MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

Penerbit USIM PRESS

Article

A Comparative Study of Principal Component Analysis with Ensemble Learning for Classification of Medical Data

Siti Amirah Batrisya Mohd Rahaizi[1], Wendy Ling Shinyie[1] and Soo-Fen Fam[2]

[1] Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Malaysia.

[2] Faculty of Technology Management and Technopreneurship, Universiti Teknikal Malaysia Melaka, 75450, Melaka, Malaysia.

Correspondence should be addressed to:
Wendy Ling Shinyie; sy_ling@student.upm.edu.my

Article Info
Article history:
Received:1 January 2026
Accepted: 15 April2026
Published: 5 May 2026

Academic Editor:
Shahrina Ismail

Malaysian Journal of Science, Health & Technology

MJoSHT2025, Volume 12, Issue No. 1
eISSN: 2601-0003

https://doi.org/10.33102/mjosht.572

***Abstract*— Dimensionality reduction is a critical component in the analysis of medical data, specifically when addressing challenges like multicollinearity, noise, and high-dimensional feature spaces that can decrease classification performance. While principal component analysis (PCA) is a traditional choice, its utility in medical datasets is often hindered by outliers, corrupted observations, and low interpretability, as principal components are linear combinations of all original variables. This research compares PCA, robust PCA (RPCA), and sparse PCA (SPCA) integrated with random forest (RF) and extremely randomized trees (ERT). A simulation study revealed that while all PCA variants struggle with low class separation, RPCA and SPCA significantly outperform standard PCA in the presence of outliers. This study utilized a diabetes dataset that underwent thorough preprocessing, including median imputation, normalization, and the synthetic minority over-sampling technique (SMOTE) to address class imbalance. Model optimization involved cross-validation of the RPCA regularization parameter and the SPCA sparsity parameter based on the area under the receiver operating characteristic (ROC) curve (AUC). At the same time, RF and ERT hyperparameters were optimized using a two-stage random and grid search approach. Final empirical results demonstrate that the RPCA-ERT model is superior, achieving an accuracy of 0.8954 and a sensitivity of 0.9434, underscoring its effectiveness in managing contaminated medical data.**

***Keywords*— Principal Component Analysis; Random Forest; Extremely Randomized Trees; Dimensionality Reduction**

I. INTRODUCTION

Type II diabetes mellitus is a global chronic disease increasing due to aging and lifestyle changes. Early diagnosis is vital to prevent severe consequences such as kidney failure and cardiovascular disease. While modern medical technology generates large-scale clinical data, including medical measurements, laboratory test results, and demographic information, these datasets are often complex, high-dimensional, and noisy.

Despite the increasing accessibility of clinical data for diabetes diagnosis, building accurate machine learning models remains challenging. Clinical data is typically high-dimensional and prone to outliers, which may result from measurement errors or biological variability and compromise model stability, computational efficiency, and classification performance.

In principal component analysis (PCA), the goal is to maximize variance by performing linear projections to reduce the dimensionality of a dataset. By transforming correlated variables into orthogonal principal components, redundancy is removed, making complex medical datasets more manageable. The effectiveness of PCA in healthcare is well-documented. For instance, Zain et al. [1] showed that naive Bayes (NB) and reduced-error pruning tree (REPTree) performance improved when PCA was used to predict breast cancer recurrence. Furthermore, BaniMustafa et al. [2] demonstrated that PCA enhanced *k*-nearest neighbours (KNN) models, increasing the area under the curve (AUC) from 98.6% to 99.6%.

Beyond disease detection, PCA is utilized in monitoring fetal health and analysing elderly mobility [3, 4]. However, standard PCA has notable limitations. It is highly sensitive to outliers, which can bias principal components and misrepresent the real structure of the dataset [5]. Additionally, because each component typically depends on all original variables, PCA produces dense loadings that are difficult to interpret [6]. These problems motivate the investigation of PCA variants that offer greater robustness and interpretability.

Robust PCA (RPCA) addresses the tendency of extreme observations to distort estimated components. Grounded in matrix decomposition theory and principal component pursuit (PCP), RPCA models the data matrix as the sum of a low-rank component (capturing systematic structure) and a sparse component (capturing outliers or corruption). This is particularly useful for biomedical data containing measurement artifacts [7].

Literature highlights RPCA's efficacy in handling noise. Gibson et al. [8] used PCP to filter for extreme environmental health exposures, while Liu et al. [9] showed that RPCA-based approaches consistently outperform alternative feature-extraction techniques. Despite these strengths, RPCA faces challenges, including high computational costs and the need for precise regularization tuning to avoid over-separating significant variation as outliers [10].

Sparse PCA (SPCA) imposes constraints on component loadings, ensuring that each component depends on only a small subset of variables. This method utilizes sparsity regularization and elastic net optimization to balance structure retention with model complexity. In medical contexts where transparency is crucial, SPCA components map more easily to clinically relevant features [11].

Studies indicate that sparse loadings do not diminish downstream effectiveness. Baytas et al. [12] used SPCA to identify major clinical attributes in electronic health records, while Rehman et al. [13] achieved 91.11% accuracy in heart failure prediction. However, a fundamental trade-off exists, where SPCA often explains less variance than standard PCA. This trade-off is often deemed acceptable because interpretability, rather than maximal variance, is the primary objective [14].

Ensemble learning is popular in healthcare due to its ability to handle complex data and achieve high accuracy. Random forests (RFs) and extremely randomized trees (ERTs) are two notable models that aggregate multiple decision trees to reduce variance. RF is effective at predicting postoperative complications and cardiovascular conditions [15, 16]. ERT extends this by introducing additional randomness into split selection, further reducing variance in noisy data. For diabetes prediction, Kate et al. [17] found that ERT (96.58%) outperformed RF. Despite their power, both methods lack transparency and can be susceptible to bias in imbalanced clinical datasets [18].

The combination of PCA and ensemble classifiers addresses high-dimensional, redundant clinical data. Ahmed et al. [19] achieved 98.05% accuracy in coronary heart disease classification using PCA-RF, while Rehan et al. [20] used PCA-ERT to identify diabetes indicators from spectral data with 92.8% accuracy.

Despite these promising results, most existing work relies on standard PCA has overlooked how robust or sparsity-driven variants influence ensemble behaviour. There is a lack of a systematic comparison of the trade-offs among accuracy, noise robustness, and interpretability within a single framework.

Therefore, a comparative study is necessary to evaluate standard PCA, RPCA, and SPCA when integrated with RF and ERT classifiers. This research seeks to fulfill this gap by examining how these PCA variants influence ensemble-based diabetes diagnosis. Specifically, the study aims to assess the sensitivity of standard PCA, RPCA, and SPCA through controlled simulations with varying sample sizes, class separations, and outlier proportions, investigate the impact of different PCA techniques on type II diabetes classification using real clinical data, and evaluate and compare the performance of RF and ERT when utilizing PCA-based feature representations.

II. Materials and Methods

This study employs a two-phase methodology to evaluate the effectiveness of various PCA-based dimensionality reduction techniques. The first phase consists of a controlled simulation study designed to assess algorithm performance under specific conditions of class separability and data contamination. The second phase involves an empirical analysis of the Pima Indians Diabetes (PIDD) dataset. For both phases, the data undergo preprocessing, including missing value imputation, outlier detection, and normalization. The dataset is split into training and test sets at 80/20. To address class imbalance within the training data, the synthetic minority oversampling technique (SMOTE) is applied. Subsequently, dimensionality reduction methods (PCA, RPCA, and SPCA) are utilized to reduce the feature space while preserving essential information. The processed data is then used to train the ensemble classifiers, RF and ERT. Lastly, the models are evaluated on the independent test set using several performance metrics, including accuracy, sensitivity (recall), precision, F1-score, and the area under the receiver operating characteristic curve (AUC).

A. Data Source

In this study, the PIDD was used to evaluate the proposed classification framework, as it is a widely used benchmark dataset in machine learning research for diabetes classification. The data were originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and contain clinical information for 768 female patients (aged 21 years and older) from a population near Phoenix, Arizona, USA [21]. The dataset comprises 768 records and 8 features. The target variable, denoted as 'Outcome', indicates the presence (1) or

absence (0) of type II diabetes. Table I shows the predictor and target variables analysed in this study.

B. Data Preprocessing

Several preprocessing steps were conducted prior to model construction. Medical datasets frequently contain irregularities such as invalid entries, extreme values, and imbalanced class distributions, which can negatively affect model performance. Consequently, careful data preparation was performed to improve data quality while preserving clinically relevant information.

TABLE I. DESCRIPTION OF PREDICTOR AND TARGET VARIABLES

Variable Name	Variable Type	Description	Measurement Scale
Outcome	Target Variable	Diabetes status (1=diabetic, 0=non-diabetic)	Binary (Categorical)
Pregnancies	Predictor Variable	Number of times the patient has been pregnant	Discrete (Numeric)
Glucose	Predictor Variable	Plasma glucose concentration (mmol/L)	Continuous (Numeric)
Blood Pressure	Predictor Variable	Diastolic blood pressure (mm Hg)	Continuous (Numeric)
Skin Thickness	Predictor Variable	Triceps skinfold thickness (mm)	Continuous (Numeric)
Insulin	Predictor Variable	Insulin level (mu U/ml)	Continuous (Numeric)
Body Mass Index (BMI)	Predictor Variable	Body mass index (kg/m^2)	Continuous (Numeric)
Diabetes Pedigree Function	Predictor Variable	Diabetes pedigree function (genetic risk indicator)	Continuous (Numeric)
Age	Predictor Variable	Age of the patient	Discrete (Numeric)

1) Handling Missing and Invalid Values:

Missing values occur when data for variables are missing, a common issue in medical datasets. In the PIDD, invalid zero values for certain physiological variables were marked as 'NA' prior to imputation. To address these, a median imputation technique was employed. This method was chosen because it is less sensitive to outliers and skewed distributions compared to mean imputation, making it more suitable for clinical variables. Furthermore, median imputation was performed separately based on the outcome group. This ensures that imputed values remain consistent with the underlying distribution of each class, maintaining clinically relevant differences between diabetic and non-diabetic individuals. By combining robustness to outliers with outcome-specific imputation, this technique preserves the observations and prevents the data structure from being distorted.

2) Outlier Detection:

Outliers are data points that deviate significantly from the rest of the dataset. This is a widespread issue in medical data due to natural variability and measurement irregularities. In this study, outliers were identified using the interquartile range (IQR) rule, which flags values more than 1.5xIQR below the first quartile or above the third quartile. This technique is preferred as it handles skewed data without requiring parametric distribution assumptions. Importantly, outliers were retained in the dataset to preserve clinically relevant variability and to evaluate how the dimensionality reduction techniques handle such noise.

3) Normalization:

PCA is inherently sensitive to the scale of the input data as it seeks to maximize variance. Consequently, variables with larger numerical scales can disproportionately influence the results, potentially obscuring the true underlying structure of the data. To solve this problem, all variables were standardized prior to dimensionality reduction. Z-score normalization was applied to ensure that each feature contributed equally to the analysis before PCA. This process involves scaling the data to have a mean of zero and a standard deviation of one, thereby creating a uniform scale across all clinical measurements.

4) Class Imbalance Handling

The target variable 'Outcome' comprises 500 non-diabetic and 268 diabetic patients, indicating a clear class imbalance. To address this, the synthetic minority oversampling technique (SMOTE) was employed. SMOTE is particularly effective at mitigating bias in model predictions caused by underrepresentation of minority classes. Unlike simple oversampling, SMOTE creates synthetic samples by interpolating between existing minority class samples. Specifically, a random sample from the minority class is selected, and its *k*-nearest neighbours are identified. A synthetic sample is then generated at a random point along the line segment connecting the original sample and a chosen neighbour. This technique enhances the model's ability to learn from the minority class, leading to more balanced and effective classification outcomes.

C. Dimensionality Reduction Methods

Dimensionality reduction is a critical preprocessing step for high-dimensional datasets, as it simplifies models, accelerates computation, and mitigates the risk of overfitting. In this study, three distinct dimensionality reduction techniques were applied to the diabetes data: standard principal component analysis (PCA), robust principal component analysis (RPCA), and sparse principal component analysis (SPCA). Each technique offers specific advantages tailored to different data characteristics and research objectives, such as handling outliers or enhancing model interpretability.

1) Principal Component Analysis

PCA is a standard dimensionality reduction technique that transforms data into a new coordinate system of principal components (PCs), which are linear combinations of the original variables. This is achieved by calculating the data's covariance matrix to determine its eigenvectors and eigenvalues. The eigenvectors define the directions of maximum variance, while the associated eigenvalues quantify the variance explained by each component. To determine the

optimal number of PCs to retain, this study evaluates the cumulative percent variance (CPV) alongside a scree plot. The CPV approach involves retaining enough components to represent a specified threshold of total variance, typically between 70% and 90%. In addition, the scree plot identifies the 'elbow' point where the eigenvalue curve levels off, suggesting the ideal number of components. Standard PCA was employed here to capture significant multivariate dependencies. This is particularly effective for medical datasets where physiological features, such as BMI, insulin, and glucose levels, are frequently correlated, enabling a simplified yet representative feature space.

2) Robust Principal Component Analysis

Despite its effectiveness at capturing data variation, standard PCA is highly sensitive to outliers, which are common in medical datasets due to extreme values or erroneous readings. To address this limitation, robust PCA (RPCA) was implemented for its ability to handle noisy data and outliers. This method decomposes the data matrix X into two distinct components, as shown in Equation 1:

$$min|L| + \lambda\|S\|_1 \text{ subject to } L+S=X \quad (1)$$

where

$\|L\|$: Nuclear norm promoting a low-rank structure

$\|S\|_1$: $\ell 1$-norm, which encourages sparsity

λ: Regularisation parameter controls the balance between the low-rank and sparse components

X: Observed data matrix

L: Low-rank matrix

S: Sparse matrix.

RPCA decomposes the observed data matrix into two components: a low-rank matrix that captures the underlying structure and a sparse matrix that contains irregularities and noise. In this study, the regularization parameter was determined via 5-fold cross-validation to maximize the area under the curve (AUC). AUC was selected as the tuning criterion because it is a threshold-independent measure of discriminative performance, making it highly suitable for medical classifications where class imbalance may exist. Optimizing the parameter based on the maximum AUC ensures that the resulting low-rank representation is effective for class separability while remaining robust to outliers. Consequently, RPCA is less influenced by anomalies, which often significantly impact model performance in biomedical settings. For instance, outliers far outside the expected physiological range can distort standard PCA outcomes. By using RPCA, this study extracts principal components that are free of data errors before proceeding to the classification stage.

3) Sparse Principal Component Analysis

While standard PCA is effective for dimensionality reduction, the resulting principal components are linear combinations of all original variables, which limits interpretability. In medical applications, this lack of sparsity makes it difficult to identify which clinical features contribute most significantly to each component. To address this limitation, Sparse PCA (SPCA) was employed in this study. SPCA extends standard PCA by introducing sparsity constraints on the component loadings, allowing each principal component to depend on only a subset of the original variables. The equation of SPCA is

$$\hat{\beta}_{SPCA} = arg\ \min_{\beta} \|X - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \quad (2)$$

where

$\hat{\beta}_{SPCA}$: Estimated sparse loading vector obtained from SPCA

X : Standardized data matrix of predictor variables

β : Loading vector for a sparse principal component

$\|.\|_2^2$: Squared Frobenius ($\ell 2$) norm measuring reconstruction error

$\|\beta\|_1$: $\ell 1$-norm enforcing sparsity in the loadings

$\|\beta\|_2^2$: $\ell 2$-norm (ridge penalty) ensures numerical stability

λ_1: Sparsity tuning parameter controlling the degree of variable selection

λ_2: Regularisation parameter controlling coefficient shrinkage.

SPCA typically retains the same number of principal components as standard PCA but enforces sparsity in the loadings, ensuring each component depends only on a subset of variables. This is particularly beneficial for high-dimensional datasets, as it facilitates dimensionality reduction while enhancing interpretability. Consistent with the RPCA strategy, the L1 sparsity parameter was selected via 5-fold cross-validation to maximize AUC. This criterion provides a threshold-independent assessment suitable for imbalanced medical data. Although the elastic net formulation includes both L1 and L2 regularization, only the L1 parameter was tuned. The L2 (ridge) penalty was fixed to ensure numerical stability and prevent overfitting, following standard implementations in the R elasticnet package to avoid overparameterization. Consequently, only a small number of influential features contribute to each principal component. By focusing on these key variables, SPCA mitigates overfitting, a frequent challenge in high-dimensional analysis [22]. Ultimately, SPCA was selected for this study because it balances effective dimensionality reduction with the clinical interpretability essential for medical applications.

D. Classification Models

After dimensionality reduction, classification models were trained to predict diabetes status using the transformed feature representations. Ensemble-based tree classifiers were selected for their high predictive accuracy, robustness to noise, and ability to capture the nonlinear relationships often found in clinical datasets. Specifically, the effectiveness of the PCA-based feature extraction techniques was evaluated using Random Forest (RF) and Extremely Randomized Trees (ERT).

1) Random Forest

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and combines their predictions through majority voting. Each tree is trained on a bootstrap sample of the data, with a random subset of features selected at each node to ensure tree diversity and reduce ensemble correlation. This randomness makes RF resistant to overfitting and highly effective for noisy, high-dimensional medical data. Furthermore, RF handles nonlinear relationships without requiring rigid distributional assumptions and remains robust against outliers and multicollinearity [23].

In this study, RF was applied to the feature spaces generated by PCA, RPCA, and SPCA. Model hyperparameters, including the number of variables per split, node size, and maximum terminal nodes, were optimized via a two-step process, which involved an initial random search followed by a refined grid search. A 5-fold cross-validation framework was utilized to evaluate candidates and ensure the selection of a robust model.

2) Extremely Randomized Trees

Extremely randomized trees (ERT), or Extra Trees, are an ensemble classifier closely related to random forest but with additional randomness introduced during tree construction. While RF identifies the optimal split from a feature subset, ERT randomizes both the feature choice and the split threshold at each node. This added randomness increases tree diversity and reduces correlation, often improving generalization on complex datasets. Furthermore, ERT maintains low computational complexity by avoiding exhaustive split-point searches, making it ideal for high-dimensional data. This randomness also enhances robustness against noise and fluctuations in the training set. Recent studies have demonstrated ERT's efficacy in biomedical classification, where feature interactions are often unstable and nonlinear [24].

To ensure a fair comparison, the ERT model was trained on the same dimensionality-reduced datasets as the RF. Key hyperparameters, including the number of features per node, minimum node size, and the total number of trees, were optimized using a randomized grid search with 5-fold cross-validation, with AUC as the primary evaluation metric.

E. Hyperparameter Optimization

Hyperparameter optimization is critical for developing reliable ensemble classifiers, particularly for high-dimensional medical datasets. To avoid the computational burden of an exhaustive grid search, this study employs a two-step optimization strategy. First, a random search is conducted to perform a broad exploration of the hyperparameter space. This approach efficiently identifies promising regions without the high complexity of exhaustive methods, a strategy proven effective for ensemble algorithms like RF and ERT. Subsequently, a grid search is utilized to systematically refine the model within the narrow ranges identified during the random search. This hybrid approach combines efficient global exploration with targeted local optimization, which is highly recommended in medical machine learning research to ensure model stability and reliable classification performance.

F. Model Evaluation Metrics

After implementing the PCA variants and ensemble classifiers, model performance was evaluated using AUC, accuracy, sensitivity, specificity, precision, and F1-score. These measures were derived from the four fundamental components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The equations for performance metrics are

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

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

AUC is a quantitative measure of the accuracy of diagnostic tests. According to the criteria established by Bradley [25], AUC values are categorized as follows: 0.50–0.60 (fail), 0.60–0.70 (poor), 0.70–0.80 (fair), 0.80–0.90 (good), and 0.90–1.00 (excellent). In medical research, the AUC is particularly advantageous because it provides a more comprehensive assessment than simple accuracy by accounting for the trade-off between the true positive rate and the false positive rate.

TABLE III. CONFUSION MATRIX

	Predicted Diabetic	**Predicted Non-Diabetic**
Actual Diabetic	True Positives (TP)	False Negatives (FN)
Actual Non-Diabetic	False Positives (FP)	True Negatives (TN)

The confusion matrix shown in Table II provides a comprehensive breakdown of the TP, TN, FP, and FN generated by each classifier. This detailed analysis facilitates a deeper evaluation of model performance and is instrumental in identifying specific patterns of misclassification. In this study, model selection was guided by a multifaceted evaluation of classification performance on the test dataset. Rather than relying on a single metric, the optimal model was identified based on consistent performance across multiple measures, including accuracy, sensitivity, precision, F1-score, and AUC. The selected model represents the most stable and reliable configuration, ensuring its suitability for clinical decision-support where both predictive accuracy and robustness are essential.

III. RESULTS AND DISCUSSION

The purpose of this study is to assess the performance of PCA, RPCA, and SPCA as feature extraction methods prior to classification.

A. Simulation Study

In this simulation study, synthetic data is generated to mimic the Pima Indian Diabetes Dataset (PIDD), featuring 8 variables and a binary target: 0 for non-diabetic and 1 for diabetic. The study evaluates how each method behaves across variations in sample size (100 and 1,000), class separability (0.1 and 3.0), and outlier proportion (0% and 40%). To eliminate class imbalance, proportions were set to 50% for each class. Observations were generated from two multivariate normal distributions with identity covariance, implying uncorrelated variables. The mean vector for non-diabetics was fixed at zero, while the mean vector for diabetics was shifted by a constant value across all features to control class separability. These simulations included two levels of separability: 0.1 for weak separability (implying high overlap) and 3.0 for strong separability (implying highly distinct classes). This research is motivated by common challenges in medical datasets, including measurement noise, data contamination, and overlapping class distributions.

After data generation, the datasets were randomly split at an 80:20 ratio for training and testing, ensuring class balance. Feature standardization was performed by applying the training set's mean and variance to the test set. This ensures that the test set is transformed using the same reference frame as the training environment, thereby preventing data leakage and ensuring a realistic performance evaluation. Dimensionality reduction was then individually applied to the standardized training datasets using PCA, RPCA, and SPCA. In all cases, the first six principal components were retained to maintain consistency across methods. The resulting components served as inputs for the RF classifier. Model performance on the test set was evaluated using accuracy (ACC), sensitivity (SEN), precision (PRE), F1-score (F1), and AUC. To account for stochastic variability, each simulation configuration was repeated 100 times, and performance metrics were averaged across these repetitions to obtain stable, reliable estimates.

TABLE IIIII. SIMULATION RESULTS OF WEAK CLASS SEPARABILITY AND NO OUTLIERS

Method	N	ACC	SEN	PRE	F1	AUC
PCA	100	0.500	0.489	0.495	0.485	0.609
RPCA	100	0.506	0.496	0.509	0.476	0.609
SPCA	100	0.515	0.519	0.511	0.512	0.606
PCA	1000	0.518	0.515	0.518	0.516	0.538
RPCA	1000	0.504	0.476	0.511	0.467	0.532
SPCA	1000	0.512	0.505	0.513	0.508	0.529

Under conditions of weak class separability and the absence of outliers, Table III shows that all PCA variants performed poorly regardless of sample size. For $N = 100$, accuracy values hovered around random guessing, ranging from 0.500 to 0.515. Sensitivity and precision were similarly low, while AUC values remained modest, lying between 0.606 and 0.609, indicating a limited discriminative ability. A similar pattern was observed for $N = 1000$. Accuracy remained approximately 0.50 across all methods, with no significant improvement in sensitivity, precision, F1-score, or AUC. The fact that increasing the sample size did not improve results indicates that strong class overlap influences model behaviour more significantly than sample size. Overall, when class separability is low and the data is clean, no single PCA variant offers a distinct advantage, a trend observed consistently across all evaluation metrics.

TABLE IVV. SIMULATION RESULTS OF WEAK CLASS SEPARABILITY AND THE PRESENCE OF OUTLIERS

Method	N	ACC	SEN	PRE	F1	AUC
PCA	100	0.493	0.535	0.493	0.567	0.597
RPCA	100	0.507	0.489	0.507	0.479	0.615
SPCA	100	0.502	0.495	0.502	0.490	0.593
PCA	1000	0.502	0.699	0.502	0.560	0.529
RPCA	1000	0.503	0.507	0.505	0.478	0.531
SPCA	1000	0.514	0.518	0.514	0.515	0.536

Under weak class separability and the presence of outliers, classification performance remained poor across all methods and sample sizes (Table IV). For $N = 100$, all PCA variants yielded approximately 0.50 accuracy. While standard PCA showed higher sensitivity (0.535), its low precision (0.493) led to an inflated F1-score driven by false positives. RPCA and SPCA similarly struggled, with AUC values for all methods remaining low (0.593–0.615). For $N = 1000$, accuracy stayed near 0.50. PCA showed high sensitivity (0.699) but poor precision, yielding a weak F1-score and an AUC of approximately 0.53. Although RPCA and SPCA provided a better balance between sensitivity and precision, overall discrimination remained weak. Ultimately, robustness to outliers cannot compensate for poor class separation, and all PCA variants fail to provide reliable classification in this scenario.

TABLE V. SIMULATION RESULTS OF HIGH CLASS SEPARABILITY AND NO OUTLIERS

Method	N	ACC	SEN	PRE	F1	AUC
PCA	100	0.999	1.000	0.999	0.999	1.000
RPCA	100	0.805	0.808	0.817	0.804	0.903
SPCA	100	1.000	1.000	1.000	1.000	1.000
PCA	1000	1.000	1.000	1.000	1.000	1.000
RPCA	1000	0.758	0.748	0.766	0.755	0.854
SPCA	1000	1.000	1.000	1.000	1.000	1.000

Under conditions of high class separability and clean data, Table V reveals noticeable performance differences among the PCA variants. At $N = 100$, standard PCA and SPCA achieved near-perfect results, with accuracy, sensitivity, precision, F1-score, and AUC all approximating 1.00. This indicates that both methods effectively captured the discriminative structure of the data. In contrast, RPCA yielded slightly lower scores across all metrics, suggesting a loss of discriminative information during the low-rank decomposition. A similar pattern emerged at $N = 1000$. While PCA and SPCA maintained perfect performance, RPCA's metrics continued to decline, with accuracy dropping to 0.758 and AUC to 0.854. This consistent reduction across all evaluation metrics suggests that the RPCA algorithm may be overly aggressive in filtering variation when the data is already clean and well-separated, leading to the removal of useful signal along with noise.

TABLE VI. SIMULATION RESULTS OF HIGH CLASS SEPARABILITY AND THE PRESENCE OF OUTLIERS

Method	N	ACC	SEN	PRE	F1	AUC
PCA	100	0.504	0.977	0.504	0.665	0.598
RPCA	100	0.762	0.785	0.767	0.758	0.847
SPCA	100	0.811	0.869	0.793	0.821	0.891
PCA	1000	0.500	0.962	0.499	0.655	0.552
RPCA	1000	0.793	0.912	0.740	0.812	0.870
SPCA	1000	0.804	0.839	0.786	0.811	0.923

Under conditions of high class separability and contaminated data, Table VI reveals distinct variations in the robustness of the PCA variants. At $N = 100$, the performance of standard PCA declined significantly, with accuracy dropping to 0.504 and AUC to 0.598, although sensitivity remained high at 0.977, precision fell to 0.504, resulting in an F1-score of 0.665. This demonstrates that standard PCA is highly sensitive to outliers, with a strong tendency to produce false positives. In contrast, RPCA demonstrated improved robustness, achieving a more balanced performance across all metrics (Accuracy: 0.762; AUC: 0.847). SPCA achieved the highest overall performance, with an accuracy of 0.811 and an AUC of 0.891. A similar trend is shown at $N = 1000$. Standard PCA continued to yield low accuracy (0.500) and poor precision (0.500), despite high sensitivity, resulting in unstable F1 scores and an AUC of 0.552. RPCA showed marked improvement, with accuracy rising to 0.793 and AUC to 0.870. However, SPCA outperformed all other methods, providing the highest AUC (0.923) and the most balanced results across all evaluation metrics.

This simulation study compared standard PCA, RPCA, and SPCA across varying sample sizes, class separability levels, and outlier contamination proportions. Overall, the results demonstrate that the effectiveness of PCA-based dimensionality reduction is highly dependent on data characteristics. When class separability is low, no PCA-based method performs satisfactorily, regardless of sample size or the presence of outliers. In scenarios where classes are well-separated and outliers are absent, standard PCA and SPCA achieve optimal results; however, RPCA underperforms due to unnecessary information loss. Conversely, in datasets with high outlier contamination, both SPCA and RPCA perform well, whereas standard PCA becomes highly unstable due to its high sensitivity and low precision.

B. Real Diabetes Data Results

This section presents empirical results from the real-world diabetes dataset. Unlike the simulation study, real-world analysis involves complexities such as missing values, class imbalance, and high variability. We evaluate different dimensionality reduction methods and ensemble classifiers on this data, comparing these findings with the patterns observed in the simulation study.

1) Missing Value Imputation: Before proceeding with model development, invalid zero entries in several clinical variables were identified and treated as missing data. The affected variables included glucose, blood pressure, skin thickness, insulin, and body mass index (BMI). Following this identification, median imputation stratified by diabetes outcome was conducted. Once the imputation process was complete, missing values were no longer present in the dataset. As shown in Table VII, the descriptive summary of the imputed data confirms the absence of missing entries across these variables, verifying that the dataset is sufficiently complete for further analysis.

TABLE VII. SUMMARY STATISTICS OF THE IMPUTED DIABETES DATASET

Variable	Min	Q1	Median	Mean	Q3	Max
Pregnancies	0.00	1.00	3.00	3.85	6.00	17.0
Glucose	44.0	99.75	117.00	121.68	140.25	199.00
Blood Pressure	24.0	64.00	72.00	72.39	80.00	122.00
Skin Thickness	7.00	25.00	28.00	29.09	32.00	99.00
Insulin	14.0	102.50	102.50	141.80	169.5	846.00
BMI	18.0	27.50	32.05	32.43	36.60	67.10
Diabetes Pedigree Function	0.0780	0.2437	0.3725	0.4719	0.6262	2.4200
Age	210	24.00	29.00	33.24	41.00	81.00

2) Outlier Detection: Outliers in the imputed data were identified using the interquartile range (IQR) rule. Table VIII summarizes the outlier counts for each predictor variable, calculated using the IQR limits. Table VIII presents the outlier detection results for each predictor feature based on the IQR method. Several features exhibited a high number of outliers. For instance, skin thickness contained 87 outliers, insulin had 51, and the diabetes pedigree function had 29. In contrast, the IQR rule identified zero outliers for glucose, indicating a relatively stable distribution. In medical datasets, outliers are common due to natural physiological variability and potential measurement noise.

TABLE VIII. OUTLIER DETECTION SUMMARY FOR EACH FEATURE

Variable	Q1	Q3	IQR	Lower Bound	Upper Bound	Number of Outliers
Pregnancies	1.0000	6.0000	5.0000	-6.50	13.50	4
Glucose	99.7500	140.2500	40.5000	39.00	201.00	0
Blood Pressure	64.0000	80.0000	16.0000	40.00	104.00	14
Skin Thickness	25.0000	32.0000	7.0000	14.50	42.50	87
Insulin	102.5000	169.5000	67.0000	2.00	270.00	51
BMI	27.5000	36.6000	9.1000	13.85	50.25	8
Diabetes Pedigree Function	0.2438	0.6263	0.3825	-0.33	1.20	29
Age	24.0000	41.0000	17.0000	-1.50	66.50	9

3) Train-Test Split and Class Distribution:

The data were split into training and test sets at 80:20, yielding 615 training and 153 test observations. Initially, the training data exhibited a class imbalance, with a higher proportion of non-diabetic cases. To address this, SMOTE was applied to balance the training set, resulting in 400 non-diabetic and 430 diabetic cases. This was an important step to ensure the models generalize equally for both classes. Conversely, the class distribution in the testing set remained unchanged to

provide an unbiased measure of model performance on new datasets.

4) Dimensionality Reduction:

The analysis focuses on the number of retained components, the proportion of variance explained, and sparsity characteristics. For RPCA and SPCA, we also report the optimal tuning parameters selected using the procedures described in Section II. Standard PCA was applied to the preprocessed training data. The number of principal components to retain was determined using a scree plot and the cumulative proportion of variance explained. As shown in Fig. 1, the first six principal components were sufficient to explain approximately 90% of the total variance. This finding indicates that the original eight features can be effectively reduced to six components while preserving most of the informational signal, which was subsequently used as input for the classification models.

Additionally, RPCA was applied to decompose the training data into a low-rank component and a sparse component representing outliers. Using the tuning procedure described in the previous section, the regularization parameter was selected based on cross-validated AUC. Among the candidate values, the optimal parameter was 0.07, yielding the highest average cross-validated AUC. Using the low-rank component obtained with this tuned parameter, PCA was subsequently applied to generate a reduced representation. To retain at least 90% of the total variance, six principal components were required, as evidenced by both the cumulative variance and the scree plot. Referring to Fig. 2, the scree plot generated from the RPCA low-rank component shows an elbow at the sixth component. Hence, six components are suitable for further analysis.

Besides, SPCA was applied using a sparsity constraint on the component loadings, which shrunk many coefficients to zero. The level of sparsity was determined based on classification performance. Tuning results indicated that the optimal sparsity parameter was 8, achieving the highest average cross-validated AUC while significantly reducing the number of non-zero loadings. Consistent with the standard PCA approach, six components were retained to capture 90% of the total variance. In this setup, SPCA enforced a sparse loading structure while providing a reduced feature representation of similar dimensionality. As shown in Table IX, the loadings derived from the SPCA technique demonstrate that each principal component is calculated using only a subset of the input variables. The primary advantage of SPCA for interpretability is its sparsity, which clarifies the contribution of individual features to each component while maintaining competitive classification performance.

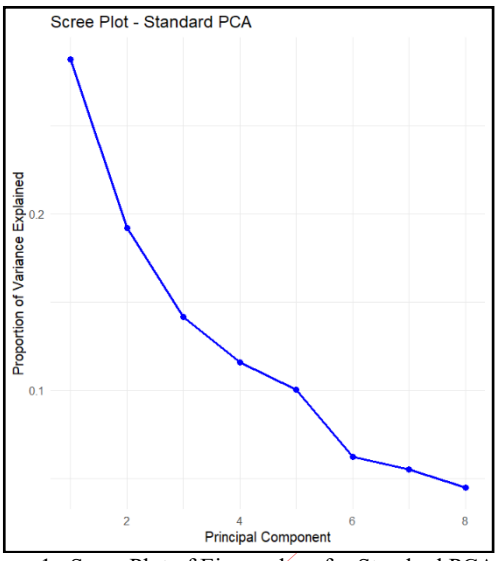

Figure 1. Scree Plot of Eigenvalues for Standard PCA

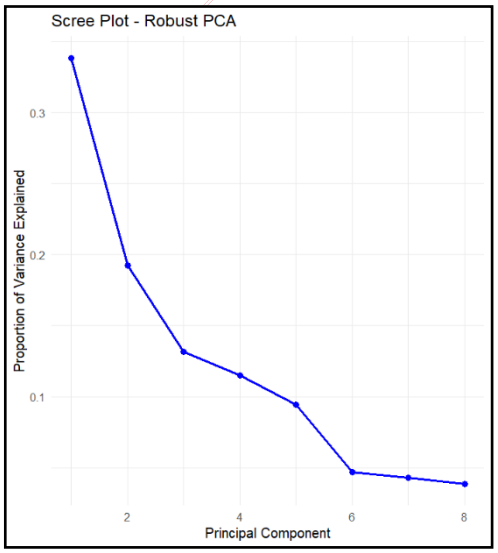

Figure 2. Scree Plot of Eigenvalues for RPCA

TABLE IX. SPARSE PCA LOADING MATRIX FOR THE SELECTED PRINCIPAL COMPONENTS

Feature	PC1	PC2	PC3	PC4	PC5	PC6
Pregnancies	-	-0.748	-	-	0.043	-0.312
Glucose	0.281	-	-0.432	-0.007	0.007	0.735
Blood Pressure	0.229	-	0.046	0.012	-0.935	-
Skin Thickness	0.723	-0.017	0.148	0.044	0.340	-
Insulin	-	-	-0.888	0.006	-	-0.352
BMI	0.576	0.163	-	-0.055	-0.042	-0.418
Diabetes Pedigree Function	-	-	-	-0.997	-	-
Age	0.121	-0.643	-	-0.011	-0.078	0.252

Table IX presents the sparse PCA loading matrix for the six retained principal components. The loading structure exhibits a high degree of sparsity, as it consists of many coefficients equal to zero. This confirms that each principal component incorporates only a limited number of features. In contrast to standard PCA, the total variance preserved by SPCA (86.2%) was lower than the approximately 90% achieved by standard PCA. This reflects the inherent trade-off between variance preservation and interpretability caused by enforcing sparsity in component loadings [26].

The first principal component (PC1) is driven predominantly by skin thickness and BMI, suggesting a link to body composition. PC2 is primarily influenced by age and pregnancy, reflecting maternal history. PC3 is characterized largely by insulin with a significant contribution from glucose, indicating insulin-based metabolic factors. PC4 is driven almost exclusively by the diabetes pedigree function, suggesting a strong link to genetic risk factors. PC5 is primarily associated with blood pressure, indicating a cardiovascular factor, while PC6 incorporates glucose, BMI, and insulin, representing a glucose-mass interaction. Overall, these results demonstrate that SPCA effectively performs feature selection by concentrating each component's contribution on a limited number of clinically meaningful variables. A summary of the dimensionality reduction and tuning parameter results for all PCA variants is provided in Table X.

TABLE X. SELECTED NUMBER OF COMPONENTS AND TUNING PARAMETER FOR PCA VARIANTS

Method	Tuning Parameter	Number of Components	Cumulative Variance Explained
PCA	-	6	0.900
Robust PCA	$\lambda = 0.07$	6	0.919
Sparse PCA	para = 8	6	0.862

5) Classification:

This section presents the classification performance of the reduced feature representations obtained from standard PCA, RPCA, and SPCA. Two ensemble-based classifiers were considered, which are RF and ERT. All models were trained on the balanced training data and evaluated on an independent test set to assess their predictive performance on unseen observations.

For each PCA variant, hyperparameter optimization was conducted separately for the RF and ERT classifiers. A random search strategy was initially employed to identify promising regions of the hyperparameter space, followed by a grid search to refine the selection of the best-performing configurations. The optimal hyperparameters were determined based on cross-validated performance on the training set. Tables XI and XII report the optimal hyperparameters and the corresponding cross-validated AUC values for both classifiers across all three PCA variants.

TABLE XI. OPTIMAL HYPERPARAMETERS AND CROSS-VALIDATED AUC VALUES FOR RANDOM FOREST (RF)

Method	Best mtry	Best nodesize	Best maxnodes	Best AUC
PCA	3	1	50	0.9033
Robust PCA	4	2	48	0.9218
Sparse PCA	3	1	48	0.9246

TABLE XII. OPTIMAL HYPERPARAMETERS AND CROSS-VALIDATED AUC VALUES FOR EXTREMELY RANDOMIZED TREES (ERT)

Method	Best mtry	Best min_node_size	Best num_trees	Best AUC
PCA	3	3	266	0.9219
Robust PCA	2	1	351	0.9353
Sparse PCA	6	1	415	0.9395

The final models were refitted using the selected hyperparameters and applied to the test dataset. This procedure ensured that the model comparison was based on optimally tuned classifiers for each dimensionality reduction method. The performance of all six models, comprising combinations of the three PCA variants and two ensemble-based classifiers, was evaluated using multiple metrics, including accuracy, sensitivity, precision, F1-score, and AUC. The classification results for these PCA-based feature representations using RF and ERT on the test dataset are presented in Table XIII.

TABLE XIII. CLASSIFICATION PERFORMANCE ON THE TEST DATASET

Method	ACC	SEN	PRE	F1	AUC
PCA-RF	0.8497	0.9245	0.7206	0.8099	0.9300
PCA-ERT	0.8889	0.9434	0.7813	0.8547	0.9404
RPCA-RF	0.8824	0.8491	0.8182	0.8333	0.9514
RPCA-ERT	0.8954	0.9434	0.7937	0.8621	0.9456
SPCA-RF	0.8627	0.9057	0.7500	0.8205	0.9326
SPCA-ERT	0.8693	0.8679	0.7797	0.8214	0.9476

Based on the results in Table XIII, both the choice of dimensionality reduction technique and the classifier significantly affect model performance across all evaluation metrics. Regarding dimensionality reduction, the specific PCA variant used has a clear impact on classification outcomes. Across both classifiers, RPCA performs relatively better and more consistently than standard PCA and SPCA. Notably, the RPCA-ERT configuration achieved the highest results in terms of accuracy (0.8954), sensitivity (0.9434), and F1-score (0.8621). This indicates that reducing the influence of contaminated samples prior to feature extraction is highly effective for outcome detection. Such performance is critical for early diabetes detection, where minimizing missed positive cases is a primary objective.

Standard PCA performs competitively, particularly when paired with ERT, achieving high sensitivity and excellent overall metrics. This reflects its effectiveness in capturing the dominant structures within the data, but its inherent sensitivity to outliers may limit performance relative to RPCA [5]. In contrast, models developed using SPCA yielded moderate results across most metrics. According to Machkour et al. [26], while SPCA enhances interpretability by imposing sparsity on principal component loadings, this penalty can reduce the

explained variance, thereby leading to lower accuracy and F1-score compared to RPCA-based models.

Depending on the classifier, all PCA-based models combined with ERT outperformed those using RF. Specifically, ERT-based models exhibited higher accuracy, F1-scores, and sensitivity. Among these, PCA-ERT and RPCA-ERT provided the highest sensitivity (0.9434). This is highly desirable in a medical context, as it minimizes false negatives during screenings. The superior performance of ERT is consistent with research by Kate et al. [17], who demonstrated that ERT can outperform RF in diabetes prediction due to improved generalization. This is largely due to lower tree correlation and greater randomness during tree construction, which are important when handling healthcare datasets characterized by noise and complex feature interactions. The alignment between these findings and previous studies demonstrates the suitability of ERT for medical classification tasks. Across all PCA variants and classifiers, the AUC remained consistently high (above 0.93), indicating that each dimensionality reduction technique provides strong overall discriminative ability.

TABLE XIV. CONFUSION MATRIX OF RPCA-ERT

Prediction	**Reference**	
	Non-diabetic	Diabetic
Non-diabetic	87	3
Diabetic	13	50

Based on the overall performance presented in Table XIV, the RPCA-ERT model was identified as the top-performing configuration. Table XIV presents the confusion matrix for the RPCA-ERT model evaluation. The model correctly identified 50 diabetic patients and 87 non-diabetic patients. However, it misclassified 13 individuals as diabetic who were actually non-diabetic (false positives) and misclassified 3 diabetic cases as non-diabetic (false negatives).

The low number of false negatives demonstrates the model's effectiveness in detecting diabetes, which is a crucial factor in medical applications where missed diagnoses can delay vital treatment. While the model produced some false positives, this trade-off is acceptable within a diagnostic context, as prioritizing sensitivity ensures that fewer diabetic cases are overlooked. In conclusion, the confusion matrix indicates that the RPCA-ERT approach provides reliable and balanced classification.

IV. CONCLUSIONS

This study investigates the performance of PCA-based variants, which are standard PCA, RPCA, and SPCA, combined with ensemble classifiers like RF and ERT for diabetes prediction. The results indicate that the RPCA-ERT combination achieves the best overall performance, demonstrating RPCA's ability to handle outliers, and the additional randomness in ERT favors generalization when handling contaminated data and complex feature interactions. While these findings underscore the importance of outlier-robust feature extraction, the study is limited by its reliance on a single benchmark dataset and its exclusive focus on tree-based ensemble methods. Further research could focus more on interpretability, specifically by combining robust dimensionality reduction with explainability techniques to enhance the utility of machine learning models for early detection of diabetes.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGEMENT

This research was supported by Universiti Putra Malaysia under the Geran Inisiatif Putra Siswazah (GP-IPS), Grant No: [9816600].

REFERENCES

[1] Z. M. Zain, M. Alshenaifi, A. Aljaloud, T. Albednah, R. Alghanim, A. Alqifari, and A. Alqahtani, "Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis," *International Journal of Advances in Intelligent Informatics*, vol. 6(3), pp. 313-327, Nov. 2020. https://doi.org/10.26555/ijain.v6i3.462

[2] BaniMustafa, S. Almatarneh, O. Bulkrock, R. Alazaidah, H. Almahasneh, and G. Samara, "Investigating Principal Component Analysis Impact on the Performance of Machine Learning Classifiers: A Health Informatics Application," in *25th International Arab Conference on Information Technology (ACIT)*, 2024, pp. 1-6, IEEE. https://doi.org/10.1109/ACIT62805.2024.10876953

[3] J. Moreira, B. Silva, H. Faria, R. Santos, and A. S. P. Sousa, "Systematic Review on the Applicability of Principal Component Analysis for the Study of Movement in the Older Adult Population," *Sensors*, vol. 23(1), 205, 2023. https://doi.org/10.3390/s23010205

[4] D. Jiménez-Narváez, V. D. C. Vaca, J. J. Loor-Duque, I. R. A. Martín, I. G. Reyes-Chacón, P. Vizcaíno, and M. E. Morocho-Cayamcela, "Predictive Modeling for Fetal Health: A Comparative Study of PCA, LDA and KPCA for Dimensionality Reduction," *IEEE Access*, vol. 13, pp. 59687-59703, Mar. 2025. https://doi.org/10.1109/ACCESS.2025.3553110

[5] S. S. S. A. Mutalib, W. N. S. W. Yusoff, A. P. Kurniati, N. A. Osman, and Z. Zulhelmy, "Robust Principal Component Analysis in Multivariate Applications," *Journal of Applied Science, Engineering, Technology, and Education*, vol. 7(2), 357-364, 2025. https://doi.org/10.35877/454RI.asci3948

[6] R. Guerra-Urzola, K. Van Deun, J. C. Vera, and K. Sijtsma, "A guide for sparse PCA: Model comparison and applications," *Psychometrika*, vol. 86(4), 893-919, 2021. https://doi.org/10.1007/s11336-021-09773-2

[7] N. Le, S. Song, Q. Zhang, and Wang, R. K, "Robust principal component analysis in optical micro-angiography," *Quantitative imaging in medicine and surgery*, vol. 7(6), 654, 2017. https://doi.org/10.21037/qims.2017.12.05

[8] E. A. Gibson, J. Zhang, J. Yan, L. Chillrud, J. Benavides, Y. Nunez, and M. A. Kioumourtzoglou, "Principal component pursuit for pattern identification in environmental mixtures," *Environmental Health Perspectives*, vol. 130(11), 117008, 2022. https://doi.org/10.1289/EHP10479

[9] J. X. Liu, Y. Xu, C. H. Zheng, H. Kong, and Z. H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12(4), 964-970, 2014. https://doi.org/10.1109/TCBB.2014.2383375

[10] P. P. Brahma, Y. She, S. Li, J. Li, and D. Wu, "Reinforced robust principal component pursuit," *IEEE transactions on neural networks and learning systems*, vol. 29(5), 1525-1538, 2017. https://doi.org/10.1109/TNNLS.2017.2671849

[11] S. Gajjar, M. Kulahci, and A. Palazoglu, "Selection of non-zero loadings in sparse principal component analysis," *Chemometrics and Intelligent*

Laboratory Systems, vol. 162, 160-171, 2017. https://doi.org/10.1016/j.jprocont.2017.03.005

[12] M. Baytas, K. Lin, F. Wang, A. K. Jain, and J. Zhou, J, "Stochastic convex sparse principal component analysis," *Journal on Bioinformatics and Systems Biology*, vol. 15, 2016. https://doi.org/10.1186/s13637-016-0045-x

[13] Rehman, A. Khan, M. A. Ali, M. U. Khan, S. U. Khan, and L. Ali, "Performance analysis of PCA, sparse PCA, kernel PCA and incremental PCA algorithms for heart failure prediction," in *International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2020, IEEE, pp. 1-5. https://doi.org/10.1109/ICECCE49384.2020.9179199

[14] J. Camacho, A. K. Smilde, E. Saccenti, and J. A. Westerhuis, "All sparse PCA models are wrong, but some are useful. Part I: computation of scores, residuals and explained variance," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, 103907, Jan. 2020. https://doi.org/10.1016/j.chemolab.2019.103907

[15] S. H. Boppana, S. S. K. Komati, R. H. Chitturi, R. Raj, and C. D. Mintz, "DiabCompSepsAI: Integrated AI Model for Early Detection and Prediction of Postoperative Complications in Diabetic Patients—Using a Random Forest Classifier," *Journal of Clinical Medicine*, vol. 14(20), 7173, 2025. https://doi.org/10.3390/jcm14207173

[16] P. Tyagi, J. Vijayashree, S. Mathur, and M. Thoke, "Integrating Machine Learning with Clinical Practice: Advancements in Heart Disease Prediction Models," in *International Conference on Data Science and Business Systems (ICDSBS)*, IEEE, 2025, pp. 1-8. https://doi.org/10.1109/ICDSBS63635.2025.11031691

[17] Kate, G. Deepika, M. N. Sravya, and S. Ganesan, "Enhanced diabetes diagnosis using ensemble classifiers with explainable AI and oversampling for imbalanced data," in *Proceedings of the 5th International Conference on Intelligent Technologies (CONIT 2025)*, IEEE, 2025. https://doi.org/10.1109/CONIT65521.2025.11167663

[18] N. A. Amiludin, M. M. Rosli, N. Ibrahim, and W. A. Hammood, "Mental Health Prediction Using Ensemble Learning Approaches with Rebalancing Technique," in *International Conference on Advanced Machine Learning and Data Science (AMLDS)*, IEEE, 2025, pp. 455-460. https://doi.org/10.1109/AMLDS63918.2025.11159407

[19] M. Ahmed, R. Hassan, S. Datto, S. Saleh, S. Islam, K. Redwan, and S. Mahmood, "A Comparative Study of Machine Learning Models for Cardiovascular Risk Prediction Using PCA-Transformed Framingham Data," in *International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, IEEE, 2025, pp. 1-6. https://doi.org/10.1109/QPAIN66474.2025.11171785

[20] Rehan, M. U. Rehman, M. Aamir, and S. Islam, "A CatBoost and ExtraTrees-based softvoting ensemble approach for non-invasive diabetes detection using hair LIBS spectral data," *Microchemical Journal*, vol. 217, 114980, Oct. 2025. https://doi.org/10.1016/j.microc.2025.114980

[21] R. Sanakal, and T. Jayakumari, "Prognosis of diabetes using data mining approach—fuzzy C means clustering and support vector machines," *International Journal of Computer Trends and Technology*, vol. 11(2), 94-98, 2014. https://doi.org/10.14445/22312803/IJCTT-V11P120

[22] R. Drikvandi, and O. Lawal, "Sparse principal component analysis for natural language processing," *Annals of Data Science*, vol. 10, pp.25–41, Feb. 2023. https://doi.org/10.1007/S40745-020-00277-X

[23] Maruotto, F. K. Ciliberti, P. Gargiulo, and M. Recenti, "Feature selection in healthcare datasets: Towards a generalizable solution," *Computers in Biology and Medicine*, vol. 196, 110812, 2025. https://doi.org/10.1016/j.compbiomed.2025.110812

[24] B. U. Rani, G. Bhavana, and S. Vemavarapu, "An enhanced microarray sample classification using machine learning," *Proceedings of the 5th International Conference for Emerging Technology (INCET)*, 2024, pp. 1-5. https://www.scopus.com/pages/publications/85200977011

[25] P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30(7), pp. 1145–1159, Jul. 1997. https://doi.org/10.1016/S0031-3203(96)00142-2

[26] Machkour, A. Breloy, M. Muma, D. P. Palomar, and F. Pascal, "Sparse PCA with false discovery rate controlled variable selection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2024, pp. 9715-9720. https://doi.org/10.1109/ICASSP48485.2024.10448237