NIJEC 4 (2) (2025) 1-12

NUML International Journal of Engineering and Computing

Journal Homepage: NIJEC.NUML.edu.pk
Print ISSN: 2788-9629

Effect of Dimensionality Reduction on Machine Learning Models Performance: A Review

Inam Ul Haq[a]*, Toffiq Saddique[b], Saqlain Abbas[c], Wasi Haider Butt[a], Umar Ajaib Khan[d]

[a] College of Electrical and Mechanical Engineering, NUST, Pakistan
[b] Staffordshire University, UK
[c] University of Chinese Academy of Sciences
[d] National University of Modern Languages, Mirpur AJK, Pakistan

ARTICLE INFO

Article History:
Received 29 October 2025
Accepted 18 December 2025
Available Online 6 January 2026

Keywords:
Dimensionality Reduction;
Feature Selection;
Machine Learning (ML); Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA

Funding:
This research received no specific grant from any agency in public, commercial or non-for-profit sector

Conflict of Interest:
The author has declared no potential conflicts of interest and falsification/fabrication of data with respect to the research, authorship, and/or publication of this article.

ABSTRACT

The performance of machine learning models are fully dependent on the training data. In today's era, where many companies transforming their businesses towards digitalization. A huge volume of raw data is generated and collected, extracting important and useful features from raw data is quite a big challenge, because it is computationally very expensive and time-consuming to train machine learning models on raw data and result in poor model's performance due to irrelevant features. So, it is necessary to filter out data by reducing the size of features or dimensions and extracting important features from a huge amount of data before implementing any machine learning model on it. In this situation, Dimensionality reduction techniques come into play to eliminate irrelevant features from the data set. In this paper, we are going to review some of the important dimensionality reduction techniques on different machine learning models and evaluate the performance of how dimensionality reduction techniques improve the overall performance of models. After reviewing the research work of different authors, we conclude our results by comparing different implemented techniques of dimensionality reduction, the results before reducing the size of features/dimensions, and after implementing dimensionality reduction techniques and figure out key findings and limitations of reviewed papers.

1. Introduction

In the past few years, Machine Learning (ML) is a fast- growing technology around the globe. Data scientists use ML for extracting useful knowledge from a large amount of data for decision making. ML algorithms allow machines to learn from the data and predict useful knowledge based on training data. Once we train our ML models on training data it predicts the results when unseen or test data is given to the ML models. But the performance or accuracy of the ML model is dependent on what type of data is used for training. Realistic raw data collected from any source consists of noise and irrelevant information that leads to a poor ML model's performance. So, Data filtration and preprocessing are necessary to avoid biased results. Selecting important features and reducing the size of the data set is quite a big challenge. Dimensionality reduction techniques help us to reduce the dimensions or features of data. By reducing the size of dimensions not only improve the performance of ML models but also useful in the data visualization process [1]. Different dimensionality reduction techniques are used for a different purpose. There are many well-known Dimensionality reduction techniques in which Principal component analysis (PCA) simply projects the n- dimensional data into lower dimensions and useful for feature extraction and data compression [2]. Linear Discriminant

https://doi.org/10.52015/nijec.v4i2.104

Analysis (LDA) aims to reduce data from n- dimensional space to lower-dimensional space by keeping achievable discriminative information and applicable to linear datasets [3]. Recursive feature elimination (RFE) extract features from a dataset by assigning different weights using an estimator. The entire feature set is used for estimator in training and features with the highest weights are selected [4].

Independent component analysis (ICA) is based on linearly separable data and is used for dimensionality reduction [5]. The wrapper approach is also used for feature selection that consists of ensembles as bagging and random subspace [6]. The hybrid dimensionality reduction (HDR) approach provides higher classification performance for noisy data [7]. A hybrid model whereas k-means and genetic algorithm were used to reduce dimensionality [8]. A fuzzy rule-based system is also used for reducing the size of dimensions and select important and useful features from a large set of data and improve the overall accuracy of ML models [9].

Kernel Principal component analysis (KPCA), and robust sparse principal component analysis (RSPCA) perform well on classification problems and also reduce the dimensions of the dataset and improve the overall accuracy of the ML model [10, 11]. A lot of work is done by researchers in the dimensionality reduction domain in this paper we are going to summarize their research by comparing the performance and accuracy of different dimensionality reduction techniques used in ML models and figure out their key findings and limitation by Highlighting the pros and cons of different dimensionality reduction. This enables the new practitioners and researchers to go through the different dimensionality reduction techniques and understanding the tradeoff between these techniques.

From the introduction part, dimensionality reduction plays a significant role in the field of machine learning to achieve low dimensional space from high feature space. In this paper, we are going to address the comprehensive literature review, and based on previous literature, we are going to answer the following questions.

Table 1: Research Questions

#	Research Questions
1	What are the advantages and limitations of the techniques and tools discussed?
2	Which technique gives us the best results?
3	Which dimensionality reduction techniques are suitable for large dataset?

2. Literature Review

Santhanam et al. [12] introduced a hybrid model for diagnosing diabetes by integrating the SVM model, whereas k-means and genetic algorithms were used to reduce dimensionality. K-means clustering technique is used for noise removal and genetic algorithm is used for feature selection which with the help of SVM classification model produced highly accurate results as compared to the previous models.

Bharti et al. [13] introduced a hybrid model for text clustering using feature selection and feature extraction. To select both the important and common features from a dataset, integrated union and intersection approaches are used. The union approach selects the important features while the intersection is applied to common features. Therefore, it ensures that lesser and all types of features are selected, and dimensionality is not increased. The results obtained from the proposed model are compared with the previously used techniques which involve Term Variance and Document Frequency. The comparative study shows that the proposed model produced more accurate and highly efficient results as compared to the other techniques.

Lin et al. [14] proposed a model based on dimension reduction and machine learning techniques to detect radio frequency RF fingerprint identification. The purpose of this research is to ensure the security of wireless devices and to work out authentication security issues. The techniques used are

KPCA, PCA, and RSPCA for dimensionality reduction whereas for machine learning SVM, random forest, grey correlation analysis, and artificial neural networks are used. The results obtained from the comparison of these studies show that the proposed model produced highly accurate and precise results as compared to the traditional techniques used for machine learning and dimensionality reduction.

Nassirtoussi et al. [15] studied that, with the help of text mining newspapers FOREX prediction was done. For this reason, a multi-component algorithm was created which addresses the features of the text mining issue listed in a specific layer individually. The primary layer called the semantic abstraction layer tackles the issue of text mining Co reference which leads to sparsity. This work creates a customized approach called heuristic hypernyms feature selection which enables words having the identical base word to be viewed as single object. The accuracy of prediction at this layer improves dramatically and is due to an acceptable drop in noise from the function room. Sentiment integration layer is the second-level layer, incorporating the capacity of analyzing sentiments using the algorithm by suggesting the sentiment weight called Sum Score, representing these as the sentiment of investors. This layer decreases the dimensions by extracting sentiments of 0 value thereby improving prediction accuracy.

Zhang et al. [16] present an unsupervised deep-learning system for the reduction of dimensions called Local Deep Function Alignment LDFA. For each sample, the author created a neighborhood and learn from the neighborhood a local SCAE for extracting the deep features. then a reinforcement to integrate each neighborhood's local characteristics but its global characteristic was used. Also, an NDFA approach to directly transform a new sample of data into the trained low-dimensional subunit was derived. The benefit of this LDFA approach is that the data sample set acquires both local and global properties local SCAEs capture local data set characteristics while global harmonization procedures encode interdependence between neighborhoods with the ultimate low- dimensional representations.

Pajouh et al. [17] studied that the ability to detect breaches and suspicious attacks inside the IoT systems is important for the flexibility of the system's framework, with growing emphasis on Internet of Things appliances and provisions. In this article, the author presented a new breach detection case study constructed based on bi-layer dimension reduction and the classification component consisting of two stages, aimed at detecting suspicious attacks for instance Remote to Local R2L and User to Root U2R attacks. This method is used to distribute high dimension data set to lower data set with fewer features by means of component analysis and linear discriminatory analysis dimensions reduction modules. They also used a 2-step classification module to classify suspicious behaviors such as Naïve Bayes and Certainty Factor versions of K-Nearest Neighbor. This model performs better as compared to the prior models developed in order to prevent U2R and R2L attacks.

Kasun et al. [18] discussed that data can often contain noise or unnecessary data which adversely affects machine learning algorithms and their prediction performance for negative matrix factorization and non- linear AE are mutilated via a poor learning rate in addition irregular prediction denotes only a tensor of initial data. This research presents a dimension reduction structure, that often signifies sections of data, with a very fast learning speed, and learns the scatter subspace amid groups. This research explores an intense learning machine AE and sparse ELM-AE for linear and non- linear dimensionality reduction system.

Du et al. [19] proposed a classification framework based on spectral and spatial features. It practices dimension reduction and deep learning techniques for spectral and spatial feature extraction, respectively. The model caters to the reduction of dimensions and classification in both layers. It is developed to measure provocative IoT activities, particularly detecting low-frequency attacks for example U2R and R2L which are potentially harmful.in terms of the rate of detection of attacks in both low frequency and typical attack frequency, this model performed much better as compared to the existing similar models. As this method uses both supervised and unsupervised

extraction approaches such as PCA and LDA respectively, therefore, due to the use of classification algorithms, the classification between various attack types and typical behaviors can be achieved.

Yang Zhang et al. and Zhidong Zhao et al. [20] the authors proposed a technique which is selecting the optimal functions. These functions increasing accuracy with the use of pathological classifications. PCS is used for feature selection. The pre-processing stage in machine training has proved that it is very effective as well as improve computational time and accuracy. This technique also helpful for the medical workers as well as staff to make decisions in the right ways and quickly by interpreting readings of CTG.

Kalia Orphanou et al. and Arianna Dagliati et al. [21] proposed a private key. The homomorphic algorithm is an encryption algorithm to ensuring medical data. It is just for the restricted Naive Bayes classification. This method also permits the data operator to access or categorize data as well as private information. Authors also test this method on the breast cancer of data and produce results that are both accurate and correct. Minas A Karaolis et al. [22] authors proposing a decision tree process which is based on the data mining techniques to predict the risk factors for CHD. This analysis was operating a Decision Tree algorithm, with different parameters which are based on risk factors. By using this approach accuracy and result of implementation shows the CHD diagnosis can be reduced.

Irem Ersoez Kaya et al. and Ayça Çakmak Pehlivanlı et al. [23] Investigation of brain tumors images by using PCA clustering-based technique. In this technique, the PCA applying the first time for MRI images. It is different types of shapes and sizes. It also applied clustering with the help of K-means and FCM. It is performing high accuracy or performance rate by the integrating of PCA as well as K-means. Sweta Bhattacharya et al. [24] proposing a PCA-Firefly algorithm that is based on the detection of datasets to classify. In this technique, dimensionality reduction is performing by the PCA-Firefly algorithm as well as the transformation of one-hot encoding. Dimensionality reduction applied with the help of XG Boost classifier. This method proposed accurate and experimental results.

Zheng Li et al. [24] authors perform for improving the quality of machine learning models as well as feature engineering to specify potentially interesting catalytic materials. According to Groce et al. [25], he said that we can illustrate genuine and complex problems with the help of classification algorithms. But sometimes it is hard to detect classifiers' faults. Another researcher Letham et al. [26] said that for the patient of atrial fibrillation Bayesian association and decision rule can be used for the estimation of stroke risk.

Zhang et al. [27] introduced a simple warning system that was used for analyzing the instances of input and then notify if the system produces an unwanted result. Wozniak et al. [28] perform deep study about classifiers work especially heterogeneous classifiers which combine different types of classifiers. Homogenous MCS classifiers like RF are key classifiers and are composed of similar types of classifiers like SVM, Maximum Likelihood, k-Nearest Neighbor, and Multilayer Perception classifiers in modern systems such as fraud detection systems, health care systems, computer security systems, and recommender systems, etc. In a classification of multilabel framework author Jun et al. [29], gives an idea about classifiers that depends upon the chain method. He said that the performance of classification is affected by the label order.

Lee et al. [30], suggested a method called memetic attribute selection that depends upon unique attributes filtration which was very helpful in the categorization of the multilabel text.

Zhang et al. [31] introduced a new procedure for the selection of multilabel attributes. According to his method, labels or attributes are categorized into two different groups. One group was considered as dependent labels and the other was considered as independent labels. The difference between these two types of labels can be analyzed by initiating a new important attribute term which is conditional common information among candidate attributes and every label will give other labels. Nam et al. [32] introduced a neural network on the behalf of classifier chain method. According to the author, a neural network algorithm works as a sequence wise prediction algorithm, and this can

be used in sequential projection tasks in different domains. Read et al. [33] discovers and explains the interaction among the model of markovain and the method of multilabel. He suggested that method of multilabel can be used for the data which is in sequential form.

2.1 Data Source and Search Strings

Table 2: Journals and Conferences

1	IEEE Explore	https://ieeexplore.ieee.org/
2	Research Gate	https://www.researchgate.net/
3	Science Direct	https://www.sciencedirect.com/
4	ACM Library	https://dl.acm.org/
5	Springer	https://link.springer.com/
6	Semantic scholar	https://www.semanticscholar.org/

("Effects" OR "Importance" AND "Dimensionality OR Features") AND ("Reduction" OR "Elimination" AND "Technique" OR "Methodology") in Machine Learning.

2.2 Data Extraction

Selected papers are listed at the end of the paper. Following key strings were used "importance of Dimensionality Reduction in ML", "Effect of Dimensionality Reduction in ML", "Feature Extraction Techniques", "Feature elimination Techniques".

Table 3: Literature Review Summary

Cited Paper	Methodologies Used	Main Findings	Limitations
[12]	SVM, K-Means and Genetic Algorithms	A hybrid model is introduced to diagnose diabetes by integrating SVM. K-means and genetic algorithms were also used to reduce dimensionality. The results showed an accuracy of 98%.	Missing values and outlier detection are not discussed. These methods are useful in enhancing accuracy.
[13]	Feature Section and Feature Extraction.	A hybrid model is proposed for text clustering using feature selection and feature extraction. Combined Union and intersection approaches are also used for dimensionality reduction. The results obtained showed improved clustering accuracy.	The proposed method is dependent on the values of parameters discussed, hence resulting in varying performance results. Moreover, the accuracy of dimension reduction can be enhanced and made more efficient.
[14]	PCA, RPCA, KPCA. SVM, Random Forest, Grey Correlation Analysis, and ANN	This study proposed a model to reduce dimensionality using PCA, RPCA, and KPCA along with random forest, SVM, grey correlation analysis, and ANN for machine learning. The results obtained show accuracy of around 90% through classification along with 70% dimensionality reduction.	The paper only concentrates on the identification of device authentication and may not work in the case of non- certified devices.
[15]	Text Mining Techniques	This study proposed a method to predict FOREX through text mining at different layers. The obtained results show an accuracy of 83.3%	The study focuses on FOREX, whereas other markets can also be researched. Moreover, testing of proposed aspects can be useful in increasing accuracy.

Cited Paper	Methodologies Used	Main Findings	Limitations
[17]	Component and Linear Discriminate Analysis. Naïve Bayes and KNN.	The study proposed a model to detect and identify suspicious attacks and breaches in IoT devices and services. The model consists of reducing dimensionality and classification. Using component and linear discriminate analysis the dimensionality is reduced and KNN and Naïve Bayes are used for the classification method. The results produced are accurate and much better than the previous research.	Classification methods can be used for other types of attacks such as U2R, R2L, etc.
[19]	Convolutional Neural Network	The study proposed a spatial and spectral feature- based model. The classification technique is used in the two-layer model to prevent attacks. The accuracy of this model is much higher as compared to the other typical models.	The parameter used are dependent and produces varying results in case of different size values.
[20]	PCA and Ada Boost	Ada Boost's model occurred for the performance of other models with the computational model's time. (2.4 to 11.6)	In the given dataset, the proposing system was not satisfied for the mixed noisy values
[21]	Naive Bayes Classifiers	In this temporal association rules using for the disease. And this disease is coronary heart disease. These rules diagnose the disease and produce high results.	These rules used for the multiple window system, because multiple window system is sluggish rather than single one.
[22]	Decision Tree	This Classification produces various risks for the disability of coronary heart disease.	Obtained algorithms are not sufficient used for reducing morbidity (CHD).
[23]	PCA and Fuzzy Means	Some algorithms like (EM-PPCA, PPCA) works effectively rather than other clustering algorithms like (K-Means, FCM).	In this accuracy as well as values are not high because the number of iterations was very less.
[24]	PCA, Firefly	These rules are used for the dimensionality reduction as well as XG Boost for classification.	It consumes time and Increase complexity in dimensionality reduction and training phases.
[25]	Classification Algorithm	We can handle complex problems with this approach. In this approach, the dataset is divided into classes of labels for a simpler form.	Sometimes it will not find classifiers' faults.
[29]	Chain Method	Performance of classification can be affected by changing the order of labels.	This rule works with too many parameters, so a nonexpert person does not know which feature is best.
[30]	Memetic Attribute Selection	This approach depends upon unique attributes filtration and uses for classification.	If the dataset does not contain unique attributes, then it will not work accordingly.
[32]	Neural Network	It is a sequence wise predicting algorithm and used a projection task.	This approach requires more data than other machine learning algorithms and it is computationally expensive.

Cited Paper	Methodologies Used	Main Findings	Limitations
[33]	Model of Markovian	It is used for systems that make changes randomly. It considers the current state as a feature state.	It will not work with data that is in random form or that is not in sequential form.

3. Discussion

The objective of this paper was to perform a comprehensive review of dimensionality reduction techniques. The articles selected for this research included different research techniques and methodology and every technique has a certain impact on the results [34]. We examined 22 latest papers, and the individual proposed method is described in detail with their corresponding results and datasets used in the respective research papers. With the help of a detailed literature review answers to the questions highlighted in the introduction part as follows.

RQ1: What are the advantages and limitations of the techniques and tools discussed?

The research study used Support Vector Machine (SVM) is used to classify the diabetes diagnoses along with K- means and Genetic algorithms to reduce dimensionality [12]. The number of attributes was reduced and only 3 to 6 attributes were selected using a genetic algorithm, resulting in improving the accuracy of the classification algorithm to about 98.79% and producing better results as compared to the previous studies. But on the other hand, the proposed study did not cater to the outliers. The other critical attributes for diabetes diagnosis such as pregnancies, PG Concentration, and Age should also be kept in consideration and were ignored in outlier data.

In another research study [13], a novel hybrid approach is introduced by using union and intersection approaches. PCA approach is used which further refines the selected features. This methodology produced much-improved results but due to its dependency on the parameters, it produced varying results in the presence of different values of these parameters. This limitation affects the results of this methodology in different situations. [14] introduced the RF Radio Frequency identification method. PCA, RSPCA, and KPCA were used to reduce dimensionality and random forest, Support Vector Machine, Artificial Neural Network (ANN), and Grey Correlation Analysis were used as machine learning algorithms. The proposed model produced the results with an accuracy of 90%. On the other hand, the proposed model only focuses on the identification during device authentication and may not work in the case of non- certified devices. Furthermore, the feature selection for dimensionality reduction also needs improvement as it does not produce much accuracy.

Another study [15], proposed a novel approach to predict FOREX through mining a textual data at a particular layer. The primary layer named Semantic Abstraction Layer deals with the issues of co-reference which usually occurs whenever two or more words insinuate the same meaning. An approach named Heuristic Hypernyms Feature Selection tackles this issue. The second layer named the sentiment integration layer proposes a sentiment score with the name Sum Score. This Sum Score defines the weight of investor's sentiment. Another purpose of this layer is to remove the values with zero scores, hence reducing dimensionality. The proposed model shows the results with an accuracy of about 83.33%. On the other hand, the research only focuses on the FOREX market whereas there is room for many other domains as well. The testing and experimentation in other domains can also be done using this model. Furthermore, abstraction can be improved by using other deep learning algorithms.

Zhao, W et al. [19] proposed a model named Spatial Spectral Feature- based Classification (SSFC). This model extracts features from high dimension hyperspectral data using a local discriminant algorithm whereas, Convolutional Neural Network (CNN) is used to retrieve high-level spatial features. The dimension reduction algorithm used produced better results as compared to the prior research. The limitation found in this paper is the scope of the learning samples. The scope of learning

data is the major aspect when it comes to deep learning models therefore, its size plays a vital role in the results of the model proposed. [32] discussed the model which uses recurrent neural networks to maximize subset accuracy. The model focuses only on the positive values, hence reducing the dimensionality. The results produced using this model show better results, but also, on the other hand, it is dependent on the input features and unique labels, thus producing varying results in different situations [35].

RQ2: Which technique gives us the best results?

Nowadays, complex techniques are being used to reduce dimensions like LDA and Principal Component Analysis, but the drawback of these two approaches is that the values or data which comes after applying PCA on the dataset or reducing the dimensions and the actual dataset dimensions cannot be mapped with each other. The same is the issue with the LDA technique.

Now question is that as PCA is a complex technique if we use the statistical standard deviation or variance approach, will it work better than PCA and LDA or not. What results would be gained by these approaches?

So, for this purpose, we will do a comparison of simple dimension reduction approaches and complex dimension reduction approaches as well. We planned to select a dataset and then we reduced the features of the dataset and check the accuracy of the dataset. For this purpose, we have selected a flower dataset named Iris. We did feature reduction by determining which features or attribute were more important and which feature were less important, which attributes were used for clustering of the selected dataset. We also checked the output and compared it with other approaches. We also calculated the error ratio or error percentage. Lastly, we only consider dimensionality reduction approaches that work with linear datasets [29], [30]. For the experiment, we used the flower dataset as mentioned above [31]. There were five dimensions in the dataset that represent a feature of each flower like the length as well width. We considered three kids of Iris flower. They are named iris_sentosa, iris_virginica, and iris_versicolor. The total number of rows in the dataset was 150. There were 50 rows for each flower type like 50 rows for iris_sentosa, 50 rows for iris_virginica, and 50 rows for iris_versicolor. The attributes that were considered for classification were the measurement of petals of flower and sepals of flower-like length and width of these attributes of the flower. We experimented by using the dimensionality reduction approaches that were considered as linear type approaches as the data was of linear type. For making groups of data, clustering approach k- means algorithm was used by putting the value of seed=4. We have a set number of clusters =3 and a number of iterations =10 for experimenting. To experiment, we use google collab which is a freely available notebook for python. Results of different techniques are given below in Table 4.

RQ3: Which dimensionality reduction techniques are suitable for large dataset?

The most popular methods for dimensionality reductions are PCA and LDA. In principle component analysis (PCA) defining new attributes which is mutually orthogonal of linear combinations with the real attributes. While linear discriminant analysis (LDA) handling the data of class frequencies are unequal as well as performance examined on irregular test data. This method provides the ratio within-class variance as well as between-class variance.

A research study [20] “Fetal state assessment-based on cardiotocography parameters” proposing a (PCA) method for the selection of features to enhance the accuracy. Feature selection performing with the help of PCA. PCA works as pre-processing before training, it provides effective results as well as examined accuracy. It is helpful for medical staff to take decision run time as soon as possible.

Table 4: Result of Different Dimensionality Reduction Technique

Dimensionality Reduction Approaches	Total Rows in Dataset	No of Dimensions	No of Item Set in Each group	No of Reduced Dimensions	Names of R Dimensions left after Reduction	Correctly Identified Item Set After Dimensionality Reduction in %	Error Percentage	Time Consumed for Grouping After Dimensionality Reduction in ms
Standard Deviation	150	5	50	1	Length of Petal	98.66	1.33	150
Variance	150	5	50	1	Length of Petal	98.66	1.33	150
Principal Component Analysis (PCA)	150	5	50	2	Length of Sepal and Width of Sepal	96	4	168
Linear Discriminant Analysis (LDA)	150	5	50	2	Function of LDA Length of Sepal & Width of Sepal	98	2	238
Factor Analysis	150	5	50	1	Length of Petal	97.3	2.66	6656
Original Dataset (Without Dimensionality Reduction)	150	5	50	-	-	92	8	2

A research study [23] "PCA based clustering for brain tumor segmentation" proposing a method pf PCA apply first time for MRI images. It is different types of shapes and sizes. It also applied clustering with the help of K- means and FCM. Its results in high accuracy by the integrating of PCA.

Bhattacharya, S., et al [24] proposing a PCA-Firefly algorithm which is based on the detection of datasets to classify. In this technique, dimensionality reduction is performing by the PCA- Firefly algorithm as well as the transformation of one-hot encoding. Dimensionality reduction was applied with the help of the XG Boost classifier. This method proposed accurate and experimental results.

PCA and LDA find the best features from the large datasets to reduce the dimension. PCA is a technique for reduction of the principal component of real data and referred eigen images, while LDA computes eigenvectors from the datasets and show it on the scatter matrices.

4. Result Analysis

According to the results of the experiment, the best approach which has given us the best results was dimensionality reduction Variance and Standard deviation approaches. From four initial dimensions in the set of features of the Iris flower set, the most important dimension was the length of the petal. By using standard deviation and variance approaches we got good results with almost 1.33% error. This means that with these approaches we have gained more than 98% of accuracy if we considered and worked with an only important set of features. Both approaches standard deviation approach and variance approach that was used within current findings, used for the identification of important dimension indicated clustering took less time for its completion. The last tuple in the above table shows that the overall accuracy that exists in the clustering was around 92%. This shows that if important dimensions or features will be included in the clustering data then both processing time and accuracy will also be improved.

5. Conclusion

In this paper, a detailed literature review is performed and based on a comprehensive literature review of different dimensionality reduction techniques highlighted in previous work, we try to answer some identified questions. In which we discussed some limitations and benefits of different methodologies used in machine learning for feature or dimension reduction and compared these techniques to answer that which technique performs better compared to others. Moreover, we analyze the effect of dimensionality reduction on machine learning model performance and how different techniques affect the accuracy of ML models. We also find that if the dataset is large and complex, PCA and LDA perform well as compared to other techniques. We also observe for linear datasets simple statistical features selection techniques give us better performance as compared to the complex dimensionality reduction techniques like PCA and LDA. Ultimately, the most effective methodology involves a process of exploration, rigorous testing, and cross-validation to determine which approach yields the optimal results for a given problem, and this depends upon several factors like size of the dataset, data characteristics, computational resources, the presence of noise and outliers and relationship between features etc [36]. Every approach/technique has its own pros and cons [37]. Lastly, a combination of different approaches complements each other and results in providing better quality outcomes.

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.

Funding

The research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Fabrication/Falsification Statement

The author(s) declare that no data have been fabricated, falsified, or manipulated in this study.

Participant Consent

The authors confirm that Informed consent was obtained from all participants, and confidentiality was duly maintained.

References

[1] Griparis, A., D. Faur, and M. Datcu. Feature space dimensionality reduction for the optimization of visualization methods. in 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2015. IEEE.

[2] Reddy, T.A., K.R. Devi, and S.V. Gangashetty. Nonlinear principal component analysis for seismic data compression. in 2012 1st International Conference on Recent Advances in Information Technology (RAIT). 2012. IEEE.

[3] Ghosh, J. and S.B. Shuvo. Improving Classification Model's Performance Using Linear Discriminant Analysis on Linear Data. in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019. IEEE.

[4] Guyon, I., et al., Gene selection for cancer classification using support vector machines. Machine learning, 2002. 46(1-3): p. 389-422.

[5] Wang, J. and C.-I. Chang, Independent component analysis- based dimensionality reduction with applications in hyperspectral image analysis. IEEE transactions on geoscience and remote sensing, 2006. 44(6): p. 1586-1600.

[6] Filali, A., C. Jlassi, and N. Arous. Dimensionality reduction with unsupervised ensemble learning using K-means variants. in 2017 14th International Conference on Computer Graphics, Imaging and Visualization. 2017. IEEE.

[7] Moon, S. and H. Qi, Hybrid dimensionality reduction method based on support vector machine and independent component analysis. IEEE transactions on neural networks and learning systems, 2012. 23(5): p. 749-761.

[8] Shi, H. and M. Xu. A data classification method using genetic algorithm and K-means algorithm with optimizing initial cluster center. in 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET). 2018. IEEE.

[9] Chen, Y.-C., N.R. Pal, and I.-F. Chung, An integrated mechanism for feature selection and fuzzy rule extraction for classification. IEEE Transactions on Fuzzy Systems, 2011. 20(4): p. 683-698.

[10] Xu, Z., et al. Cross-version defect prediction via hybrid active learning with kernel principal component analysis. in 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER). 2018. IEEE.

[11] Nabhan, M., Y. Mei, and J. Shi, High Dimensional Process Monitoring Using Robust Sparse Probabilistic Principal Component Analysis. arXiv preprint arXiv:1904.09514, 2019.

[12] Santhanam, T. and M. Padmavathi, Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 2015. 47: p. 76-83.

[13] Bharti, K.K. and P.K. Singh, Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Systems with Applications, 2015. 42(6): p. 3105-3114.

[14] Lin, Y., et al., The individual identification method of wireless device based on dimensionality reduction and machine learning. The Journal of Supercomputing, 2019. 75(6): p. 3010-3027.

[15] Nassirtoussi, A.K., et al., Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 2015. 42(1): p. 306-324.

[16] Zhang, J., J. Yu, and D. Tao, Local deep-feature alignment for unsupervised dimension reduction. IEEE transactions on image processing, 2018. 27(5): p. 2420-2432.

[17] Pajouh, H.H., et al., A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. IEEE Transactions on Emerging Topics in Computing, 2016.

[18] Kasun, L.L.C., et al., Dimension reduction with extreme learning machine. IEEE transactions on Image Processing, 2016. 25(8): p. 3906-3918.

[19] Zhao, W. and S. Du, Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. IEEE Transactions on Geoscience and Remote Sensing, 2016. 54(8): p. 4544-4554.

[20] Zhang, Y. and Z. Zhao. Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost. in 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP- BMEI). 2017. IEEE.

[21] Orphanou, K., et al., Incorporating repeating temporal association rules in naïve bayes classifiers for coronary heart disease diagnosis. Journal of biomedical informatics, 2018. 81: p. 74-82.

[22] Karaolis, M.A., et al., Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on information technology in biomedicine, 2010. 14(3): p. 559-566.

[23] Haq, I.U., Saddique, T., Basharat, M., Butt, W.H. (2025). A Hybrid S/W Requirement Elicitation Approach to Improve Quality of Requirements. In: Nagar, A., Jat, D.S., Mishra, D., Joshi, A. (eds) Intelligent Sustainable Systems. Worlds4 2024. Lecture Notes in Networks and Systems, vol 1180. Springer, Singapore. https://doi.org/10.1007/978-981-97-9324-2_7.

[24] Bhattacharya, S., et al., A Novel PCA-Firefly based XGBoost classification model for Intrusion Detection in Networks using GPU. Electronics, 2020. 9(2): p. 219.

[25] Groce, A., et al., You are the only possible oracle: Effective test selection for end users of interactive machine learning systems. IEEE Transactions on Software Engineering, 2013. 40(3): p. 307-323.

[26] Letham, B., et al., Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 2015. 9(3): p. 1350-1371.

[27] Zhang, P., et al. Predicting failures of vision systems. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[28] Woźniak, M., M. Graña, and E. Corchado, A survey of multiple classifier systems as hybrid systems. Information Fusion, 2014. 16: p. 3-17.

[29] Jun, X., et al., Conditional entropy based classifier chains for multi-label classification. Neurocomputing, 2019. 335: p. 185-194.

[30] Lee, J., et al., Memetic feature selection for multilabel text categorization using label frequency difference. Information Sciences, 2019. 485: p. 263-280.

[31] Zhang, P., G. Liu, and W. Gao, Distinguishing two types of labels for multi-label feature selection. Pattern Recognition, 2019. 95: p. 72-82.

[32] Nam, J., et al., Maximizing subset accuracy with recurrent neural networks in multi-label classification. Advances in neural information processing systems, 2017. 30: p. 5413- 5423.

[33] Read, J., L. Martino, and J. Hollmén, Multi-label methods for prediction with sequential data. Pattern Recognition, 2017. 63: p. 45-55.

[34] Inam-Ul-Haq, W. A., Shakoor, M., & Butt, W. H. (2024). Systematic Literature Review on Methodologies for Improving Software Quality in Software Development Process.

[35] Inam-Ul-Haq, W. Abbas and W. H. Butt, "Systematic Literature Review on Requirement Management Tools," 2022 International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan, 2022, pp. 1-6, doi: 10.1109/ICETST55735.2022.9922932.

[36] Ul Haq and W. Haider Butt, "The State of Practices in Requirement Elicitation: An Improved Methodology for Pak Software Industry," 2022 17th International Conference on Emerging Technologies (ICET), Swabi, Pakistan, 2022, pp. 76-82, doi: 10.1109/ICET56601.2022.

[37] Haq, I. (2025). Sustainable Development Of Vehicular Fog Computing Using Karlskrona Manifesto. Numl International Journal Of Engineering And Computing Учредители: National University of Modern Languages, 3(2).