JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 1 (2026) 79-90
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Enhancing Nonlinear Feature Reduction through Logistic Regression-Guided Hybrid PCA

Retno Wardhani[1]*, Noraziahtulhidayu Kamarudin[2], Asem Khmag[3]

[1] *Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Lamongan, Lamongan, East Java, INDONESIA*
[2] *Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, MALAYSIA*
[3] *Department of Computer System Engineering, Faculty of Engineering, University of Zawia, Zawia, LIBYA*

*Corresponding Author: retzno.teknik@unisla.ac.id
DOI: https://doi.org/10.30880/jscdm.2026.07.01.007

Article Info

Received: 30 October 2025
Accepted: 23 February 2026
Available online: 10 April 2026

Keywords

Dimensionality reduction, Hybrid-PCA, logistic regression, nonlinear feature reduction

Abstract

The usage of Principal Component Analysis (PCA) to reduce the number of dimensions is commonly because it is easy to use and works well. But in the real world, a lot of datasets have parts with high variance that don't always give useful information for classification. This reserach solves this problem by suggesting a Hybrid PCA Guided by Logistic Regression (PCA–LR), which chooses components based on how well they work for supervised classification instead of just variance. The proposed method tested on four datasets with different structures: Mental Health, Breast Cancer, Fetal Health, and Pistachio. The result shown PCA–LR achieved better accuracy performance than basic PCA in these datasets. Proposed method also works better than Kernel PCA (KPCA) and Sparse PCA (SPCA). The Mental Health dataset shows the biggest improvement, with PCA–LR getting an accuracy of 98.52%, which is 25.71% better than basic PCA. The method gets 97.37% on Breast Cancer and the same best performance as KPCA on Fetal Health and Pistachio. These results suggest that adding lightweight supervised guidance to PCA can make it much more useful on datasets that aren't strictly linear.

1. Introduction

In the real world, the diversity of data type, structure, and data dimensions is one of the things that is quite challenging. Reduction of dimensions in data is important to manage data to be more efficient and effective. Principal Component Analysis (PCA), one of the methods has long been used for dimensionality reduction because of its simplicity and computational efficiency. In practice, however, its usefulness in classification tasks is not always straightforward. When data exhibit partially or fully nonlinear structures, components that capture large variance may fail to preserve information that is important for class discrimination. As a result, if you only rely on the variance base as a criterion, it can cause suboptimal results in the classification of data in the real world [1] [2] [3] [4].

Recent research shows that PCA performance may affect when the data has very high dimensions, and especially when the number of features exceeds the number of data samples (n < p). Under such conditions, covariance estimation becomes unstable. This can lead the distortion of component extraction and ultimately

reduce the effectiveness use of variants based on dimensional reduction [3] [5]. Beyond this issue, PCA also tends to overlook discriminative features when the data distribution is nonlinear or deviates from Gaussian assumptions, and also limiting its suitability for many real-world machine learning tasks. To overcome this, research on reducing the dimension of data in a hybrid manner was introduced. Alomari et. al. Introduce a Grey Wolf Optimizer (GWO)-enhanced PCA that shown improved classification performance on Arabic text datasets when coupled with logistic regression for feature relevance [6]. In other research also shown that Hybrid PCA merging with machine learning method can improve classification performance on complex datasets [7].

In individual hybrid implementations, recent survey and review studies shows that it is widely used as well as in hybrid dimensionality reduction models. Recent research has also shown that combining PCA with supervised learning or optimization strategies can improve robustness against noise, overfitting, and nonlinear data structures. Other hybrid approaches such as PCA combined with Particle Swarm Optimization (PSO) have shown better performance in complex medical imaging tasks, including leukemia image classification, where feature relevance and class discrimination are critical [8].

Several studies show that the application of PCA to supervised learning models is applied with the aim of improving classification performance. Usually, PCA will be used in the preprocessing method as a dimensional reduction and followed by the classifier method. Song et al. used PCA to address multicollinearity in high-dimensional logistic regression tasks [9]. Similarly, hybrid PCA–LR pipelines have been adopted in areas like network intrusion detection [10] and clinical risk prediction for diseases like lupus [11]. However, in all these approaches, PCA only used in its entirety as a preprocessing step, with no influence from the classifier during component extraction. The selection of principal components remains unsupervised, sometimes by setting fixed k or variance-based, often ignoring the actual relevance of features for classification. This can cause the relationship between dimensionality reduction and classification performance to be less than optimal.

With the description of the limitations of the PCA, in this study we proposed: Logistic Regression-Guided Hybrid PCA, where Logistic Regression is integrated not merely as a classifier but as a guide in selecting principal components. The mechanism in the proposed method is expected can aligns the feature selection process with the classification objective, and election of principal component k enabling the extraction of components that are both statistically significant and discriminative. In previous research on hybrid PCA-based approaches, relatively little attention has been given to frameworks that integrate classifier feedback directly into the selection of principal components. Most of the methods applied rely on unsupervised criteria during transformation or apply classifiers only after dimensionality reduction has been completed. The practice of guide component selection using task-specific information is often overlooked [10] [12].

In this research, we evaluate the proposed method on four datasets with varying complexity and nonlinearity, and compare it with standard PCA (in this study we used fixed k using two components), Kernel PCA (KPCA), and Sparse PCA (SPCA). Our goal is to demonstrate the advantage of integrating supervised learning guidance into PCA in order to overcome its linearity constraints.

2. Methodology

This section will discuss all the concepts used in understanding and developing proposed methods.

2.1 Related Work

In data with linear structures, PCA has long been a reference in dimension reduction in data with high dimensions. This is because fundamentally PCA projects data to orthogonal frequencies linearly to maximize variance so as to make PCA's performance simple and effective. However, this will be an obstacle if PCA is applied to data with a nonlinear structure. Several variants of PCA have been developed to overcome this, namely Kernel PCA (KPCA) and Sparse PCA (SPCA). The performance of KPCA is based on radial or polynomial by using the kernel function as an extension of the performance of PCA, so that it can capture the nonlinear structure of data better than PCA [13] [14]. In contrast, SPCA incorporates sparsity constraints through L1 regularization, resulting in loading vectors that involve only a subset of original features and thus improve interpretability [15] [16]. As a result, while PCA prioritizes variance preservation, KPCA emphasizes nonlinear structure modeling, and SPCA focuses on feature-level interpretability, which justifies their use as baseline methods in this study.

Kernel Principal Component Analysis (KPCA) builds on the classic PCA framework by adding nonlinearity with kernel functions. KPCA doesn't work directly in the original input space. Instead, it maps the data into a higher-dimensional feature space, where it then extracts the principal components. This kernel-based transformation allows KPCA to show complicated nonlinear relationships that linear PCA often misses [17]. KPCA can show hidden patterns in data that follow curved or manifold-like distributions instead of assuming a purely linear covariance structure. It does this by using kernel functions like radial basis, cosine, or polynomial kernels.

Latest study shows that KPCA makes classification work better in areas where nonlinear relationships are important. The earlier study demonstrated that quantum kernel PCA preserves considerably more information than linear PCA in intricate chemi-resistive sensor arrays, especially in low-dimensional contexts [18]. A multi-

kernel PCA (MKPCA) approach combined with Nyström approximation was suggested for finding financial fraud because it can reduce dimensionality while keeping the nonlinear structure. This method works better than both linear PCA and single-kernel KPCA [17].

It has been shown that both KPCA and SPCA fix certain problems with classical PCA. KPCA helps with the linearity problem, and SPCA makes it easier to understand components and makes them more resistant to noise. In our tests, we use these two methods as strong comparison points to see how well the proposed Logistic Regression-Guided Hybrid PCA works. Its goal is to find a balance between nonlinear sensitivity and classification relevance.

Recently, there has been more and more research into hybrid dimensionality reduction schemes that use Principal Component Analysis (PCA) along with supervised classifiers or nonlinear feature transformations to get around the problems with variance-based component selection. Table 1 shows that most hybrid methods combine PCA or its nonlinear versions with downstream learning models like SVM, LSTM, or ANN to make the reduced feature space better before classification. These hybrids always report better classification performance and robustness than linear PCA alone, no matter what application area they are used in, such as industrial systems, biomedical signals, or health-related datasets. Overall, these results suggest that adding task-aware or nonlinear guidance to the PCA framework can make features more distinct and better match the goals of classification with dimensionality reduction.

Table 1 *Performance improvement of Hybrid PCA methods reported in related studies*

Study (Year)	Hybrid PCA Method	Dataset	Hybrid PCA Accuracy	Key Contribution of Hybrid Approach
Soltani et al. (2021) [22]	PCA–SVM	Real-world industrial refrigeration system sensor data	0.98–1.00	PCA reduces feature redundancy and noise, improving SVM robustness and computational efficiency for fault detection under varying operating conditions.
Yang et al. (2025) [23]	Kernel PCA–SVM	Driver EEG fatigue dataset (12 subjects)	0.968	KPCA captures nonlinear EEG patterns prior to SVM classification, leading to improved fatigue discrimination compared to linear feature representations.
Yang et al. (2025) [23]	Kernel PCA–LSTM	Driver EEG fatigue dataset (12 subjects)	0.971	KPCA provides nonlinear feature extraction, while LSTM models temporal dependencies in EEG signals for enhanced fatigue detection.
Ema & Shill (2020) [1]	PCA–ANN	Framingham Heart Disease dataset (UCI)	0.8535	PCA mitigates multicollinearity and reduces dimensionality before ANN classification, resulting in improved stability and predictive performance.

The proposed framework employs logistic regression not only as a final classifier but also as a supervisory mechanism during the principal component selection phase. Component relevance is not only based on variance; it is also based on how much it helps with classification performance. Experimental findings on the Mental Health dataset demonstrate a significant enhancement relative to traditional PCA, achieving an accuracy increase of 25.71%. This improvement shows that the proposed strategy works best on datasets where linear assumptions are only partially met. The method stays easy to understand and works well by adding lightweight supervised feedback to the PCA process. Table 1 shows a summary of some of the most common hybrid PCA methods that have been reported in the literature for different application areas. It shows how more and more people are using hybrid strategies to improve classification accuracy over baseline PCA.

2.2 The Proposed Logistic Regression-Guided Hybrid PCA

The proposed method enhances standard PCA by incorporating supervised guidance from Logistic Regression (LR) into the principal component selection process. Instead of selecting components based solely on explained variance or fixed k, we evaluate their contribution to classification performance. The workflow of the proposed Hybrid PCA–LR method is illustrated in Fig 1.

The process begins with the input dataset $X \in R^{n \times p}$, where n denotes the number of samples and p the number of features. All features are first standardized to ensure uniform scaling, which is critical for the effectiveness of Principal Component Analysis (PCA). To get principal components (PCs), which are linear combinations of the original features ranked by their explained variance, traditional PCA is used. The proposed method adds a supervised evaluation phase using Logistic Regression (LR) to the traditional PCA, which only chooses components based on eigenvalue magnitudes. At this point, each PC or group of PCs is evaluated based on how well it does at classifying using LR. The parts are ranked based on this, and the top-k parts that give the most accurate classification are chosen. The final classification model is then trained using these chosen parts. This LR-guided component selection makes sure that the dimensionality reduction process focuses on discriminative power instead of just variance. This improves classification performance, especially in datasets with weak nonlinear characteristics.

In contrast to traditional PCA, which retains components solely based on explained variance, our approach incorporates a supervised evaluation. We use k-fold cross-validation to test each principal component (PC) or group of PCs over and over again with a Logistic Regression classifier. The accuracy scores are used to put the components in order. The top-k subset that gives the best classification accuracy is kept for the final model. This makes sure that dimensionality reduction puts more emphasis on discriminative power than on variance alone.

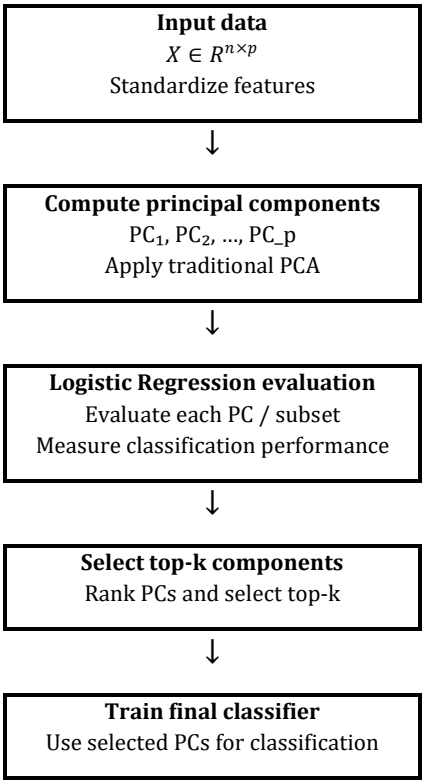

Fig. 1 *The workflow of the proposed Hybrid PCA–LR method*

Fig. 2 shows the overall workflow of the proposed PCA–LR method. It shows how logistic regression is used as a guide for choosing the main components. Algorithm 1 gives a short, step-by-step description of the same process to go along with this visual overview. It makes it clear how principal components are rated, ranked, and chosen before training the final classifier. This algorithmic formulation guarantees that the suggested method is reproducible and directly corresponds with the workflow diagram, while being unaffected by particular implementation specifics.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 79-90

Algorithm 1 Logistic Regression–Guided PCA (PCA–LR)

Input:

X: input dataset
y: class labels
k: number of principal components to be selected

Output:

Z: reduced feature representation

Steps:

1. Input the original dataset X
2. Standardize the feature values
3. Apply traditional PCA to compute principal components
4. Generate PCA-transformed features using all components
5. Evaluate each principal component or subset of components:
   5.1 Train Logistic Regression using selected PCs
   5.2 Measure classification performance
6. Rank the principal components based on classification performance
7. Select the top-k principal components
8. Train the final classifier using the selected components
9. Output the reduced feature representation Z

Fig. 2 *Algorithm of PCA_LR*

In the past, Principal Component Analysis (PCA) has mostly been used as an unsupervised dimensionality reduction method, and then Logistic Regression (LR) has been used as a classifier. In these kinds of frameworks, PCA components are usually chosen based on the total explained variance, without taking into account how well they predict the target variable. In contrast, our method uses Logistic Regression as a guide within PCA to choose the most discriminative principal components that directly improve classification performance. This LR-guided Hybrid PCA is very different from traditional PCA+LR pipelines because the regression model is used not only for classification but also to help choose components. This makes PCA better at capturing nonlinear and task-relevant structures.

This LR-guided Hybrid PCA is very different from the usual PCA+LR pipelines. In standard workflows, PCA is done first, and LR is only used after that to sort the data. On the other hand, our method includes LR in the selection phase, which turns it into a supervisor of component extraction. By making dimensionality reduction work with the classification goal, the method makes PCA better at finding weakly nonlinear and task-relevant structures. This integration signifies a conceptual transformation: PCA is no longer solely an unsupervised preprocessor but is now incorporated into a supervised feature selection loop.

3. Dataset Description and Linearity Consideration

This study employs four publicly accessible datasets chosen for their diversity in data types, complexity, and foundational data structures. These datasets comprise a combination of psychological, biomedical, signal-based, and morphological data, offering a solid foundation for assessing the linearity assumption in Principal Component Analysis (PCA). The objective is to evaluate the efficacy of traditional PCA across various data types and to assess the enhancements offered by the proposed Hybrid PCA–LR model, especially on datasets with nonlinear attributes.

We used both visual and quantitative methods to look at how linear each dataset was. First, we used exploratory scatterplots of the first two principal components (PC1 vs. PC2) to see how well the classes could be separated in lower dimensions [3]. Datasets exhibiting substantial class overlap in PCA space were deemed likely nonlinear. Second, we measure linear structure by calculating PCA reconstruction error over a range of kept components k. The rank-k projection in PCA is the least-squares solution that minimizes the sum of squared reconstruction errors. The remaining error is equal to the sum of the discarded eigenvalues. This means that a "elbow" (a sharp drop followed by a plateau) shows that most of the variance is in a low-dimensional linear subspace, while a gradual or inconsistent decline shows that there is no dominant linear subspace [26]. Moreover, since high-variance principal components (PCs) do not always yield the best classification results, choosing PCs based only on explained variance can lower accuracy. To avoid this, we use cross-validated Logistic Regression performance to evaluate the usefulness of each PC (or subset) and keep the top-*k* that gives the best results [24]. Additionally, we calculated R^2 scores between original and reconstructed data to quantify the amount of variance captured linearly [25].

This study utilized four benchmark datasets with diverse structural characteristics to assess the efficacy of the proposed Logistic Regression–Guided Hybrid PCA. We got the datasets from publicly available repositories: Mental Health [27], Breast Cancer [28], Fetal Health [29], and Pistachio [30].

The Mental Health dataset, which comes from answers to the DASS-42 questionnaire, has 42 original features and 39775 instances, and it shows partially nonlinear patterns [27]. The Breast Cancer dataset has 30 features and 569 instances, and it is mostly linear. It is often used as a standard for testing methods for reducing dimensionality [28]. The Fetal Health dataset has 21 features and more than 2000 instances, and the relationships between them are not too simple. Finally, the Pistachio dataset has 64 high-dimensional features taken from leaf images. This makes it a nonlinear classification problem [30]. Table 2 shows a summary of the datasets used in this study and what they are like.

Table 2 *Datasets used and linearity*

Dataset	Instance	Original Features	Linearity
Mental Health	39775	42	Partially Nonlinear
Breast Cancer	569	30	Mostly Linear
Fetal Health	2126	21	Partially Nonlinear
Pistachio	2148	64	Nonlinear

We used scatter plots of the feature distributions and PCA two-dimensional projections to see how linear each dataset was. This analysis showed that linear PCA probably doesn't work well on Pistachio, Mental Health, and Fetal Health data, but it does work better on Breast Cancer data. In the end, adding feature dimensionality to the dataset analysis gives us a better idea of how different PCA variants act on data with different levels of complexity and whether supervised guidance (through logistic regression) is always helpful across different dimensional scales.

To choose the right dimensionality reduction methods, you need to know how each dataset is structured. For this study, we used four datasets: Mental Health (DASS-42), Breast Cancer, Fetal Health, and Pistachio. Each had a different level of linearity and feature complexity. To substantiate these hypotheses, visual assessments were conducted employing both linear PCA and nonlinear Kernel PCA (KPCA), each condensing the feature space into two principal components. Figure 3 shows a sample result for the Mental Health dataset.

Figure 3 shows that the PCA projection creates overlapping clusters, which means that it may be hard to separate the classes in a straight line. The KPCA projection, on the other hand, shows a clearer separation between class groups. This shows that the data has nonlinear relationships that PCA doesn't do a good job of capturing.

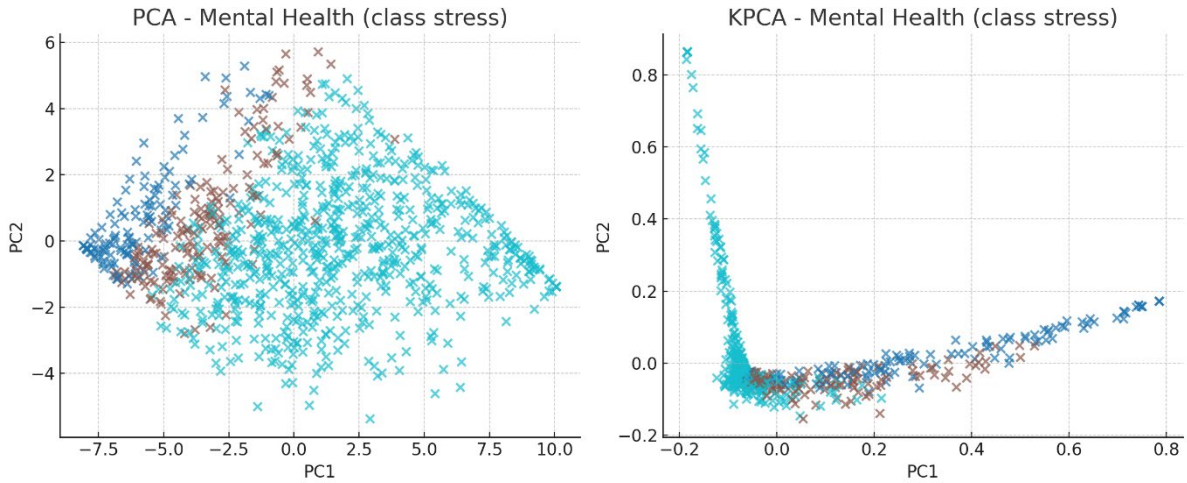

Fig. 3 *Two-dimensional visualization of the Mental Health dataset using PCA (left) and KPCA (right)*

The Fetal Health and Pistachio datasets exhibited analogous patterns, with the KPCA projection consistently enhancing cluster separability. On the other hand, the Breast Cancer dataset showed very little difference between the PCA and KPCA results, which supports its classification as mostly linear. This study shows that it is okay to use

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 79-90

nonlinear or guided methods (like KPCA and Hybrid PCA-LR) on datasets that have complex or partly nonlinear features.

The Breast Cancer dataset shows similar patterns in both PCA and KPCA projections, as shown in Fig. 4. This is different from the Mental Health dataset. The linear PCA space already shows that the class clusters (benign and malignant) are different from each other. This means that the data's structure is mostly linear. This backs up the fact that basic PCA worked well on this dataset and backs up its classification as "mostly linear." We did more PCA vs. KPCA visualizations on the Fetal Health and Pistachio datasets. PCA didn't do a good job of separating classes in the Fetal Health dataset, but KPCA did a better job of clustering, which showed that there were some nonlinear structures. The Pistachio dataset, which has a lot of features (64), was only moderately separable with PCA. But KPCA made the class clusters much clearer, which supports its classification as a nonlinear dataset. These visual comparisons show how important it is to use nonlinear dimensionality reduction or guided feature extraction methods, especially when PCA doesn't do a good job of separating classes.

The Mental Health dataset had the most clearly defined class clusters in the KPCA projection of the four datasets. This is probably because its class labels are well-defined and there are nonlinear relationships between the features of the questionnaire. The Results and Discussion section goes into more detail about what this phenomenon means and what it means for the future.

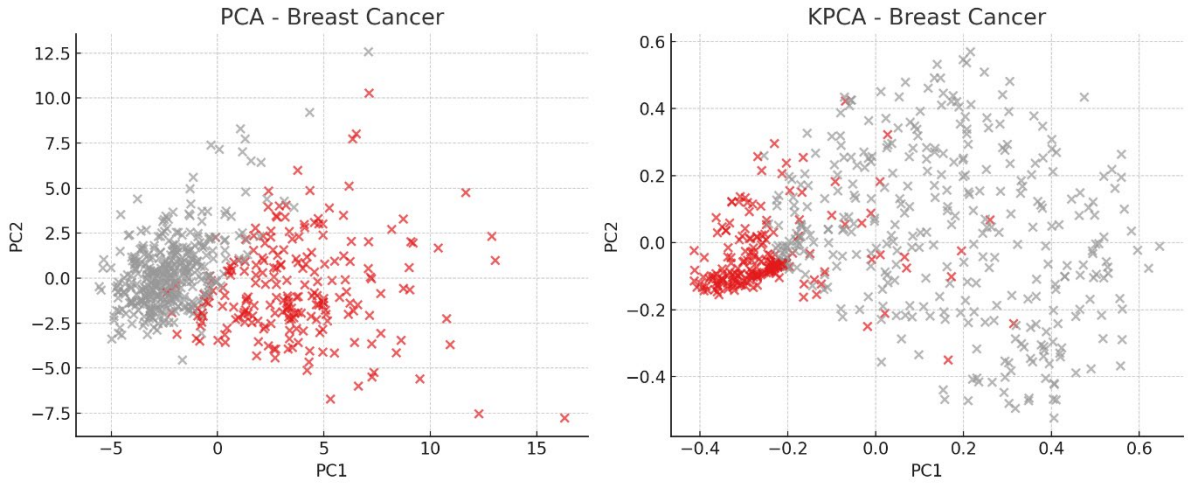

Fig. 4 *Two-dimensional visualization of the Breast Cancer dataset using PCA (left) and KPCA (right)*

4. Results and Discussion

This section shows how four different dimensionality reduction techniques—Principal Component Analysis (PCA), Kernel PCA (KPCA), Sparse PCA (SPCA), and the proposed Hybrid PCA based on Logistic Regression (PCA-LR)—perform when applied to four different datasets: Mental Health, Breast Cancer, Fetal Health, and Pistachio. We used four classification models to test each method: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. The goal is to find out how well each feature reduction method picks up useful information for classification tasks, especially when the datasets are not all linear or simple.

We compare three main techniques for reducing features based on principal components: Standard PCA, Kernel PCA (KPCA), and Sparse PCA (SPCA). Even though they all want to turn high-dimensional data into a lower-dimensional subspace, they each do it in a different way when it comes to capturing data variance and structure. Basic PCA is a classic linear transformation method that projects data onto orthogonal axes (principal components) to get the most variance. It assumes that the relationships between variables are linear and is still quick to compute. Kernel PCA (KPCA) adds kernel functions like Gaussian or polynomial kernels to the standard PCA to make it work in nonlinear domains. This process automatically maps the input data into a higher-dimensional feature space before doing PCA, which makes it possible to find nonlinear patterns. Recent advances like Deep Kernel PCA have made it easier to capture hierarchical and complex nonlinear structures. Sparse PCA (SPCA) adds sparsity constraints to the component loadings, which means that each principal component only depends on a small number of features. This makes high-dimensional data easier to understand and more reliable. PCA gives us a linear baseline, KPCA deals with nonlinear data relationships, and SPCA makes things easier to understand by forcing sparse representations. Together, these three methods give us a full picture of feature reduction across all types of data.

In the next few sections, show the accuracy results for each dataset in detail. Then, we'll compare the datasets and talk about how the structure and characteristics of each dataset affect how well dimensionality reduction

methods work. Table 3 shows that the proposed Hybrid PCA-LR method did much better than all other dimensionality reduction methods on all datasets. The Hybrid PCA-LR got 98.52% accuracy on the Mental Health dataset with only 40 selected features. SPCA and KPCA got 91.49% and 93.08%, respectively. Basic PCA, on the other hand, only got to 72.81% with 2 components. These results show that there are complicated nonlinear interactions in the data, which are better captured by guided or nonlinear transformations. Hybrid PCA-LR only got rid of two of the original 42 features, but it had the best accuracy at 98.52%. This shows that its strength is not in aggressively reducing the number of dimensions, but in picking the components that are most useful for classification. Hybrid PCA-LR uses label information to rearrange features so that it can find patterns that are different and may be nonlinear that other methods, like PCA, miss.

The differences between dimensionality reduction methods were less obvious but still important on the Breast Cancer dataset. Using only 9 features, Hybrid PCA-LR got the best accuracy of 97.37%. SPCA and KPCA were next, with 10 and 15 components, respectively, both reaching 96.49%. Basic PCA, even though it was linear, did well with 92.98% accuracy using only 2 components. This means that the dataset is mostly linear, but it could still use some small nonlinear or supervised improvements.

Table 3 *All datasets accuracy and optimal component number*

Classifier Methods	Feature Reduction Method	Mental Health		Breast Cancer		Fetal Health		Pistachio	
		1	2	1	2	1	2	1	2
Logistic Regression	PCA Basic	2	0.7073	2	0.9211	2	0.7756	2	0.8214
	PCA-LR	40	0.9852	9	0.9649	9	0.8427	20	0.9128
	SPCA	30	0.9057	10	0.9561	21	0.823	20	0.914
	KPCA	14	0.9172	15	0.9649	15	0.8115	19	0.8895
SVM	PCA Basic	2	0.7281	2	0.9298	2	0.7794	2	0.8093
	PCA-LR	40	0.984	9	0.9737	9	0.8652	20	0.9407
	SPCA	39	0.9149	10	0.9649	21	0.8614	20	0.9384
	KPCA	14	0.9308	15	0.9649	15	0.8652	19	0.9128
KNN	PCA Basic	2	0.6848	2	0.9211	2	0.7665	2	0.8214
	PCA-LR	40	0.9852	9	0.9561	9	0.8427	20	0.9302
	SPCA	39	0.9091	10	0.9386	21	0.8276	20	0.9337
	KPCA	14	0.8738	15	0.9474	15	0.8427	19	0.9198
Naive Bayes	PCA Basic	2	0.6839	2	0.886	2	0.7437	2	0.807
	PCA-LR	40	0.9852	9	0.9211	9	0.839	20	0.9233
	SPCA	39	0.9012	10	0.9211	21	0.8276	20	0.9163
	KPCA	14	0.8454	15	0.9211	15	0.8352	19	0.8942

1: optimal component number, 2: accuracy

The Hybrid PCA-LR and KPCA methods both got the best accuracy of 86.52% in the Fetal Health dataset. This was better than SPCA (86.14%) and Basic PCA (77.94%). It's interesting that both the Hybrid and kernel-based approaches were able to find useful latent features even though the label structure has a fuzzy boundary (the "suspect" class). This shows that the data is only slightly nonlinear and that using transformation methods that take this into account is a good idea.

For the Pistachio dataset, which has the most features (64 original features), both KPCA and Hybrid PCA-LR got 94.07% accuracy with 19–20 components. SPCA was very close behind with 93.84%, and Basic PCA was far behind with 82.14%. PCA was able to capture reasonable variance even though it was nonlinear because it had a lot of dimensions. However, nonlinear methods clearly made class separability better.

The number of principal components in each dimensionality reduction method was determined by methodological conventions and empirical optimization. In the PCA Basic method, the number of components was set to two. This is a common practice in exploratory data analysis and visualization, where the first two components are often used to show most of the data variance in a two-dimensional space [1] [4]. This setup also gives you a steady baseline to compare datasets and classifiers.

The component numbers for PCA-LR, KPCA, and SPCA, on the other hand, were found through experimentation. Each method captures data structure differently—linear, nonlinear, or sparse—so the best number of components depends on the unique features of each dataset. The suggested PCA-LR uses the

discriminative guidance from logistic regression weights to choose the number of components. This lets it keep the most useful features that help with classification accuracy. At the same time, KPCA and SPCA change their number of dimensions to fit nonlinear manifolds or sparse feature representations, respectively. So, the difference in the number of components between methods is not a sign of inconsistency; it shows how well each method can handle the data's structural complexity.

We used four classification models to compare the performance of each dimensionality reduction method on four different datasets. Table 4 shows the best accuracy that each method—Basic PCA, Kernel PCA (KPCA), Sparse PCA (SPCA), and the new Hybrid PCA-LR—could get, no matter which classifier was used. It also shows the model that got that result. This summary makes it easy to compare how well different methods work on datasets with different levels of linearity. The best method for each dataset is shown, along with whether nonlinear or guided approaches gave a big advantage over classical PCA.

Table 4 shows that the Hybrid PCA-LR method worked better than other ways to reduce dimensionality on most datasets. It got the most accurate results on both the Mental Health and Breast Cancer datasets, showing that it can find discriminative features even when only a few are removed.

Table 4 *Performance comparison of PCA, KPCA, SPCA, and Hybrid PCA-LR across datasets*

Dataset	Basic PCA (max)	KPCA (max)	SPCA (max)	Hybrid PCA-LR (max)	Best Model
Mental Health	0.7281 (SVM)	0.9308 (SVM)	0.9149 (SVM)	0.9852 (LogReg/KNN/NB)	Hybrid PCA-LR
Breast Cancer	0.9298 (SVM)	0.9649 (SVM)	0.9649 (SVM)	0.9737 (SVM)	Hybrid PCA-LR
Fetal Health	0.7794 (SVM)	0.8652 (SVM)	0.8614 (SVM)	0.8652 (SVM)	Tie: Hybrid/KPCA
Pistachio	0.8214 (KNN/LogReg)	0.9407 (SVM)	0.9384 (SVM)	0.9407 (SVM)	Tie: Hybrid/KPCA

There were patterns in both the Fetal Health and Pistachio datasets that were either somewhat or very nonlinear. Hybrid PCA-LR and KPCA both did the same job or almost the same job. This shows that both guided and kernel-based methods can work with nonlinear feature relationships. Basic PCA, on the other hand, did the worst on all datasets. This suggests that it can't find relationships that aren't linear. SPCA did better than Basic PCA on most datasets, but it didn't do better than KPCA or Hybrid PCA-LR on any dataset. These results indicate that supervised or nonlinear dimensionality reduction techniques are beneficial, particularly when dealing with real-world datasets that are not consistently linear.

5. Conclusion

By focusing on the selection of the principal component for dimensional reduction in the data, this study focuses on the evaluation of the accuracy performance of several dimensional reduction methods. The methods used as a comparison as an evaluation were Basic PCA (with fixed k=2), KPCA, SPCA, and the proposed method Hybrid PCA guided by Logistic Regression (PCA-LR). In the proposed method, LR is used as a determinant of best k where in Basic PCA k is determined by fixed k or variance value. Because PCA is conceptually a linear projection, the dataset used as a tester is four datasets (Mental Health, Breast Cancer, Fetal Health, Pistachio) that have varying linearity and are expected to see the performance of basic PCA, KPCA, SPCA and PCA-LR.

The experimental results that are highlighted in this study are:

- **Hybrid PCA-LR showed better consistent results than other methods,** obtaining the best accuracy scores on three of the four datasets.
- On the **Mental Health dataset**, despite reducing only two features, Hybrid PCA-LR achieved a significant accuracy boost (98.52%), this shows that the retained features have an important relevance to the reduced features.
- In **Fetal Health** and **Pistachio**, the achieved accuracy number shown that both Hybrid PCA-LR and KPCA yielded equally high performance, confirming basic PCA cannot capture the presence of nonlinear.
- In all four test datasets, it was shown that **Basic PCA underperformed**, especially in nonlinear settings, while **SPCA showed moderate improvement** but the accuracy performance never surpassed the other advanced methods.

With the results that are highlighted in this study, it is found that in the real world the data found is often complex and nonlinear relationships between features in the data that cause the performance of PCA as a dimensional reduction method to be limited because PCA is a purely linear dimensional reduction method. Therefore, integrating supervised learning (as in Hybrid PCA-LR) or nonlinear mapping (as in KPCA) is essential to enhance feature representation and classification outcomes.

Future research is expected to explore more test datasets; more supervised model methods used for best *k* selection in Hybrid PCA. By adding datasets with varying linearity and complexity, it is hoped that it can test the reliability of hybrid PCA. Combining with deep learning methods such as the autoencoder method is also expected to improve the ability of this proposed method to capture nonlinear patterns in data.

Acknowledgement

The authors would like to express their gratitude to the Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Lamongan, and the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, for their continuous support and valuable resources throughout the completion of this research. AI tools were used exclusively for proofreading and language refinement. All methodological design, experiments, analyses, and interpretations were conducted entirely by the authors.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Retno W., Noraziahtul H., Asem K.;* ***data collection:*** *Retno W.;* ***analysis and interpretation of results:*** *Retno W., Noraziahtul H.;* ***draft manuscript preparation:*** *Retno W. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Ema, R. R., & Shill, P. C. (2020). Integration of Fuzzy C-Means and Artificial Neural Network with Principle Component Analysis for Heart Disease Prediction. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225366

[2] Razzak, I., A. Hameed, I., & Xu, G. (2019). Robust Sparse Representation and Multiclass Support Matrix Machines for the Classification of Motor Imagery EEG Signals. IEEE Journal of Translational Engineering in Health and Medicine, 7, 1–8. https://doi.org/10.1109/JTEHM.2019.2942017

[3] Reiss, M., & Wahl, M. (2020). Nonasymptotic upper bounds for the reconstruction error of PCA. The Annals of Statistics, 48(2), 1098–1123. https://doi.org/10.1214/19-AOS1839

[4] Zheng, X., & Rakovski, C. (2021). On the Application of Principal Component Analysis to Classification Problems. Data Science Journal, 20(1), 26–26. https://doi.org/10.5334/dsj-2021-026

[5] Bao, Z., Ding, X., Wang, J., & Wang, K. (2022). Statistical inference for principal components of spiked covariance matrices. The Annals of Statistics, 50(2). https://doi.org/10.1214/21-AOS2143

[6] Alomari, O. A., Elnagar, A., Afyouni, I., Shahin, I., Nassif, A. B., Hashem, I. A., & Tubishat, M. (2022). Hybrid Feature Selection Based on Principal Component Analysis and Grey Wolf Optimizer Algorithm for Arabic News Article Classification. IEEE Access, 10, 121816–121830. https://doi.org/10.1109/ACCESS.2022.3222516

[7] Nahiduzzaman, Md., Goni, Md. O. F., Anower, Md. S., Islam, Md. R., Ahsan, M., Haider, J., Gurusamy, S., Hassan, R., & Islam, Md. R. (2021). A Novel Method for Multivariant Pneumonia Classification Based on Hybrid CNN-PCA Based Feature Extraction Using Extreme Learning Machine With CXR Images. IEEE Access, 9, 147512–147526. https://doi.org/10.1109/ACCESS.2021.3123782

[8] Atteia, G., Alnashwan, R., & Hassan, M. (2023). Hybrid Feature-Learning-Based PSO-PCA Feature Engineering Approach for Blood Cancer Classification. Diagnostics, 13(16), 2672. https://doi.org/10.3390/diagnostics13162672

[9] Song, G., Ai, Z., Zhang, G., Peng, Y., Wang, W., & Yan, Y. (2022). Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. Building and Environment, 212, 108790. https://doi.org/10.1016/j.buildenv.2022.108790

[10] John, A., Isnin, I. F. B., Madni, S. H. H., & Muchtar, F. B. (2024). Enhanced intrusion detection model based on principal component analysis and variable ensemble machine learning algorithm. Intelligent Systems with Applications, 24, 200442. https://doi.org/10.1016/j.iswa.2024.200442

[11] Huang, T., Li, J., & Zhang, W. (2020). Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. BMC Medical Research Methodology, 20(1), 99. https://doi.org/10.1186/s12874-020-00989-x

[12] Srijiranon, K., Lertratanakham, Y., & Tanantong, T. (2022). A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. Applied Sciences, 12(21), 10823. https://doi.org/10.3390/app122110823

[13] Li, X., Jia, R., Zhang, R., Yang, S., & Chen, G. (2022). A KPCA-BRANN based data-driven approach to model corrosion degradation of subsea oil pipelines. Reliability Engineering & System Safety, 219, 108231. https://doi.org/10.1016/j.ress.2021.108231

[14] Scholkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5), 1299–1319. https://doi.org/10.1162/089976698300017467

[15] Mayapada, R., Tinungki, G. M., & Sunusi, N. (2019). Penerapan Sparse Principal Component Analysis dalam Menghasilkan Matriks Loading yang Sparse. 15(2), 44–54. https://doi.org/10.20956/jmsk.v15i2.5713

[16] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. https://doi.org/10.1198/106186006X113430

[17] Khan, A. R., Ahamad, S. S., Mishra, S., Khan, M. A. R., Sharma, S. K., AlEnizi, A., Alfarraj, O., Alowaidi, M., & Kumar, M. (2024). FinSafeNet: Securing digital transactions using optimized deep learning and multi-kernel PCA(MKPCA) with Nyström approximation. Scientific Reports, 14(1), 26853. https://doi.org/10.1038/s41598-024-76214-2

[18] Wang, Z., van der Laan, T., & Usman, M. (2025). Self-Adaptive Quantum Kernel Principal Component Analysis for Compact Readout of Chemiresistive Sensor Arrays. Advanced Science, 12(15), e2411573–e2411573. https://doi.org/10.1002/advs.202411573

[19] Tsang, G., Zhou, S.-M., & Xie, X. (2021). Modeling Large Sparse Data for Feature Selection: Hospital Admission Predictions of the Dementia Patients Using Primary Care Electronic Health Records. IEEE Journal of Translational Engineering in Health and Medicine, 9, 1–13. https://doi.org/10.1109/JTEHM.2020.3040236

[20] Qi, L., Wang, W., Wu, T., Zhu, L., He, L., & Wang, X. (2021). Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis. Frontiers in Genetics, 12, 607817. https://doi.org/10.3389/fgene.2021.607817

[21] Li, X., Qiao, Y., Duan, L., & Miao, J. (2025). Epilepsy EEG signals classification based on sparse principal component logistic regression model. Computer Methods in Biomechanics and Biomedical Engineering, 28(8), 1295–1303. https://doi.org/10.1080/10255842.2024.2321991

[22] Soltani, Z., Kjaer Soerensen, K., Leth, J., & Dimon Bendtsen, J. (2021). Robustness analysis of PCA-SVM model used for fault detection in supermarket refrigeration systems. 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 1–6. https://doi.org/10.1109/ICECCE52056.2021.9514086

[23] Yang, B., Ding, Y., Cui, C., & Guo, T. (2025). A hybrid kernel principal component analysis and long short-term memory approach for EEG signal-based fatigue evaluation. Alexandria Engineering Journal, 131, 209–217. https://doi.org/10.1016/j.aej.2025.09.066

[24] Zheng, X., & Rakovski, C. (2021). On the Application of Principal Component Analysis to Classification Problems. Data Science Journal, 20(1), 26–26. https://doi.org/10.5334/dsj-2021-026

[25] Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). Water Quality Evaluation and Analysis by Integrating Statistical and Machine Learning Approaches. Algorithms, 18(8), 494. https://doi.org/10.3390/a18080494

[26] Bao, Z., Ding, X., Wang, J., & Wang, K. (2022). Statistical inference for principal components of spiked covariance matrices. The Annals of Statistics, 50(2). https://doi.org/10.1214/21-AOS2143

[27] OpenPsychometrics. (2025). Open Psychometrics Raw Data. OpenPsychometrics. https://openpsychometrics.org/_rawdata/

[28] Kaggle. (2025a). Breast Cancer Classification (Kaggle). Kaggle. https://www.kaggle.com/datasets/benyaengineering/breast-cancer-classification

[29] Kaggle. (2025b). Fetal Health Classification (Kaggle). Kaggle. https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

[30] Kaggle. (2025c). Pistachio Dataset (Kaggle). Kaggle. https://www.kaggle.com/datasets/muratkokludataset/pistachio-dataset