. ANNs have gained significant attention in recent years due to their strong predictive capability and flexibility across various classification and prediction tasks. Recent studies highlight the effectiveness of ANN architectures, particularly Multilayer Perceptrons (MLPs), in learning nonlinear decision boundaries from data (A. R. Khan et al., 2025). In this study, a Multilayer Perceptron (MLP) architecture is employed, consisting of an input layer, one or more hidden layers, and an output layer. Fully connected neurons and nonlinear activation functions enable the network to approximate complex functions (Aziz et al., 2022).

The ANN architecture employed in this study follows a standard Multilayer Perceptron (MLP) structure consisting of an input layer, one hidden layer, and an output layer. Let $x = (x_1, x_2, \ldots, x_d)^T$, denote an input feature vector of dimension d. The input layer contains d neurons corresponding to the input features, the hidden layer consists of h neurons, and the output layer contains $c$neurons representing the class labels.

Each neuron in a layer is fully connected to neurons in the subsequent layer via weighted connections. Information flows forward through the network in a process known as forward propagation, where neuron activations are computed sequentially from the input layer to the output layer. For a hidden neuron j, the net input is computed as in Eq. (9).

$$z_j = \sum_{i=1}^{d} w_{ij} x_i + b_j \tag{9}$$

where w_{ij} represents the weight connecting input neuron i to hidden neuron j, and b_j is a bias term. The neuron output is obtained by applying a nonlinear activation function $f(.)$ in Eq. (10).

$$a_j = f\left(z_j\right) \tag{10}$$

In recent ANN implementations, the Rectified Linear Unit (ReLU) activation function is commonly used in hidden layers due to its computational efficiency and effectiveness in mitigating vanishing gradient problems (Mienye & Swart, 2024). At the output layer, Sigmoid or Softmax activation functions are employed depending on whether the classification task is binary or multi-class.

The learning objective of the ANN is to minimise a loss function that measures the discrepancy between predicted outputs and true class labels. For classification tasks, cross-entropy loss is widely adopted and is defined as in Eq. (11).

$$L = \sum_{k=1}^{C} y_k \log\left(\hat{y}_k\right) \tag{11}$$

where y_k denotes the true class label and $\hat{y}_k$ represents the predicted probability for class k. Cross-entropy loss provides stable gradients and efficient convergence in modern neural network training (Ulku & Akagündüz, 2022).

2.2.3. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification and regression tasks. SVM constructs an optimal decision boundary that separates data points of different classes with the maximum possible margin, thereby improving generalisation performance on unseen data. Recent studies confirm that SVM remains a strong baseline classifier, particularly for structured and medium-sized datasets, due to its solid theoretical foundation and robustness to overfitting (Nandi et al., 2023). The decision boundary in SVM is defined by a subset of critical training samples known as support vectors, which lie closest to the separating hyperplane.

Let $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}, x_i \in \square^{d}, y_i \in \{-1,+1\}$ denote a labeled training dataset. The objective of a linear SVM is to find a hyperplane $w^T x + b = 0$ that separates the two classes with maximum margin. This is achieved by solving the following optimisation problem given in Eq. (12).

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{12}$$

subject to $y_i\left(w^T x_i + b\right) \geq 1, i = 1,2,\ldots,n.$ Maximising the margin corresponds to minimising $\|w\|^2$, which leads to better generalization (Chhajer et al., 2022). For real-world data that may not be perfectly separable, SVM introduces slack variables ξ_i to allow misclassification. The optimisation problem becomes,

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{13}$$

subject to $y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0.$

When data are not linearly separable, SVM employs kernel functions to implicitly map input data into a higher-dimensional feature space where linear separation becomes possible. This approach is known as the kernel trick. Common kernel functions include linear kernel, polynomial kernel and radial basis function (RBF) kernel. Among these, the **RBF kernel** is widely used due to its ability to model nonlinear decision boundaries given in Eq. (14).

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right) \tag{14}$$

where γ determines the influence of individual training samples. Recent literature emphasizes careful tuning of C and γ to achieve optimal performance.

Once trained, the SVM classifier assigns a class label to a new data point x using the decision function in Eq. (15).

$$f(x) = sign\left(\sum_{i \in SV} \alpha_i y_i K(x_i, x_j) + b\right) \tag{15}$$

where SV denotes the set of support vectors and α_i are Lagrange multipliers are obtained during training. Only support vectors contribute to the decision function, making SVM efficient during prediction (Chhajer et al., 2022). The regularisation parameter C controls the trade-off between maximising the margin and minimising classification error. A larger C places greater emphasis on correct classification, while a smaller C promotes a wider margin.

2.3. Model Configurations

In this study, various feature extraction methods were combined with machine learning algorithms to create nine classification models. Based on the feature representation that was employed, the models were divided into three groups. The baseline models, which include K-means, Support Vector Machine (SVM), and Artificial Neural Network (ANN), were created using the original dimensions. Principal Component Analysis (PCA) and Robust MRCD-PCA (RMPCA) were used for dimensionality reduction before classification, producing models based on PCA and RMPCA. All models were evaluated using the same dataset, experimental settings, and performance metrics to ensure a fair comparison. Table 2 summarizes the classification models and dimension extraction techniques considered in this study.

Table 2. Summary of model configurations.

Category	Description
Baseline models	K-means, ANN, SVM
PCA-based models	PCA+K-means, PCA+K-ANN, PCA+SVM
Robust MRCD-PCA models	RMPCA+K-means, RMPCA+ANN, RMPCA+SVM

2.4. Performance Evaluation

The performance of the nine models was evaluated using accuracy, precision, recall, and F1-score, which are widely used metrics in classification problems. Since this study is based on binary classification, Eqs. (16) – (19) can be used to evaluate the performance of the model (Asif et al., 2025; Singh et al., 2022).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{17}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{18}$$

$$\text{F1 Score} = \frac{2*(\text{Recall*Precision})}{(\text{Recall+Precision})} \tag{19}$$

3. Results and Discussion

3.1. Comparative Performance Analysis

In this section, the results of the comparative performance analysis of nine models for breast cancer classification are presented and discussed. Table 3 reports the accuracy, precision, recall, and F1-score used to evaluate the performance of these models. The highest values are bold to indicate the better model.

Table 3 shows that SVM outperforms ANN and K-means among the baseline models, indicating that supervised and margin-based classifiers are well-suited for breast cancer classification tasks, as they can maximize class separation. In contrast, the comparatively weaker performance of K-means highlights the limitations of this unsupervised clustering method. With a lack of label information, K-means struggles to learn complex decision boundaries in such settings (An et al., 2023; I. K. Khan et al., 2024).

From an overall perspective, incorporating dimensionality reduction techniques improves the classification performance of machine learning models for breast cancer classification, which is consistent with findings reported in previous studies of Koca and Aktepe (2024) and Athisayamani et al. (2025). This improvement is observed in both PCA and RMPCA models, which generally achieve higher performance compared to baseline approaches. However, the performance gains of PCA-based models are not consistent across all classifiers, as the baseline SVM slightly outperforms its PCA-based counterpart. This result reflects a known limitation of classical PCA, which is sensitive to outliers and deviations from normality, factors that commonly arise in medical datasets due to biological variability and measurement errors (Chen et al., 2020; Hubert et al., 2005).

The most dependable and consistent performance is shown by RMPCA-based models, especially when paired with supervised classifiers like SVM and ANN. The better performance of RMPCA-based models implies that robust dimensionality reduction successfully reduces the impact of outliers. Robust dimensionality reduction improves the SVM by offering cleaner and more informative feature representations. Enhanced accuracy achieved through the proposed method may lead to more reliable diagnoses, reducing unnecessary psychological burden and supporting better patient management.

Table 3. Classification performance of machine learning models for test data.

Model		Accuracy (%)	Precision	Recall	F1-score
Baseline models	K-means	89.47	0.901	0.872	0.883
	ANN	96.49	0.967	0.957	0.962
	SVM	**98.25**	**0.986**	**0.976**	**0.981**
PCA-based models	PCA + K-means	90.35	0.910	0.880	0.890
	PCA + ANN	97.37	0.980	0.960	0.970
	PCA + SVM	97.37	0.970	0.970	0.970
Robust MRCD-PCA models	RMPCA + K-means	91.23	0.920	0.890	0.900
	RMPCA + ANN	97.37	0.970	0.970	0.970
	RMPCA + SVM	**98.25**	**0.990**	**0.980**	**0.980**

Figure 2 reflects the findings shown in Table 3, providing a visual representation of the classification performance across different models. As illustrated in Figures 2(a)–(c), the baseline model still shows significant mixing of the two groups, indicating that the separation between classes is still unclear and poorly defined. This suggests that the model without dimensionality reduction is less effective in handling the complexity and potential outliers in the data.

In contrast, a clearer separation between the two groups can be observed when the PCA and RMPCA-based models are used. The combination of dimensionality reduction techniques improves the data structure, allowing the model to better distinguish between classes. Notably, the RMPCA-based model shows a more consistent and clear clustering pattern, indicating increased robustness against outliers.

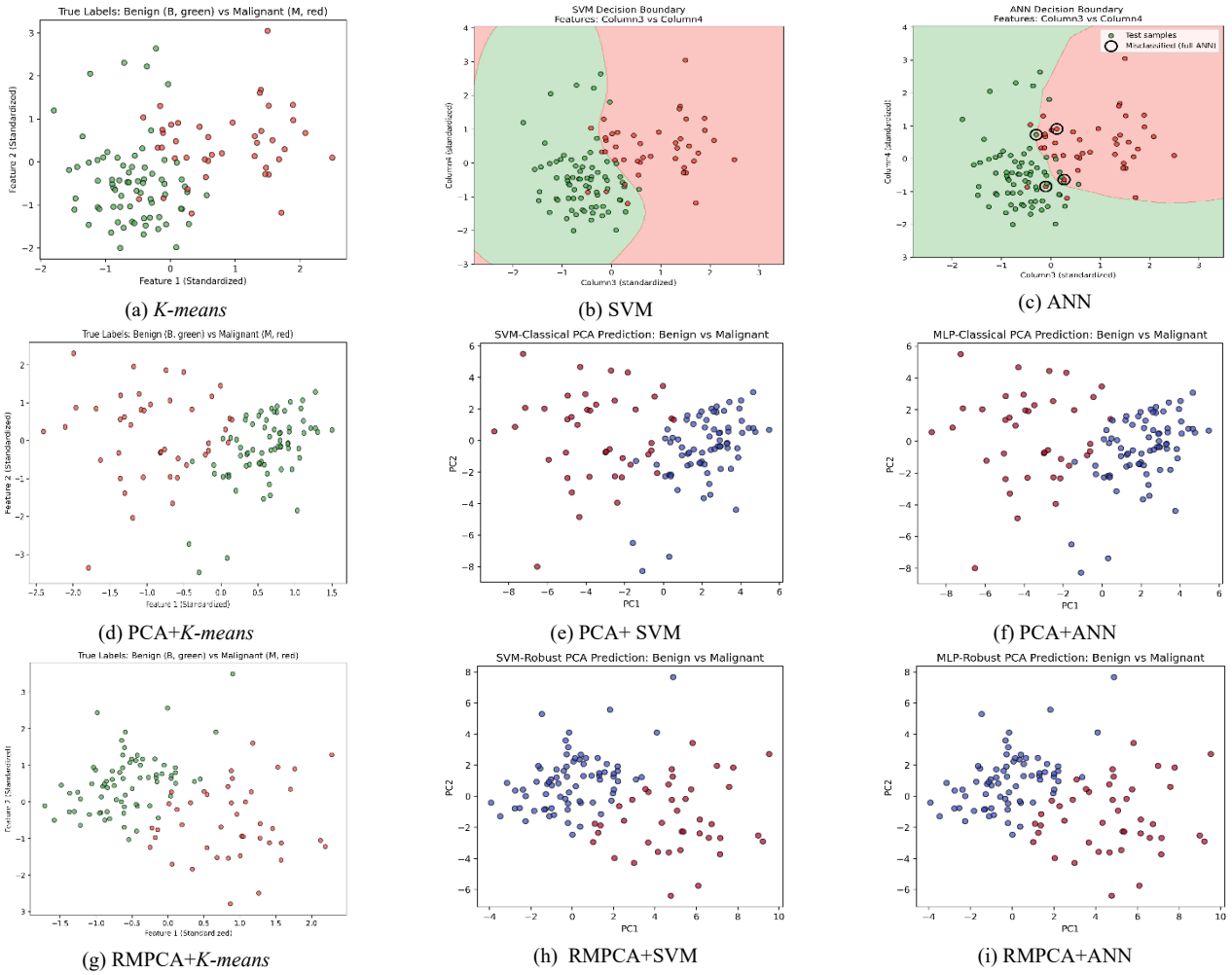
Overall, these findings demonstrate that combining RMPCA with machine learning models leads to more stable and reliable classification performance. The improved separation highlights the effectiveness of the proposed approach in capturing the underlying data structure. Therefore, RMPCA serves as a valuable preprocessing technique for breast cancer classification, especially in high-dimensional settings where traditional methods may contain outliers.

3.2. Performance Analysis by Class

In this section, the class-wise performance of the two best-performing models identified in Section 3.1 is analysed to provide a more detailed understanding of their classification behaviours. Specifically, the baseline SVM and RMPCA + SVM models are evaluated in distinguishing between benign and malignant breast cancer cases, highlighting differences in predictive performance across classes. This class-level analysis is important, as overall accuracy may not fully reflect performance for each category. By examining precision, recall, and F1-score, a more comprehensive evaluation is achieved. The detailed classification results for both models are presented in Table 4.

Table 4. Performance analysis for SVM and RMPCA+SVM by class

Model	Class	Precision	Recall	F1-score
SVM	Benign	**0.973**	**1.000**	0.986
	Malignant	**1.000**	**0.952**	0.976
RMPCA + SVM	Benign	0.970	**1.000**	**0.990**
	Malignant	**1.000**	0.950	**0.980**



(a) *K-means* (b) SVM (c) ANN

(d) PCA+*K-means* (e) PCA+ SVM (f) PCA+ANN

(g) RMPCA+*K-means* (h) RMPCA+SVM (i) RMPCA+ANN

Figure 2. Comparison of classification visualisations for *K*-means, ANN, and SVM classifiers under original, PCA-based, and Robust PCA-based models.

From Table 4, benign samples are accurately identified by the baseline SVM model, which achieves a perfect recall score. The high precision and F1-score further indicate consistent classification performance for this class. For malignant cases, the baseline SVM model attains 100% precision, indicating that all samples predicted as malignant are correctly classified. However, the slightly lower recall for malignant cases suggests that some malignant samples remain undetected.

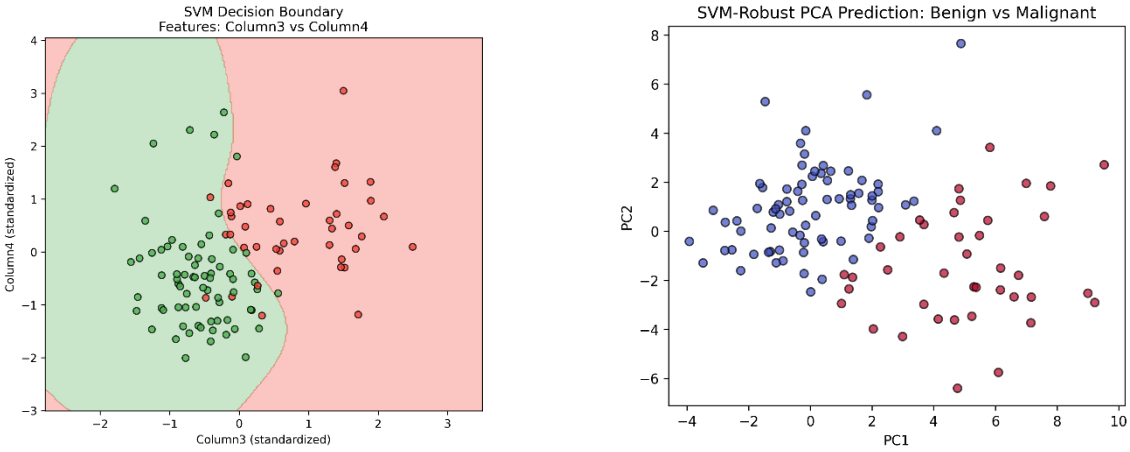
Comparable and marginally better class-wise performance is demonstrated by the RMPCA + SVM model. This model maintains perfect recall for the benign class, similar to the baseline SVM, while achieving slightly higher precision and F1-score values, indicating more consistent classification. This improvement suggests that incorporating RMPCA enhances dimensionality reduction by reducing the influence of errors or outliers, which are common in real-world medical datasets (Alessandrini et al., 2022).

Overall, both models demonstrate excellent discriminative ability. As shown in Table 4, the RMPCA + SVM model demonstrates a slight improvement over the baseline SVM, particularly in terms of F1-score for both benign and malignant classes. Although the improvement is marginal, it indicates that the incorporation of RMPCA enhances the stability and robustness of the classification model, leading to more balanced performance.

In Figure 3(a), the SVM trained on the original dimensions shows noticeable overlap between benign and malignant samples, indicating less distinct class separation in the original dimensions. In contrast, Figure 3(b) demonstrates that applying RMPCA prior to SVM classification results in a clearer separation between the two classes, with reduced overlap and more compact class distributions. This visual improvement supports the quantitative results in Tables 3 and 4, confirming that robust dimensionality reduction enhances feature representation and contributes to more stable and reliable classification performance in breast cancer data.

The results demonstrate that the proposed Robust MRCD-PCA (RMPCA) improves classification performance compared to the classical PCA approach. Beyond technical performance, this improvement has important socio-cultural implications. In clinical settings, reducing false positives is crucial, as misdiagnosis may lead to unnecessary psychological stress, anxiety, and social burden among patients (Samuel et al., 2026). Therefore, the robustness of the proposed method may contribute to more reliable diagnostic outcomes and improved patient well-being.

The analysis in this study is conducted using a benchmark dataset to ensure a standardized evaluation of the proposed method. The primary objective is to assess the effectiveness of the Robust MRCD-PCA (RMPCA) in improving classification performance compared to the classical PCA approach. The results demonstrate that the robust method provides more reliable classification outcomes, indicating its ability to handle potential outliers.



(a) SVM

(b) RMPCA+SVM

Figure 3. Visualisation of SVM using (a) original features and (b) Robust PCA.

4. Conclusion

This study compared baseline machine learning models with their PCA and RMPCA-based models to assess the effect of dimensionality reduction on breast cancer classification using breast cancer Wisconsin diagnostic dataset, obtained from the UC Irvine Machine Learning Repository. The results demonstrate that dimensionality reduction generally

improves classification performance compared to using the original dimensions. As evidenced in Tables 3 and 4, RMPCA-based models exhibit more consistent and reliable performance across different classifiers, highlighting the advantage of robust dimensionality reduction in handling contaminated medical datasets.

Further analysis indicates that SVM-based classifiers achieve excellent classification performance among the evaluated machine learning models, demonstrating the effectiveness of margin-based learning for structured medical data. While incorporating PCA improves the performance of certain models, these improvements are not consistent across all classifiers, likely due to PCA's sensitivity to outliers. In contrast, RMPCA mitigates the influence of outliers, resulting in more stable and consistent classification performance, particularly when combined with supervised classifiers such as SVM. In addition to improving classification performance, the proposed method has the potential to support more reliable clinical decision-making, which may reduce the socio-cultural burden associated with misdiagnosis.

Overall, the findings show that RMPCA serves as an effective preprocessing step for improving the robustness and accuracy of machine learning-based breast cancer classification. Beyond demonstrating performance gains, this study provides a systematic comparison of baseline, PCA and RMPCA-based classifiers, offering insight into the role of robust dimensionality reduction in medical classification tasks. Future research may extend this work by evaluating additional medical datasets, exploring alternative robust dimensionality reduction techniques, and integrating advanced learning models to further enhance classification performance and generalizability.

Acknowledgements: This research was supported by the Ministry of Higher Education (MOHE) Malaysia through the Fundamental Research Grant Scheme (Early Career Researcher) (FRGS-EC/1/2024/STG06/UMT/02/7).

Funding Statement: Ministry of Higher Education (MOHE) Malaysia through the Fundamental Research Grant Scheme (Early Career Researcher) (FRGS-EC/1/2024/STG06/UMT/02/7).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Contribution: Abu Bakar: Conceptualization, Methodology, Software, Visualization, Writing-Reviewing and Editing. Abd Mutalib: Methodology, Analysis, Methodology, Writing-Original Draft. Pratama: Writing-Reviewing and Editing. Lola: Writing-Reviewing and Editing.

References

Adiwijaya, Wisesty, U. N., Lisnawati, E., Aditsania, A., & Kusumo, D. S. (2018). Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. *Journal of Computer Science*, *14*(11), 1521–1530. https://doi.org/10.3844/jcssp.2018.1521.1530

Aggarwal, C. C. (2015). *Machine Learning for Data Mining*. Springer.

Alessandrini, M., Biagetti, G., Crippa, P., Falaschetti, L., Luzzi, S., & Turchetti, C. (2022). *Robust-PCA and LSTM Recurrent Neural Network*. 1–18.

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, *23*(9), 4178. https://doi.org/10.3390/s23094178

Asif, S., Wenhui, Y., Ur-Rehman, S., Ul-ain, Q., Amjad, K., Yueyang, Y., Jinhai, S., & Awais, M. (2025). Advancements and prospects of machine learning in medical diagnostics: Unveiling the future of diagnostic precision. In *Archives of Computational Methods in Engineering* (Vol. 32, Issue 2). https://doi.org/10.1007/s11831-024-10148-w

Athisayamani, S., Tamilazhagan, S., Singh, A. R., Hwang, J., & Joshi, G. P. (2025). A novel double machine learning approach for detecting early breast cancer using advanced feature selection and dimensionality reduction techniques. *Scientific Reports*, *15*, 22971.

Aziz, M. A. A., Ismail, S., & Allias, N. (2022). Deep learning in face recognition for attendance system: An exploratory study. *Journal of Computing Research and Innovation*, *7*(2), 74–81. https://doi.org/10.24191/jcrinn.v7i2.288

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, *30*(1), 113–128. https://doi.org/10.1007/s11222-019-09869-x

Cakmak, Y., & Pacal, I. (2025). Enhancing breast cancer diagnosis: A comparative evaluation of machine learning algorithms using the Wisconsin dataset. *Journal of Operations Intelligence*, *3*(1), 175–196. https://doi.org/10.31181/jopi31202539

Chen, X., Zhang, B., Wang, T., Bonni, A., & Zhao, G. (2020). Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinformatics*, *21*, 1–20. https://doi.org/10.1186/s12859-020-03608-0

Chhajer, P., Shah, M., & Kshirsagar, A. (2022). The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction. *Decision Analytics Journal*, *2*, 100015. https://doi.org/10.1016/j.dajour.2021.100015

Gholami, M., Mouhoub, M., & Sadaoui, S. (2023). Biogeography-based optimization for feature selection. *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*, *36*. https://doi.org/10.32473/flairs.36.133230

He, L., Yang, Y., & Zhang, B. (2023). Robust PCA for high-dimensional data based on characteristic transformation. *Australian and New Zealand Journal of Statistics*, *65*(2), 127–151. https://doi.org/10.1111/anzs.12385

Hubert, M., Rousseeuw, P. J., & Branden, K. Vanden. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

Kartini, D., Amin Badali, R., Muliadi, Nugrahadi, D. T., Indriani, F., & Wahyu Saputro, S. (2025). Dimensionality reduction using principal component analysis and feature selection using genetic algorithm with support vector machine for microarray data classification. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(1), 154–166. https://doi.org/10.35882/ijeeemi.v7i1.49

Khan, A. R., Razak, M. S. B. A., Yusuf, B. B., Shafri, H. Z. B. M., & Mohamad, N. B. (2025). Harnessing artificial neural networks for coastal erosion prediction: A systematic review. *Marine Policy*, *178*, 106704. https://doi.org/10.1016/j.marpol.2025.106704

Khan, I. K., Daud, H. B., Zainuddin, N. B., Sokkalingam, R., Abdussamad, Museeb, A., & Inayat, A. (2024). Addressing limitations of the K-means clustering algorithm: Outliers, non-spherical data, and optimal cluster selection. *AIMS Mathematics*, *9*(9), 25070–25097. https://doi.org/10.3934/math.20241222

Koca, Y. B., & Aktepe, E. (2024). Effect of dimension reduction with PCA and machine learning algorithms on diabetes diagnosis performance. *Turkish Journal of Engineering*, *8*(3), 447–456. https://doi.org/10.31127/tuje.1413087

Li, S. (2023). A study on the crucial indicators for breast cancer detection using machine learning algorithm. *Journal of Physics: Conference Series*, *2646*, 012042. https://doi.org/10.1088/1742-6596/2646/1/012042

Midi, H., Suhaiza, J., Aslam, M., Syahida, H., & Amielda, E. (2025). Improved robust principal component analysis based on Minimum Regularized Covariance Determinant for the detection of high leverage points in high dimensional data. *Sains Malaysiana*, *54*(8), 2087–2097. https://doi.org/10.17576/jsm-2025-5408-17

Mienye, I. D., & Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, *15*(755).

Mohammed, M. A., Akawee, M. M., Saleh, Z. H., Hasan, R. A., Ali, A. H., & Sutikno, T. (2023). The effectiveness of big data classification control based on principal component analysis. *Bulletin of Electrical Engineering and Informatics*, *12*(1), 427–434. https://doi.org/10.11591/eei.v12i1.4405

Nandi, B., Jana, S., & Das, K. P. (2023). Machine learning-based approaches for financial market prediction: A comprehensive review. *Journal of AppliedMath*, *1*(Special Issue – Mathematical Models and Applications), 1–18. https://doi.org/10.59400/jam.v1i2.134

Rahman, M. A., Khan, M. S. H., Watanobe, Y., Prioty, J. T., Annita, T. T., Rahman, S., Hossain, M. S., Aitijjo, S. A., Taskin, R. I., Dhrubo, V., Hanip, A., & Bhuiyan, T. (2025). Advancements in breast cancer detection: A review of global trends, risk factors, imaging modalities, machine learning, and deep learning approaches. *BioMedInformatics*, *5*(46), 1–70. https://doi.org/10.3390/biomedinformatics5030046

Samuel, H., Le, S., Dev, A., & Bell, S. (2026). Psychosocial impact of false-positive surveillance for hepatocellular carcinoma: A qualitative study. *Supportive Care in Cancer*, *34*(362), 2–7.

Shrifan, N. H. M. ., Akbar, M. F., & Nor Ashidi, M. I. (2022). An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*, 6365–6376. https://doi.org/10.1016/j.jksuci.2021.07.003

Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics*, *12*(116), 1–22. https://doi.org/10.3390/diagnostics12010116

Ulku, I., & Akagündüz, E. (2022). A survey on deep learning-based architectures for semantic segmentation on 2D images. *Applied Artificial Intelligence*, *36*(1), 1–45. https://doi.org/10.1080/08839514.2022.2032924

Yaqoob, A., Verma, N. K., Mir, M. A., Tejani, G. G., Eisa, N. H. B., Mamoun Hussien Osman, H., & Shah, M. A. (2025). SGA-driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study. *Scientific Reports*, *15*, 10944. https://doi.org/10.1038/s41598-025-95786-1

Zahariah, S., & Midi, H. (2023). Minimum regularized covariance determinant and principal component analysis-based method for the identification of high leverage points in high dimensional sparse data. *Journal of Applied Statistics*, *50*(13), 2817–2835. https://doi.org/10.1080/02664763.2022.2093842

Zhu, A., Hua, Z., Shi, Y., & Tang, Y. (2021). An improved K-means algorithm based on evidence distance. *Entropy*, *23*(1550), 1–15.