Journal of Autonomous Intelligence (2023) Volume 6 Issue 2
doi: 10.32629/jai.v6i2.586

ORIGINAL RESEARCH ARTICLE

PCSVD: A hybrid feature extraction technique based on principal component analysis and singular value decomposition

Vineeta Gulati*, Neeraj Raheja

Department of Computer Science and Engineering, Maharishi Markandeshwar (Deemed to be University), Ambala 133207, India

*** Corresponding author:** Vineeta Gulati, vineetaarora04@gmail.com

ABSTRACT

Feature extraction plays an important role in accurate preprocessing and real-world applications. High-dimensional features in the data have a significant impact on the machine learning classification system. Relevant feature extraction is a fundamental step not only to reduce the dimensionality but also to improve the performance of the classifier. In this paper, the author proposes a hybrid dimensionality reduction technique using principal component analysis (PCA) and singular value decomposition (SVD) in a machine classification system with a support vector classifier (SVC). To evaluate the performance of PCSVD, the results are compared without using feature extraction techniques or with existing methods of independent component analysis (ICA), PCA, linear discriminant analysis (LDA), and SVD. In addition, the efficiency of the PCSVD method is measured on an increased scale of 1.54% accuracy, 2.70% sensitivity, 3.71% specificity, and 3.58% precision. In addition, reduce the 15% dimensionality and 40.60% RMSE, which are better than existing techniques found in the literature.

Keywords: support vector classifier; machine learning; independent component analysis; linear discriminant analysis; chronic kidney disease dataset; dimensionality reduction

ARTICLE INFO

Received: 3 April, 2023
Accepted: 1 June, 2023
Available online: 18 July, 2023

1. Introduction

Machine learning is a rapidly emerging research area with a bright future. This part of AI science plays one of the most important roles in many aspects of our lives, including healthcare. For example, the number of patients in hospitals is increasing, which means that it is becoming increasingly difficult to analyze and collect all patient data[1]. Machine learning provides an excellent solution to this problem, facilitating automated data analysis and strengthening the healthcare system[2]. Applications of machine learning in healthcare include diagnosis, drug availability, and drug customization. In machine learning problems, there is not only one factor to predict the outcome of the disease but many factors that jointly contribute to the final prediction[3]. One cannot conclude a prediction based on the analysis of a single factor. The whole factor, whether small or large, gives the final prediction. It depends entirely on the features it contains. The more unnecessary features there are, the more difficult it is to predict the result. It may happen that some of the features are similar or redundant. In this case, the dimensionality reduction algorithm can be of great help. The dimensionality reduction algorithm is an unsupervised learning technique[4]. It analyzes the lower dimensions of the numerical input data and preserves the important relationships

in the data. When a given data set has so many features, a machine learning model starts to perform poorly instead of making it more complex[5]. The reason is that the information lost by rejecting some features is compensated for by accurate mapping in low-dimensional space. When a model has fewer features, its performance improves.

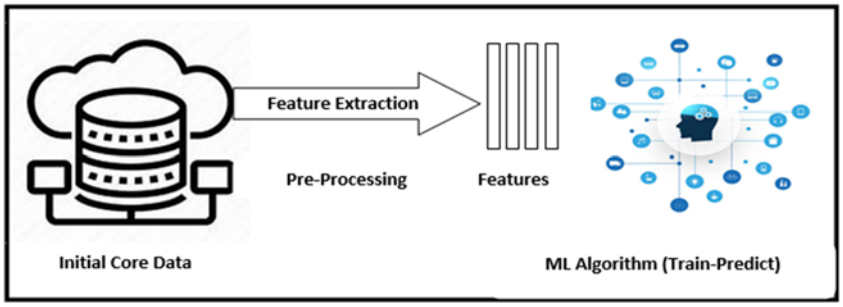

Figure 1. Modeling steps of machine learning classification system.

Extraction is a process by which distractions and inconsistencies are removed from a data set. This is done so that downstream classifiers can work better. In this way, various applications can be performed[6]. The main task of a feature extractor is to simply represent the data according to its class labels. Feature extraction is used in machine learning and deep learning models. It is the process of converting raw data into numerical features[7]. This is used to obtain the information in the original dataset. In simpler terms, feature extraction, as the name suggests, is the process of creating a subset of features by merging existing features. The result of feature extraction is new features that are a linear combination of the existing features. The new features that result from feature extraction have completely different values compared to the original features. Our main goal in feature extraction is to obtain fewer features because machine learning models can provide more accurate data when there are fewer features[8]. These few features can now be easily used to retrieve data. Feature extraction is an extremely helpful method to select a particular variable and combine it with some other variables to reduce the amount of data. In this case, the result is evaluated using techniques such as recall and precision. Therefore, feature extraction plays a fundamental role in this process. In some machine learning projects, effective feature extraction can also be used to solve underfitting and overfitting problems[6]. Since only necessary and important data needs to be extracted, feature extraction is extremely beneficial because it provides a clear and improved visualization of the data. In addition, training a model can be made more efficient when feature extraction methods are used[6].

Researcher's contribution

- The aim of this work is to find a more effective dimensionality reduction technique among all available techniques. The authors first investigated and analyzed popular feature extraction methods, including PCA, ICA, SVD, and LDA.

- A hybrid PCSVD method was proposed to reduce the dimensionality of the chronic kidney disease dataset.

- The authors compared the proposed work with previously developed feature extraction methods and the full feature set. SVC was used for the classification system and achieved 98.75% accuracy.

- Finally, the new PCSVD method outperformed previously reported methods in the literature by reducing dimensionality by 15%.

The subsequent sections of this paper are organized as follows: Section 2 reviews related work. Section 3 explains the methodology for the proposed approach. The results of the experiment are discussed in Section 4, and Section 5 provides conclusions and outlooks.

2. Related work

Table 1 shows the feature extraction techniques previously implemented by different researchers using different chronic disease datasets as well as the standard chronic kidney disease dataset.

Table 1. Feature extraction techniques used by various researchers.

Author	Dataset	Feature extraction/selection method	Classifier
Islam et al.[8]	Standard chronic kidney disease dataset	PCA	XGBoost–98.3%
Venkatesan et al.[9]	Standard chronic kidney disease dataset	14 features are extracted through recursive feature elimination technique	XGBoost–98.9%
Swain et al.[10]	Standard chronic kidney disease dataset	Chi squared test	SVM: 99.3% with 9 selected features
Ebiaredoh-Mienye et al.[11]	Standard chronic kidney disease dataset	Information gain	AdaBoost: 99.8% with 18 selected features
Jerop and Segera[12]	Standard chronic kidney disease dataset, respiratory diseases, breast cancer, heart disease	PCA, LDA	SVM–95.94%
Navaneeth and Suchetha[13]	Standard chronic kidney disease dataset	SVD, PCA	SVD-SVM–87.32% PCA-KNN–84.32%
Inayatullah and Qayyurn[14]	Standard chronic kidney disease dataset	Cross-validation measures are implemented on the dataset	KNN–92%, SVM–97% NB–98%
Reddy and Devi[15]	CKD dataset	Analyze DT, KNN, SVM, and SGD classifier with the 16 features of the dataset	KNN–69% LR–75% SVM–75% SGD–75%
Jain and Singh[16]	Standard chronic kidney disease dataset	PCA, ReliefF and Hybrid of PCA and ReliefF	SVM–97.4%
Gu[17]	Standard chronic kidney disease dataset	PCA, TSNE, SVD	SVM–84%, ANN–93%
Gharibdousti et al.[18]	Standard chronic kidney disease dataset	Analyze the accuracy of classifiers with different-different features, and achieve the best result with nine features	NB–96.7% LR–98% SVM–98%
Bouzalmat et al.[19]	ATT face database and the Indian face database (IFD)	PCA, LDA, ICA	SVM PCA–90.24% LDA–93.9% ICA–91%
Reza and Ma[20]	Wisconsin breast cancer, and wine data of UCI (University of California, Irvine) database, namely, data collected from Australian crabs	ICA, PCA, LDA	SVM, NB PCA–65.0%, 65.5% LDA–61.5%, 66.5% ICA–62.0%, 67.5%
Ramachandran et al.[21]	Different structures or multidimensional Datasets	PCA, ICA, SVD, LDA	Presented the comparative study of linear and non-linear dimensionality reduction techniques based on different parameters
Li et al.[22]	Real user consumption records	Feature extraction methods	Discussed the impact of feature extraction techniques on machine learning algorithms

3. Material and methods

The data set used for the experiment is described in detail in this section, after which the process of the proposed dimensionality reduction methods is presented. Then, the methods used to classify the data and the machine learning performance indicators are presented.

3.1. Description of the dataset

The UCI (University of California, Irvine) standard chronic kidney disease dataset contains 400 instances with 25 features (11 numeric, 14 nominal) used for the experimental work[21]. The authors analyze each feature based on variance ratio and information gain to determine how many distinct features are symmetric and non-negative and how they are compressed into a few components with respect to each original feature, as shown in **Table 2**. **Figure 2** and **Figure 3** show the comparative analysis of each feature in the dataset.

Table 2. Information gain and variance ratio of features in chronic kidney disease dataset.

Sr. No.	Feature name	Feature code	Information gain	Variance ratio
1	Hemoglobin	Hemo	0.431195	0.031324
2	Specific gravity	Sg	0.367656	0.052535
3	Albumin	Al	0.360552	0.049486
4	Packed cell volume	Pcv	0.353455	0.018299
5	Serum creatinine	Sc	0.345120	0.037414
6	Red blood cell count	Rc	0.304990	0.013780
7	Sodium	Sod	0.273210	0.033138
8	Hypertension	Htn	0.232037	0.013597
9	Diabetes mellitus	Dm	0.213579	0.012601
10	Potassium	Pot	0.208297	0.031963
11	Blood glucose random	Bgr	0.184061	0.043435
12	Blood urea	Bu	0.158384	0.039185
13	Sugar	Su	0.140848	0.046182
14	Blood pressure	Bp	0.139370	0.066808
15	Anemia	Ane	0.138195	0.007469
16	Appetite	Appet	0.129965	0.010536
17	White blood cell count	Wc	0.092783	0.017343
18	Pus cell	Pc	0.081545	0.023800
19	Pedal edema	Pe	0.074148	0.009113
20	Age	Age	0.061907	0.071505
21	Pus cell clumps	Pcc	0.054137	0.022586
22	Red blood cells	Rbc	0.053046	0.027174
23	Bacteria	Ba	0.038314	0.020170
24	Coronary artery disease	Cad	0.000000	0.011403

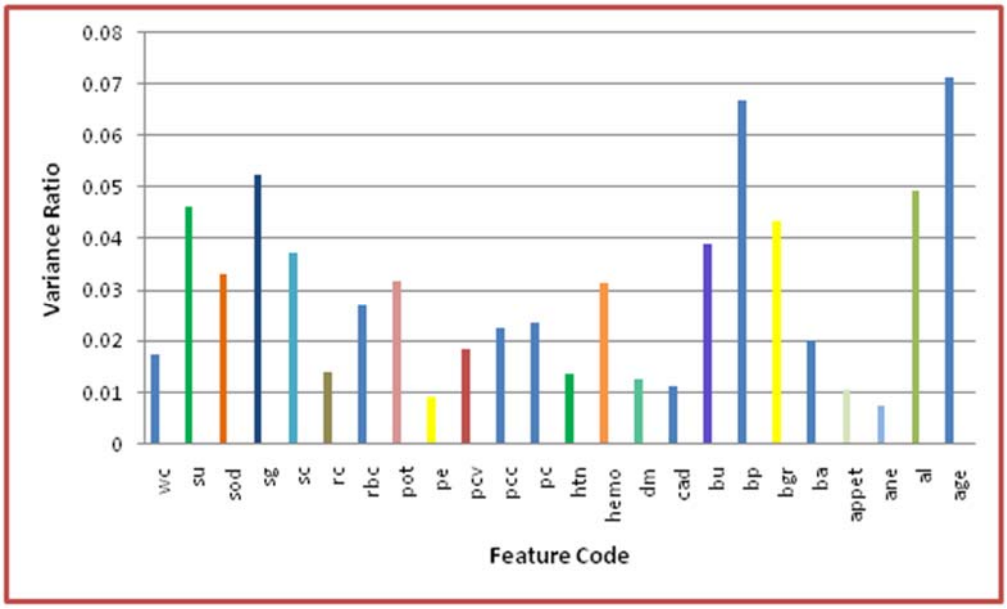

Figure 2. Comparative analysis of information gain in chronic kidney disease dataset.

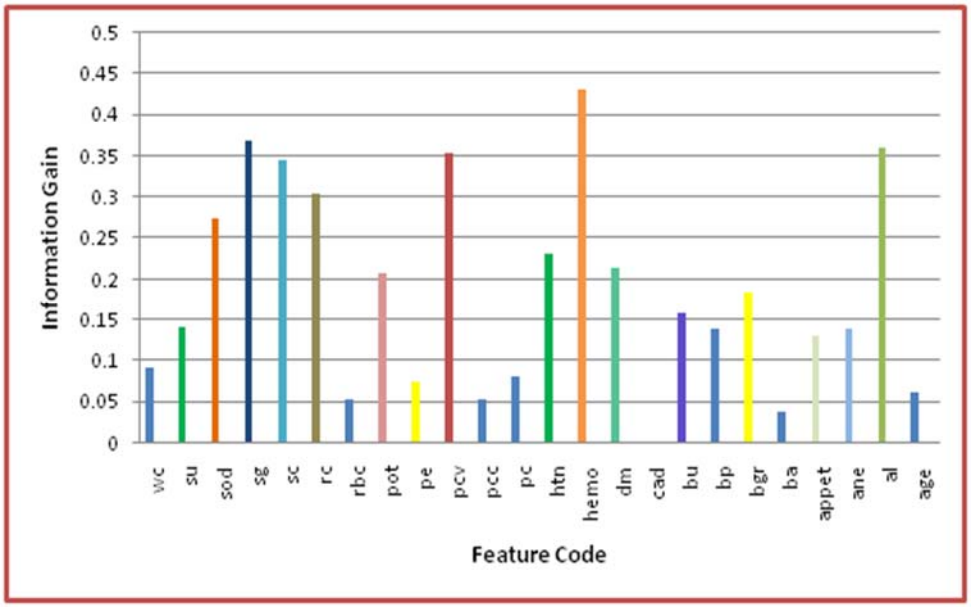

Figure 3. Comparative analysis of variance ratio in chronic kidney disease dataset.

3.2. Methodology

The proposed technique PCSVD is a hybrid of the two most commonly used feature extraction algorithms, principal component analysis (PCA) and singular value decomposition (SVD), to generate the best PCSVD components or features that not only reduce dimensionality but also increase the performance graph of the classification system. First, we compute a standardized matrix to create the covariance matrix using principal component analysis. Then we perform matrix decomposition using the singular value decomposition method to find the best eigenvalues, not only using a single technique, but also using the hybrid concept to reduce the dimensionality and improve the performance. **Figure 4** and **Figure 5** show the block diagram and flowchart of the proposed work, respectively.

Algorithm 1 PCSVD

1: {
2: Perform Hold_out Validation and split the dataset into x and y
3: training set: $x^{(1)}$, $x^{(2)}$ .... $x^{(n)}$
4: Number of components= x_size
5: P (number of variances in the components) = number of features
6: For i in the range (1, P):
7: Calculation of mean and standardized matrix Z
8: $Z = X^t.N.X$//generate covariance matrix
9: $X = U.S.V^T$//Perform singular decomposition, U & V are matrix of eigenvectors, S as diagonal matrix of the eigenvalues
10: $P^* = S_1, S_2, S_3 \ldots S_P$//Sorting the eigenvalues $\sum$ in descending order according to their eigenvector
11: PCSVD = $Z.P^*$//Generate the updated version of x, each combination independent of the other
12: Select the larger variance values and ignore the smaller ones to reduce dimensionality
13: }

The basic principle of the PCSVD algorithm is described here:

Step 1: Compute the mean and standardization for the training set: $x^{(1)}$, $x^{(2)}$ .... $x^{(n)}$ having n number of samples and p number of variances.

Step 2: Generate the covariance matrix $Z = X^t.N.X$, a combination of eigenvectors and eigenvalues. Categorize as a symmetric matrix that can be diagonalized and decomposed like SVD decomposition.

Step 3: Perform singular decomposition $X = U.S.V^T$, U and V are left and right singular matrices of eigenvectors, and S is the diagonal matrix of the eigenvalues.

Step 4: Sort the eigenvalues S in descending order according to their eigenvector, $P^* = S_1, S_2, S_3 \ldots S_p$.

Step 5: Generate the updated version of X, each combination independent of the other, through PCSVD $= Z.P^*$.

Step 6: Select the larger variance values and ignore the smaller ones to reduce dimensionality.

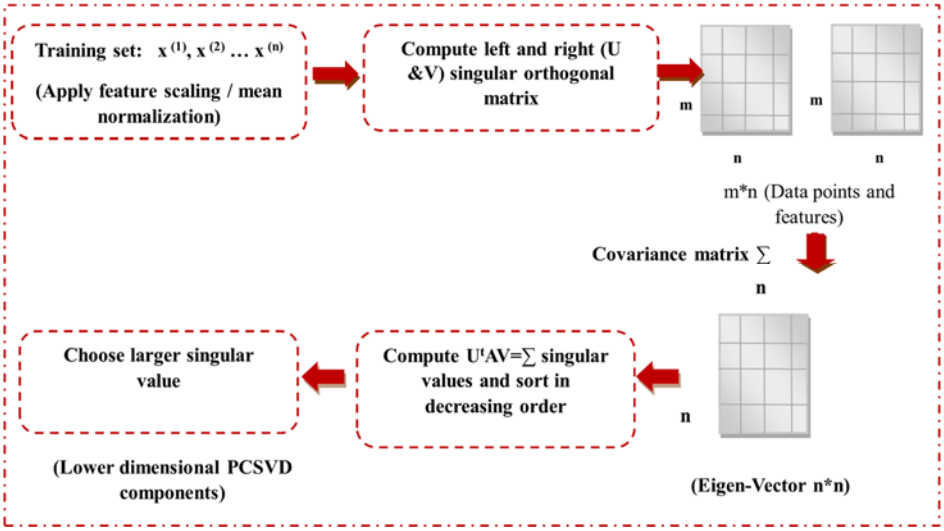

Figure 4. Block diagram of PCSVD technique.

Description of commonly used feature selection algorithms is as follows:

Principal component analysis (PCA): PCA is a dimensionality reduction method primarily used to minimize the dimensionality of large data sets. It produces high features by combining more than two features[23]. In simpler terms, a principal component is a normalized linear combination of native features. It maximizes the difference in the data. There are many applications for PCA, including anomaly and outlier detection. Thus, it is an approach to detect a pattern in data and examine the data to highlight its similarities and differences. First, PCA calculates the mean of all variables present in the data set to pass the data through the origin and subtract the mean values[24].

Training set: $x^{(1)}, x^{(2)} \ldots x^{(n)}$

- Preprocessing (feature scaling/mean normalization)

$\mu j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$ (Replace each $x_j^{(i)}$ with $x_j - \mu_j$)

- Calculate the covariance matrix; it is a measure between two dimensions and shows how the two variables vary together: $\sum = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})$ (If the non-diagonal elements in this covariance are positive, we can suppose that x and y variables rise together.)

- Compute the eigenvector of the matrix.

The most important eigenvector has the direction in which the variable is strongly correlated, so the eigenvector with the highest eigenvalues is chosen as the PCA, and the other dimension values are ignored.

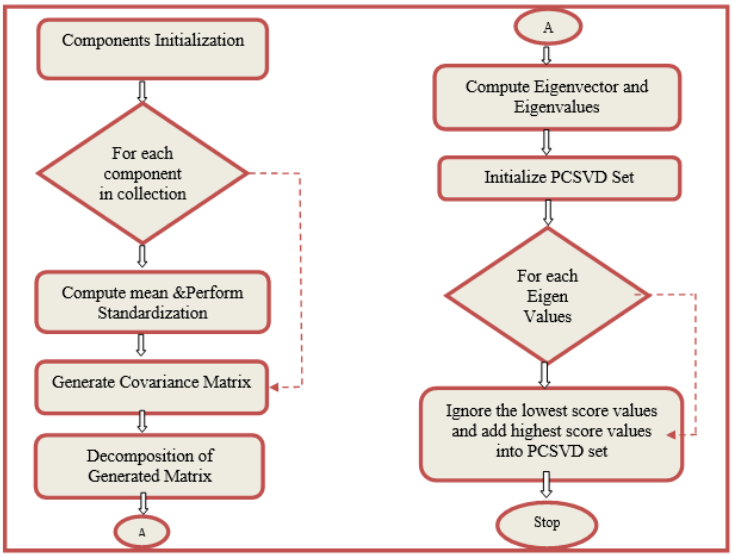

Figure 5. Flow chart of proposed PCSVD technique.

Independent component analysis (ICA): Independent component analysis is another type of linear dimensionality technique. Its focus is on accurately identifying each independent component. By focusing on each component in a mixture, useful data is collected, and noise in the mixture is removed[25]. Let us take an example of independent component analysis. Suppose there is audio material with the sound of two birds that needs to be analyzed. For this purpose, the ICA technique is effectively used. Transform a set of vectors into a maximally independent set. Generate features in linearly separable form, e.g., $X = As$, where A is in matrix form and the goal of ICA is to find a segregation matrix W that approximates A–1.

Assumptions: The independent components are:

- Statistically independent $p(x, y) = p(x)p(y)$
- Non-Gaussian $x_i = \sum_j a_{ijs_j}$ or $x_i = \sum_j W_{ij} S_j$

Goal: For X, solve W such that $\{S_i\}$ is maximally independent.

- The whitening of the data set involves the removal of any kind of correlation. Eigenvalue decomposition: $x = ED^{\frac{-1}{2}}E^T$
- Determine W for each 1 to n number of components up to:

$$W_P^T W_P + 1 \approx 1$$

Linear discriminant analysis (LDA): Linear discriminant analysis is another feature extraction method used specifically for dimensionality reduction. In most cases, it is used to solve classification problems[26]. The result of linear discriminant analysis leads to a maximum partition of classes. The goal is to reduce the n-dimensional feature space to a small subspace k ($k \leq n - 1$) while preserving the unequal class information[6].

- Compute within class scatter matrix: $Sw = S_1 + S_2$ (S_1 is the covariance matrix for class $c1$ and S_2 is that for class $c2$).

$S_1 = \sum_{x \in c1}(x - \mu 1)(x - \mu i)^T$

$S_2 = \sum_{x \in c2}(x - \mu 2)(x - \mu i)^T$

- Calculate the scatter matrix between the classes:

$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$

- Find the best LDA projection vector, as in principal component analysis, using the eigenvector with the largest eigenvalues.

$Y = W^T X$ (W is the projection vector and X is the input data)

Singular value decomposition (SVD): This is a widely used method to decompose a matrix into different submatrices. The result of singular value decomposition gives many useful properties to the original matrix[27]. SVD is a matrix factorization technique that reduces the number of features in a data set by reducing the freedom of dimension from n-dimensions to k-dimensions ($k < n$). The SVD is applicable to any real-valued matrix. It is more suitable to measure the similarity between features by examining the similarity pattern contained in the word "Co-Occurrence". It is always possible to decompose a real-valued matrix A into: $A = U \in V^t$. U and V are the orthogonal matrix: $A^t = A^t A = I$, where U is the left singular matrix and V is the right singular matrix. $\sum$ is the diagonal matrix with singular values ($r \times r$). Thus, a given data matrix is decomposed into a long, sparse, and diagonal matrix because decomposing such $a\ d$ matrix enables the discovery of latent, hidden features that can help in classification or clustering[28].

Support vector classifier (SVC): The proposed technique is PCSVD compared to linear feature extraction techniques for the classification process using an SVC classifier. SVC is used for machine learning classification problems and is the basis of support vector machine (SVM), where the classes of the dataset are divided into two classes and an SVM prediction rule with hyperplane is followed[29].

$$Y = \int \begin{matrix} 0\ if\ W^T + b < 0 \\ 1\ if\ W^T + b > 0 \end{matrix}$$

Find the hyperplane with the maximum separation distance of the training data. The method is often referred to as C-SVC because the regularization factor C is used for optimization, which governs the tradeoff between a smooth evaluation margin and an appropriate classification of the training points[30]. Here, C is the outcome-determining parameter; if C is small, then the consequence for misclassified points is low, so an evaluation frontier with a large margin is chosen; otherwise, in the case of a high penalty, an evaluation frontier with a smaller margin is chosen.

Performance metrics: After model fitting, the model must be evaluated so that we can use metrics to find out how effective the model is. These can vary depending on the task being handled by the machine learning algorithms[31]. When estimating the performance of any machine learning model, we need to consider which database to use to estimate the model's performance. Since the model turns to the training set during the training process of machine learning, the output of the model can only be predicted if its performance is evaluated based on the training data[32]. Thus, to estimate the generalization error, the machine learning model must be evaluated on data that it has not yet seen. Therefore, it is ideal to evaluate the performance of the ML (machine learning) model using test data. Below are performance metrics for evaluating the effect of feature extraction techniques[33]:

- True positive (TP): These are the items that are correctly classified as "yes" or "have the disease"
- True negative (TN): These are the items that are classified as not having the disease
- False positive (FP): The model predicted "yes" but the patient does not have the disease
- False negative (FN): The model predicted no, but the patient has the disease
- Sensitivity (TPR) measures the rate of improvement of positives, and complementary specificity (FPR) measures the rate of improvement of negatives.

$$\text{Specificity} = \frac{TN}{FP+TN},\ \text{Sensitivity} = \frac{TP}{FN+TP}$$

- Accuracy is one of the most common evaluation metrics; it measures the number of correct calculations relative to all predictions made[11].

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

- Root mean square error (RMSE) represents the deviation between the model prediction and the actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2}$$

- The area under the curve (AUC) is used for binary classification and measures which categorization method is better for classification based on thresholds.

4. Results and discussion

The results of the experiment are presented and discussed in this section. The chronic kidney disease dataset from UCI machine learning database is used for this study. The proposed work was implemented in the Jupyter Notebook integrated development environment using the Anaconda Navigator platform of the Python programming language. The authors begin by pre-processing a machine learning model by imputing missing values with their mean; categorical or nominal values are converted to numeric form, as datasets may contain missing values or irrelevant or redundant information. Dimensionality reduction follows to eliminate all irrelevant information. For this purpose, the dataset is subjected to the feature extraction techniques PCA, ICA, LDA, SVD and the proposed hybrid formation SVPC-LDA.

Results of features extracted by different feature extraction methods

- Feature extraction by principal component analysis (PCA): This is one of the most used feature extraction techniques, which preserves as much information as possible and extracts features that are uncorrelated after assignment, cannot be further reduced, and have a large variance. This technique is based on correlation and variance ratios and generates eight new PCA components or features that are independent of each other.

- Feature extraction by singular value decomposition (SVD): This is an exact decomposition of the original matrix using orthogonal features that contain important, non-redundant information about the observations and uses Latent Semantic Analysis, which assumes that the words that are most frequent in one topic are less frequent in the other topic. This technique generates eight SVD components by selecting vectors corresponding to the largest singular values.

- Feature extraction by independent component analysis (ICA): It is another type of linear dimensionality technique. Its main focus is on the accurate identification of each independent component. By focusing on each component in a mixture, collecting useful data, and removing noise in the mixture, eight ICA components are generated.

- Feature extraction by linear discriminant analysis (LDA): It finds arcs that exploit partitioning between multiple classes; the goal is to evolve the (n-dimensional) feature space to a less significant subspace k ($k = n - 1$), preserving the class-unequal information[24]. In this technique, feature reduction is performed by using class information.

- Feature extraction by the proposed PCSVD method: Seven new PCSVD components were extracted using PCA and SVD techniques. Here, not only was the dimensionality reduced but also the accuracy was improved. **Table 3** presents the number of features extracted by each method.

A later phase of the experiment explored comparative analysis of the classifiers to determine which existing and hybridized techniques were most effective in reducing the dimensionality of the data. The proposed technique, common feature extraction methods, and the entire feature set were used to evaluate the

effectiveness of the classifier SVC. **Table 4** shows the performance of the classifier SVC. The accuracy achieved with complete features is 65%; with ICA, PCA, and SVD, it is 97.25%; with LDA, it is 97.50%; and with the proposed PCSVD, it is 98.75%. The proposed PCSVD feature extraction technique improves the accuracy by 51.92% and reduces the dimensionality by 70.83% compared to the full feature set. Compared to existing feature extraction techniques, the accuracy increased by 1.54% and the dimensionality was reduced by 15%.

Table 3. Number of features extracted by each feature extraction technique.

FE techniques	Total number of features	Number of features extracted
PCA	25	8
ICA	25	8
SVD	25	8
LDA	25	(Preserve only class information)
Proposed PCSVD	25	7

Table 4. Performance of SVC classifier with different feature extraction techniques.

FE techniques	Accuracy (%)	RMSE (%)	Sensitivity (%)	Precision (%)	Specificity (%)	AUC (%)
All features	65.00	59.16	97.75	92.00	92.25	50.00
PCA	97.25	15.72	96.15	93.33	97.75	98.07
ICA	97.25	15.72	96.15	93.33	97.75	98.07
SVD	97.25	15.72	96.15	93.33	98.75	98.07
LDA	97.50	15.81	98.07	96.42	96.42	97.25
PCSVD	98.75	11.18	98.75	96.68	100	99.03

Figures 6–**10** illustrate the performance comparison of the classifiers of SVC with the proposed PCSVD feature extraction techniques and with other existing techniques in terms of accuracy, RMSE, AUC, specificity, and sensitivity.

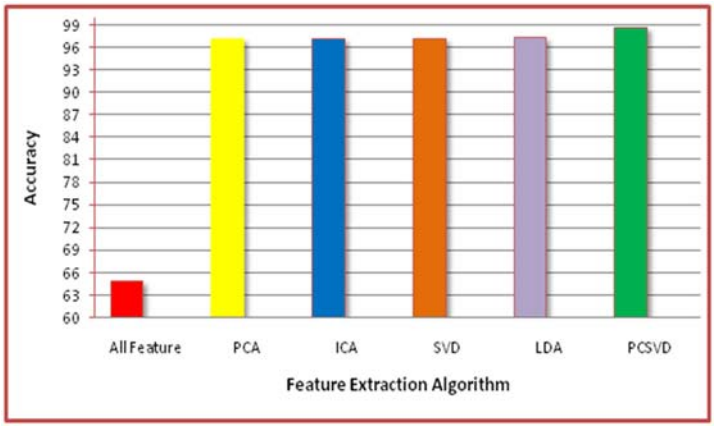

Figure 6. Comparison of accuracy of PCSVD with other feature extraction techniques.

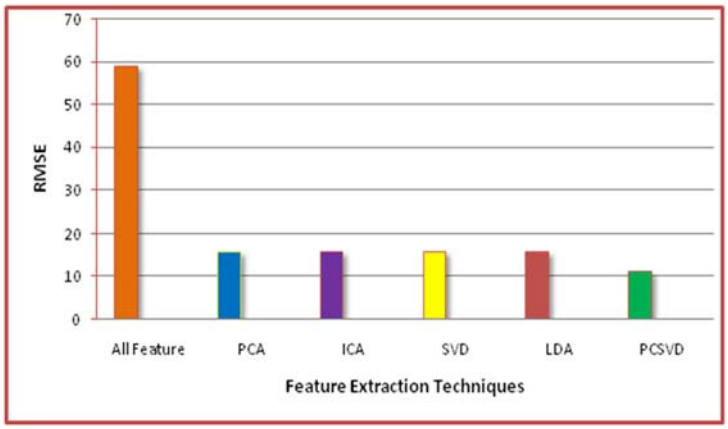

Figure 7. Comparison of RMSE of PCSVD with other feature extraction techniques.

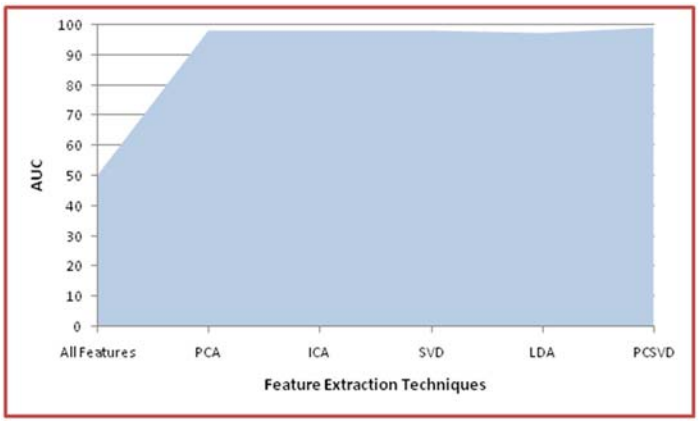

Figure 8. Comparison of AUC of PCSVD with other feature extraction techniques.

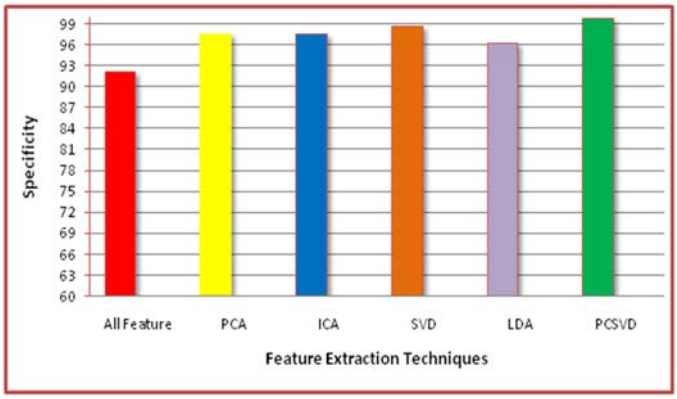

Figure 9. Comparison of specificity of PCSVD with other feature extraction techniques.

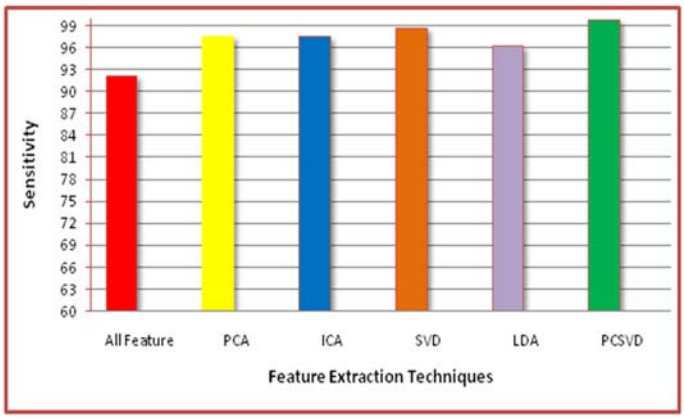

Figure 10. Comparison of sensitivity of PCSVD with other feature extraction techniques.

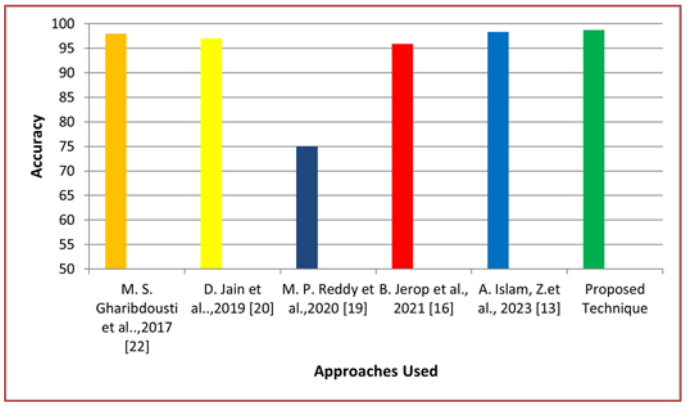

Figure 11. Improvement in accuracy of PCSVD technique.

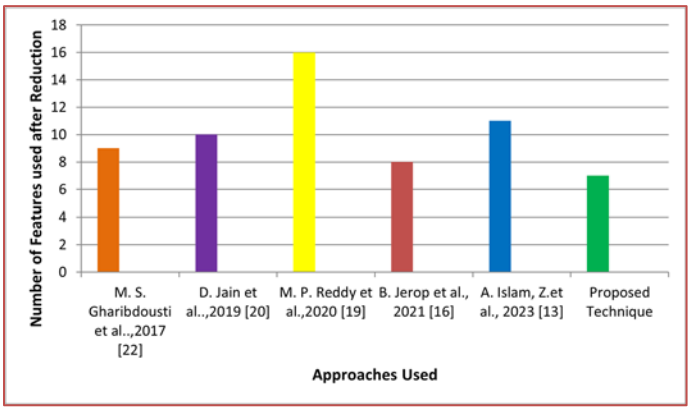

Figure 12. Improvement in dimensionality reduction of PCSVD technique.

5. Conclusion

Techniques for feature extraction and dimensionality reduction have a significant influence on the machine learning classification system. The key aspect of this research is that the authors proposed a hybrid feature extraction technique, PCSVD, and compared it with existing feature extraction techniques. It was implemented using the machine learning support vector classifier (SVC) on the standard chronic kidney disease dataset. The performance of the hybrid PCSVD technique was analyzed using a scale of measurement parameters, accuracy, sensitivity, specificity, AUC, and precision. PCSVD effectively reduces the overall dimensionality of 70.83% of the features and provides an improved reliable result of 98.75% accuracy, which is superior to existing approaches found in the literature.

For future work, the authors plan to implement and analyze the PCSVD technique with other machine learning algorithms such as Naive Bayes, logistic regression, decision tree, etc. The authors also plan to propose a hybrid formation of existing feature extraction techniques and compare them all to develop a predictive model for chronic kidney disease diagnosis.

Author contributions

Conceptualization, VG and NR; methodology, VG; software, VG; validation, VG and NR; formal analysis, VG; investigation, VG; resources, VG; data curation, VG; writing—original draft preparation, VG writing—review and editing, VG; visualization, VG; supervision, NR.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Gulati V, Raheja N. Comparative analysis of machine learning techniques based on chronic kidney disease dataset. *IOP Conference Series: Materials Science and Engineering* 2021; 1131(1): 012010. doi: 10.1088/1757-899X/1131/1/012010
2. Winter G. Machine learning in healthcare. *British Journal of Healthcare Management* 2019; 25(2): 100–101. doi: 10.12968/bjhc.2019.25.2.100
3. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015; 349(6245): 255–260. doi: 10.1126/science.aaa8415
4. Ayesha S, Hanif MK, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion* 2020; 59: 44–58. doi: 10.1016/j.inffus.2020.01.005
5. Dulhare UN, Ayesha M. Extraction of action rules for chronic kidney disease using Naïve bayes classifier. In: Proceedings of 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC); 15–17 December 2016; Chennai, India. pp. 1–5.

6. Storcheus D, Rostamizadeh A, Kumar S. A survey of modern questions and challenges in feature extraction. In: Proceedings of the 1st International Workshop "Feature Extraction: Modern Questions and Challenges"; 11 December 2015; Montreal, Canada. pp. 1–18.
7. Velliangiri S, Alagumuthukrishnan S, Joseph SIT. A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science* 2019; 165: 104–111. doi: 10.1016/j.procs.2020.01.079
8. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics* 2023; 14: 100189. doi: 10.1016/j.jpi.2023.100189
9. Venkatesan VK, Ramakrishna MT, Izonin I, et al. Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Applied Sciences* 2023; 13(5): 2885. doi: 10.3390/app13052885
10. Swain D, Mehta U, Bhatt A, et al. A robust chronic kidney disease classifier using machine learning. *Electronics* 2023; 12(1): 212. doi: 10.3390/electronics12010212
11. Ebiaredoh-Mienye SA, Swart TG, Esenogho E, Mienye ID. A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. *Bioengineering* 2022; 9(8): 350. doi: 10.3390/bioengineering9080350
12. Jerop B, Segera DR. An efficient PCA-GA-HKSVM-based disease diagnostic assistant. *BioMed Research International* 2021; 2021: 1–10. doi: 10.1155/2021/4784057
13. Navaneeth B, Suchetha M. A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. *Biomedical Signal Processing and Control* 2020; 62: 102068. doi: 10.1016/j.bspc.2020.102068
14. Inayatullah, Qayyurn H. An improved comparative model for chronic kidney disease (CKD) prediction. In: Proceedings of 2020 14th International Conference on Open Source Systems and Technologies (ICOSST); 16–17 December 2020; Lahore, Pakistan. pp. 1–8.
15. Reddy MP, Devi TU. Prediction of diagnosing chronic kidney disease using machine learning: Classification algorithms. *International Journal of Innovation Technology and Exploring Engineering* 2020; 9(4): 1922–1924. doi: 10.35940/ijitee.f3989.049620
16. Jain D, Singh V. A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification. *International Journal of Computers and Applications* 2021; 43(6): 524–536. doi: 10.1080/1206212X.2019.1577534
17. Gu S. *Applying Machine Learning Algorithms for the Analysis of Biological Sequences and Medical Records* [Master's thesis]. South Dakota State University; 2019.
18. Gharibdousti MS, Azimi K, Hathikal S, Won DH. Prediction of chronic kidney disease using data mining techniques. In: Proceedings of Industrial and Systems Engineering Conference; 20–23 May 2017; Pittsburgh, Pennsylvania. pp. 2135–2140.
19. Bouzalmat A, Kharroubi J, Zarghili A. Comparative study of PCA, ICA, LDA using SVM classifier. *Journal of Emerging Technologies in Web Intelligence* 2014; 6(1): 64–68. doi: 10.4304/jetwi.6.1.64-68
20. Reza MS, Ma J. ICA and PCA integrated feature extraction for classification. In: Proceedings of 2016 IEEE 13th International Conference on Signal Processing (ICSP); 6–10 November 2016; Chengdu, China. pp. 1083–1088.
21. Ramachandran R, Ravichandran G, Raveendran A. Evaluation of dimensionality reduction techniques for big data. In: Proceedings of 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC); 11–13 March 2020; Erode, India. pp. 226–231.
22. Li L, Wu Y, Ou Y, et al. Research on machine learning algorithms and feature extraction for time series. In: Proceedings of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); 8–13 October 2017; Montreal, Canada. pp. 1–5.
23. Tanwar S, Ramani T, Tyagi S. Dimensionality reduction using PCA and SVD in big data: A comparative case study. In: Proceedings of Future Internet Technologies and Trends: First International Conference, ICFITT 2017; 31 August–2 September 2017; Surat, India. pp. 116–125.
24. Gulati V, Raheja N, Gujral RK. Pica-A hybrid feature extraction technique based on principal component analysis and independent component analysis. In: Proceedings of 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT); 7–9 October 2022; Bangalore, India. pp. 1–6.
25. Almeida AR, Almeida OM, Junior BFS, et al. ICA feature extraction for the location and classification of faults in high-voltage transmission lines. *Electric Power Systems Research* 2017; 148: 254–263. doi: 10.1016/j.epsr.2017.03.030
26. Sarhan M, Layeghy S, Moustafa N, et al. Feature extraction for machine learning-based intrusion detection in IoT networks. *Digital Communications and Networks* 2022; in press.
27. Kadhim AI, Cheah YN, Hieder IA, Ali RA. Improving TF-IDF with singular value decomposition (SVD) for feature extraction on Twitter. In: Proceedings of 3rd International Engineering Conference on Developments in Civil and Computer Engineering Applications; 26–27 February 2017; Erbil, Iraq.
28. Sujatha R, Ephzibah EP, Dharinya S, et al. Comparative study on dimensionality reduction for disease diagnosis using fuzzy classifier. *International Journal of Engineering and Technology* 2018; 7(1): 79–84. doi: 10.14419/ijet.v7i1.8652
29. Janani J, Sathyaraj R. Diagnosing chronic kidney disease using hybrid machine learning techniques. *Turkish*

Journal of Computer and Mathematics Education (TURCOMAT) 2021; 12(13): 6383–6390.
30. Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease—A machine learning perspective. *IEEE Access* 2021; 9: 17312–17334. doi: 10.1109/ACCESS.2021.3053763
31. Zelaya CVG. Towards explaining the effects of data preprocessing on machine learning. In: Proceedings of 2019 IEEE 35th International Conference on Data Engineering (ICDE); 8–11 April 2019; Macao, China. pp. 2086–2090.
32. Alam S, Yao N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory* 2019; 25(3): 319–335. doi: 10.1007/s10588-018-9266-8
33. Huang J, Li Y, Xie M. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology* 2015; 67: 108–127. doi: 10.1016/j.infsof.2015.07.004