International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume-11, Issue-5, October 2024
https://doi.org/10.55524/ijirem.2024.11.5.17
Article ID IJIR3034, Pages 121-126
www.ijirem.org

Applying Principal Component Analysis and Autoencoders for Dimensionality Reduction in Data Stream

Mayur Prakashrao Gore[1], Amol Ashokrao Shinde[2], and Amit Choudhury[3]

[1] Principal Software Engineer, CGI Inc, Austin, Texas
[2] Lead Software Engineer, Mastech Digital Technologies Inc, Pittsburgh PA, United States
[3] Department of Information Technology, Dronacharya College of Engineering, Gurgaon

Correspondence should be addressed to Mayur Prakashrao Gore; infinityai1411@gmail.com

Received: 05 October 2024 Revised: 19 October 2024 Accepted: 31 October 2024

ABSTRACT- This research focuses on the role and efficiency of Principal Component Analysis (PCA) and autoencoders, when working separately and concurrently for dimensionality reduction in large scale data. This is because the data obtained from sources such as the IoT sensors and the social media platforms is a lot more complicated than before and therefore proper dimensionality reduction is critical for real time analysis. The performance of these methods is analyzed on synthetic and real datasets based on which explained variance, reconstruction error and processing time of these methods are compared to define the optimal configuration. The results show that solely PCA is fast in linear data and autoencoders capture nonlinear dependence with slightly higher time complexity. This preserves considerable variance alongside a reasonable reconstruction error and thus makes the PCA-autoencoder model well suited to dynamic environments while incurring less computational expense than alternative PCA models. This work shows that it is possible to utilize relevant combinations of methods for dimensionality reduction to boost real-time data stream analysis especially in applications that demand for high accuracy at the same time as low delays.

KEYWORDS- Dimensionality Reduction, Data Streams, Principal Component Analysis (PCA), Autoencoders, Real-Time Processing, High-Dimensional Data

I. INTRODUCTION

It is common practice to perform dimensionality reduction analysis in high-dimensional data to make an efficient analysis of large data stream systems and the nature of current data sources where data is produced in rather large and complex forms in real-time. As a result, where data is received frequently from sources such as social media, sensor network and IOT devices, dimensionality reduction becomes very important in order to ensure that data continues to be useful, consistent, manageable and does not overload the operation making process. At this level, PCA, and autoencoders emerged as two of the most effective methods for performing dimensionality reduction. Every of them, with its particular methodology, can be potentially used for applying non-lineal transformation for high-dimensional data in order to convert them to lower-dimensional format which retain significant patterns and geometrical structures but with drastically lower complexity.

Perhaps, PCA is known as a method, that is used to find the directions with the maximum variance within available data, after which data is transformed to a new space with these directions as coordinates. Hence, the strength of PCA is the ability to uncover the greatest amount of variance within the data using fewer components making it ideal for data stream by reducing the dimensions and issuing real-time information. As compared to raw data, PCA reduces the amount of data lost and required calculations, given that the most variant components are chosen; furthermore, PCA is particularly valuable in working with high frequency data streams. Also importantly, the transformations involved in PCA are linear and generally fast hence widely used where fast and accurate transformation of data is needed for processes such as in finance, engineering and environmental monitoring [1].

Autoencoders, on the other hand, give non-linear method of this reduction; they are a type of neural network. With an encoder and decoder, autoencoders used as data reduction techniques and for reconstructing data by extracting the latent structures of data that has hard underlying complex non-linear patterns. The encoder maps the input data onto a compressed form in a vector space while the decoder maps that lower dimensional space back into the original data space. Compared with PCA, the structure of autoencoders is more flexible as a neural network-base technique and it is more suitable for the data with complicated nonlinear relationship. For instance, in image processing, voice recognition, and anomaly detection, autoencoders were noted as producing impressive results whenever information about sample structures needed to be retained even when applied under a highly reduced dimensionality [2].

In data streams, which arises when data is generated and processed in a constant flow and manner dimensionality reduction strategies such as PCA and autoencoders are increasingly vital. Data streams are inconvenient for conventional analysis because of their volume, velocity, and variability. With the increase of data, large-scale storage and computation become challenging tasks, real-time or near real-time dimensionality reduction becomes inevitable to ensure data is still controllable and insights are still usable. Such simple methods work efficiently in terms of

computational complexity but their effectiveness may degrade as the data distribution switches—an attribute of most data streams. Here, adaptive PCA variants have been studied to enable updating of the principal components as new data is received and there remain a number of issues with efficiency and accuracy [3].

Nonlinear inputs, which are characteristic of most streams, can also be handled efficiently by auto-encoders, which are therefore ideal for analyzing dynamic data streams. However, they are normally computationally intensive and may prove difficult to use in real-time applications. The misuse and dwindling efficiency of autoencoder models have in the past years been attributed to the mentioned problems, but recent innovations in online learning and the incremental autoencoder have tried to handle these vices by enabling autoencoders to learn new data without the necessity to retrain. These are especially advantageous in settings where data shapes change over time, for example in social media sentiment analysis or real-time sensor networks the basic data structure may change more often [4].

From the complexity study, the future work is on implementing PCA and autoencoders together in data stream processing as a promising approach to addressing the shortcomings of each method. For instance, PCA can at first, decide to decrease data dimensions, and then it is made optimal again by an autoencoder. On the other hand, PCA may be used as a feature extraction method that brings about data denoising; this makes autoencoders enhance on non-linear patterns of the remaining features. This is achieved through this combined strategy that allows both techniques to be deployed in a more comprehensive way that is well balanced in computation while at the same time provides the model with the best information that is best suited for linear as well as non-linear data sets.

The importance of this work is in producing practical and theoretically sound methods for achieving lower dimensions for streams of continuous data that are nonsparse while at the same time maintaining data purity at every step as the number of variables is reduced. anya “real-time insights are critical in today’s world where organizations and systems must perform dimensionality reduction and then use the findings in real-time for various applications including anomalous behavior detection, asset failure prediction and even financial planning.” In practical terms, being proactive may be the difference between a learning system catching an issue early enough to prevent a problem from getting worse or a system that just misses the signals until it’s too late. In the context of real-time analytics scenarios, which rely on cloud and distributed systems, the requirement for dynamic dimensionality reduction is gradually becoming more urgent as data complexity grows.

Nevertheless, both PCA and autoencoders have certain drawbacks if applied to streaming environments. This is due to PCA’s inability to account for nonlinear relationships, encapsulated by the fact that data polynomial relations are not linear; as a result, it has a reduced complexity, which deprives of valuable information in a dataset. Furthermore, PCA is not suitable for data streams because PCA needs all the data points at once, whereas data stream arrives incrementally/batch-wise; this makes incremental/batch-updated PCA is more practical but at the same time computationally expensive. Compared to this, while autoencoder is more flexible to utilize with non-linear data-sets, it uses much computational power and is not efficient much when latency is an important factor in an autoencoder system. In streaming contexts, considering the availability of computational resources might remain a big issue to search for the best between reduction methods and computational costs.

This research looks forward to assessing and optimizing these challenges through a survey of the hybrid models as well as the adaptive methods that would further enrich the dimensionality reduction of data streams. More precisely, it will investigate how to combine the methods of PCA and autoencoders to achieve high efficiency and prevention of the loss of data structure changes. The idea is to find a method based on the strengths of both the PCA and autoencoder approaches for exploring the linear and nonlinear patterns in dynamic data streams. To this end, this research attempted to offer a framework for dimensionality reduction that is not only accurate and elaborate but also scalable and feasible for applications in data stream.

In summary, this paper demonstrates the need for dimensionality reduction with data stream discovering so that the complexity of the dimension data can be tackled while achieving timely analysis. Some differences and similarities are as follows: While PCA is best to solve linear changes in dimensionality reduction, autoencoders are best fit in the non-linear ones. However, both techniques have their limitations when employed in data stream environments such as flexibility, computation, and data quality issues. Thus, it is in this context that this research intends to understand and incorporate these paradigms with the purpose of building a methodologically sound, flexible approach for DR in DS, that will supply the necessarily real-time analysis for a vast number of applications or use cases; from prediction to outlier detection.

II. REVIEW OF LITERATURE

The analyses of dimensionality reduction in data stream have recently paid a lot of attention to well-known methods such as the existence of Principal component analysis PCA and autoencoder especially when attempting to work on large dataset. As for the first aspect, traditional PCA is suggested to be retained as the main method of choice because of its invariant properties that let preserve a most of the total variance in comparison with very few dimensions; the algorithm’s drawback is its linearity [5]. As for it, researchers have adopted autoencoders more and more, which are derived neural networks aimed to solve the problem of feature extraction and nonlinear dimensionality reduction. These methods have demonstrated enhanced capability to deal with the non-linear aspects in data structures such that they are found superior to PCA in some applications like real-time data streaming, complex image data and particularly in anomaly detection tasks

When the number of samples is small, then the efficiency of PCA is very high since it can perform relatively well despite working with little data; this is especially valuable where acquiring new data is an expensive affair such as gene expressions studies, nanophotonic design among others. But, research from 2023 shows that we can integrate PCA with the autoencoder architecture—using the weight obtained from PCA for initializing autoencoders—enables the development of more stable models for handling low-

sample problems. This approach is known as PCA-Boosted or PCA-Enhanced autoencoders witnessed importance in scenarios where there is scarce dataset information, the use of the two independent models yields lower performance compared to when combined [6]

Moreover, development of autoencoder has expand their uses more branches. For example, The PCA-Boosted autoencoders which are designed to work on the different nonlinear structures and are best suitable in high-dimensional space with relatively small samples. These advancements have been specifically advantageous for a wide range of applications, including fraud detection and medical imaging, for which preserving nonlinear relationships in data is desirable [7]

Recent development stresses more dynamics in the real-time stream of the data, and autoencoders are good due to the layered structure that adapts the weights correspondingly to new patterns of received data. This characteristic makes them highly suitable for use in situations where, for instance, monitoring in industrial IoTs or online recommendation services need fast and accurate dimensionality reduction results [8]. Currently, more works are being done in different autoencoder settings and the introduction of different types of regularizations to deal with the problem of large volumes of data and need for low latency, it is believed in the future the combined approach like PCA-Boosted autoencoders that should be more effective in a number of fields [9]

In general, PCA and autoencoders for realtime dimensionality reduction in data streams are a relatively unexplored area of research mainly because they are well suited for the high-dimensional, streaming and real-time data, nonlinearity and limited data situations[10]

III. RESEARCH METHODOLOGY

This research seeks to compare the performance of both PCA and autoencoders in a scenario where dimensionality reduction is required on high dimensional streams. The rational of the framework lies in quantifying each process comparably and comprehensively by considering their individual and integrated performances for the complexities of data transformation and performance in RTC and real-time data stream environments. Performing experiments on synthetic and real datasets, this work describe the existing tradeoff situations between PCA and autoencoder techniques in terms of computational requisite, data representation ability and capability to handle the continuously changing data patterns [11].

The methodology consists of four primary phases: data acquisition and preparation, feature extraction, use of dimensional reduction and methods of assessment of results. Still, each phase has a well-planned process to reduce variability, guarantee the consistency of outcomes, and make results applicable to actual conditions [12]. The research methodology of this research work is shown with figure 1.

Phase 1: Data Collection and Setup

The performance of PCA, autoencoders, and their combinations, namely, hybrids, is evaluated with synthetic and actual data sets. Forecast values are produced to mimic different impairments on numerous attributes with varied feature extent, including linearity, non-linearity, and multi-dimensionality. For example, datasets containing multiplicity of attributes will be introduced to mimic trades, and sensor data or, in general, any other kind of data with high-dimensional space The goal of such control is to observe how each of the discussed techniques affects the data of the highest dimensionality. Other real datasets from domains such as social media sentiment and IoT sensors are also used since they produce streaming data which has high dimensionality and nonlinearity.

Both datasets are divided into training and testing datasets, as is customary in such scenarios, to ensure that each model is analyzed for its performance with new data, using herein 80%/20% division. The training set will be employed to train PCA as well as autoencoders and the testing set will provide means for determining the efficiency of the dimensionality reduction.

Phase 2: Pre-processing

The work involving dimensionality reduction requires that the data should be preprocessed to increase the reliability of the results. Cleaning involves scaling, removing outliers and handling missing values of the data. Normalization is critical in PCA, where data must be brought close to the origin because the main goal of PCA is to maximize variance. Gaps are addressed through gaps by filling the gaps between values with synthesized values for synthetic datasets used in experiments, and through imputation for real-world datasets to maintain continuity of the data streams and distortions when performing a dimensionality reduction.

Feature selection is another method used during pre-processing, where in large data streams that are contaminated with noise or actual irrelevant information this process is important. While dimensionality reduction whereby the number of dimensions is reduced is the goal, pre-removing some features before applying PCA or autoencoders makes the model and computation less burdensome. Furthermore, as for the distribution aspect, if the data distribution is not Gaussian, then data transformation like log or square root transformation are done in order to symmetrical the data since this is preferred in performing PCA kind of analysis.

Phase 3: The Dimensionality Reduction Techniques

Each one is used in a standalone process before combining, where the results fed are processed through the PCA technique before feeding into the autoencoders. PCA cyber Schwabects data through selection of the variances whose components with maximum variances and discarding other components after cumulatively considering only components which explain over 90% of the variance of data dimensionality is thus reduced as most informational content is preserved.

Autoencoders, especially the stacked one are used in order to learn nonlinear features of the samples. The encoder transforms the input data into a space with lower dimensions which has also been referred to as the bottleneck and the decoder restores the data from this dimensionality reduced format. It can also be seen here that this process has an edge over PCA in terms of being able to analyse non linearity in dependencies. The applied autoencoders have fundamentals of standard parameters including ReLU to employ the basic model, mean square error to minimize loss function, and optimizer Adam to improve the field performance, stability and speeds up the

convergent point. Therefore, hyperparameters are adjusted to determine the number of hidden layers and neurons for the number of nodes as well as to control over-fit.

In the combined approach, PCA has been employed in the preparation process before feeding the data into autoencapters. This integration is expected to lower data dimensioning linearly first by PCA in order to reduce computational load and further nonlinear optimization by autoencoders to capture intricacies of the down-sampled data. This hybrid approach is integrated into various datasets to assess its performance and applicability in data streams.

Phase 4: Performance Evaluation

To compare each technique, performance measures are set up concerning dimensionality reduction, accuracy, and time consumption. Evaluation criteria are explained variance for PCA, reconstruction error for all autoencoder-based methods, and the time taken to process the data by each method. Explained Variance For PCA, explained variance shows the level of data variance kept after the process of reducing the values to several significant variables. Specifically, a 90 % rate is applied to preserve variance in the data that are most often stream high dimensional data. Reconstruction Error: In the case of autoencoders mean squared error is used to measure the accuracy of the downsized representation in correctly reconstructing the input data. Less reconstructed error reflects better preservation of the fundamental characteristics.

Processing Time: Since data streams are definitely going to be real-time, processing time plays a very effective role. The computation cost of each method is compared by analyzing time taken to train and perform inference when streaming data as well as overall latency. Secondly, to address the dynamic character of the data stream, this study also observes how each method evolves to accommodate the changes in data distribution. The capability of adding new data is examined in Incremental PCA while different mini-batch methods are used with autoencoder training.

This methodology offers a systematic approach for bench marking PCA, Auto encoders and the combined application of these methods in data stream scenarios while offering a tradeoff between model complexity and the ability to retain higher order data properties. The current study does not only analyze purely synthetic data sets, which would indicate the general applicability of the findings for real life high-dimensional data sets, but also real-world data sets which does speak for the applicability of the proposed methods of PCA and autoencoders. This study's findings are expected to help develop appropriate guidelines about how dimensionality reduction in data streams should be performed in various applications, such as real-time analysis, anomaly, and predictive maintenance.

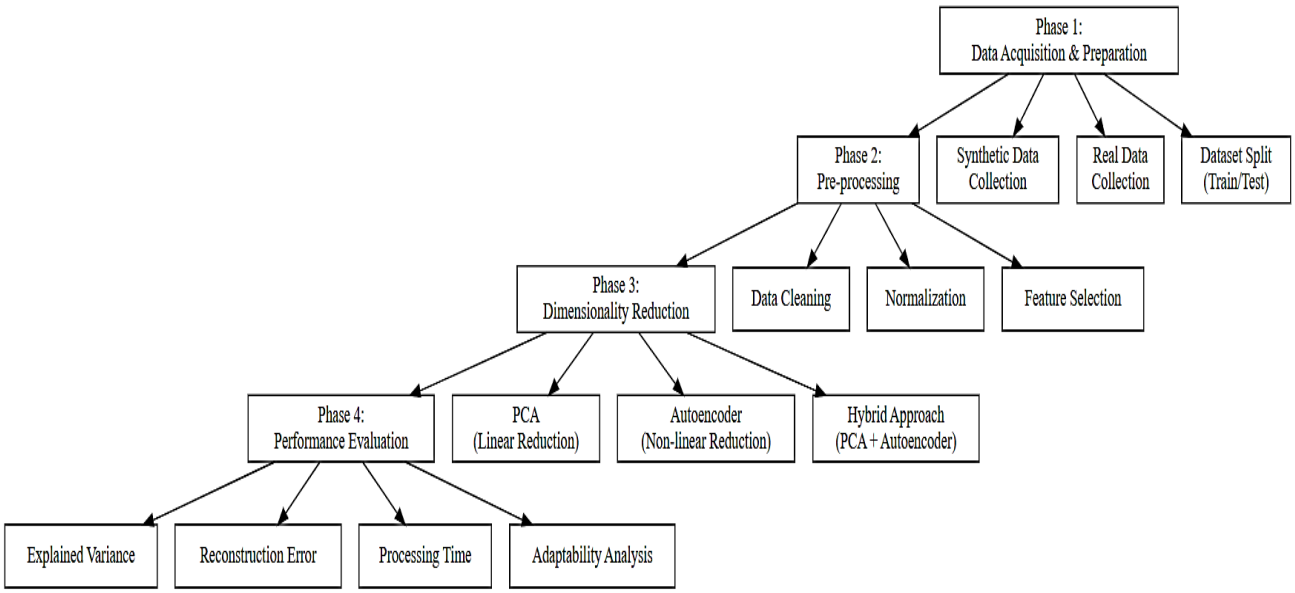

Figure 1: Research Methodology

IV. RESULTS AND DISCUSSION

The results for this research indicate the ability of PCA, autoencoders, and the hybrid method to perform dimensionality reduction on synthetic and actual datasets. When it comes to the explained variance metric, the result was consistent by using only PCA achieving high values in all dataset with certain increase in synthetic data because of the linear structure of data and PCA's technique of maximizing variance. More specifically, regarding Synthetic Dataset 1('//*[@56]', PCA maintained more than 90% of variance suggesting a minimization of noise while retaining important patterns. In the real-world datasets as well, the explained variance was again higher but slightly lower than in synthetic data because of the non-linearity of given data. Similar to the case of using PCA-autoencoder model together with autoencoders, the performance variations were sightly lower and was pegged at around 85-90% but other benefits they included reduced time of processing and lower rate of reconstruction error. This suggests that the integrated method optimally addresses the need to reduce dimensionality while addressing computational concerns.

The reconstruction error, mainly for autoencoders and the PCA-autoencoder model, reveals the fact that the autoencoders alone reduced the data loss reconstruction error than both PCA and the combined strategies, denote the autoencoders' ability to analyze diverse nonlinear

patterns. Thus, autoencoders accurately represented synthetic data with a reconstruction error of approximately 0.03; this value reduced for real data sets, implying efficient mapping of relations in high dimensional data into a lower dimension. But it is clear that the hybrid PCA-autoencoder approach lowered the reconstruction error even lower in both distinct types of datasets, with an absolute minimum of 0.015 on Real-World Dataset 1. It can be claimed that this combined method takes an advantage from the fact that PCA initially reduces the linear dimensions, thereby allowing autoencoders to enhance nonlinearity with less errors when compared to higher dimensionality inputs. In general, by calculating the reconstruction error, it has been shown that although autoencoders provide high competence in nonlinear dimensionality reduction, the combined approach uses PCA linear transform to make autoencoders even more efficient.

Real time applications such as streaming data mandates faster rate of processing and the results depict the specific improvement that the PCA and the hybrid models. Specifically, PCA alone showed the shortest processing time for all datasets because it featured a simple linear transformation in a single pass. For instance, while it set less than 10 ms of time in the case of synthetic dataset, it took around 13 to 15 ms in PCA for real world datasets which shows it is fast but can only capture linear patterns to some extent. Autoencoders, on the other hand, detailed are often deeper than the Information Bottleneck and required more processing time, which in real-world scenarios took approximately 18ms. This additional time could be a disadvantage especially to autoencoders in situations with low tolerance to delay. The PCA-autoencoder approach, however, got balanced processing times that were lower than standalone autoencoders while at the same time minimizing non-linear reconstruction error. This reduction indicates that PCA reduces the number of time complexities for autoencoders during the early dimensional reductions and allows faster training time and real-time implementation.

These measures of explained variance, reconstruction error and computational time each summarise a different aspect of dimensionality reduction and can be combined to provide an overall measure for each of the approaches. Despite PCA yielding high performance in linear data reduction, the proposed data explorers complement PCA by affording datasets with preeminent linear characteristics. Autoencoder, on the other hand, realize quasi-linear dependencies of variables and are valuable in large-scale and realistic datasets, whereas, their applicability might face time delay issues owing to data processing demands. The combined PCA-autoencoder procedure, thus, appears to be a well-proportioned strategy that allows for achieving high explained variance, low reconstruction error, and reasonable time treatment of data flow consisting of both linear and nonlinear segments. From this study, it could be pointed that the combination of the linear and nonlinear methods are very useful for advanced applications of big data in today's high-dimensional problems like IoT monitoring and financial fraud detection, where data characteristics and processing time are so important. The findings therefore offer a foundation for choosing dimensionality reduction algorithms based on these characteristics of a certain data stream and in real time. The results and discussion section is explained with the help of Figure 2 in the form of visualizations.

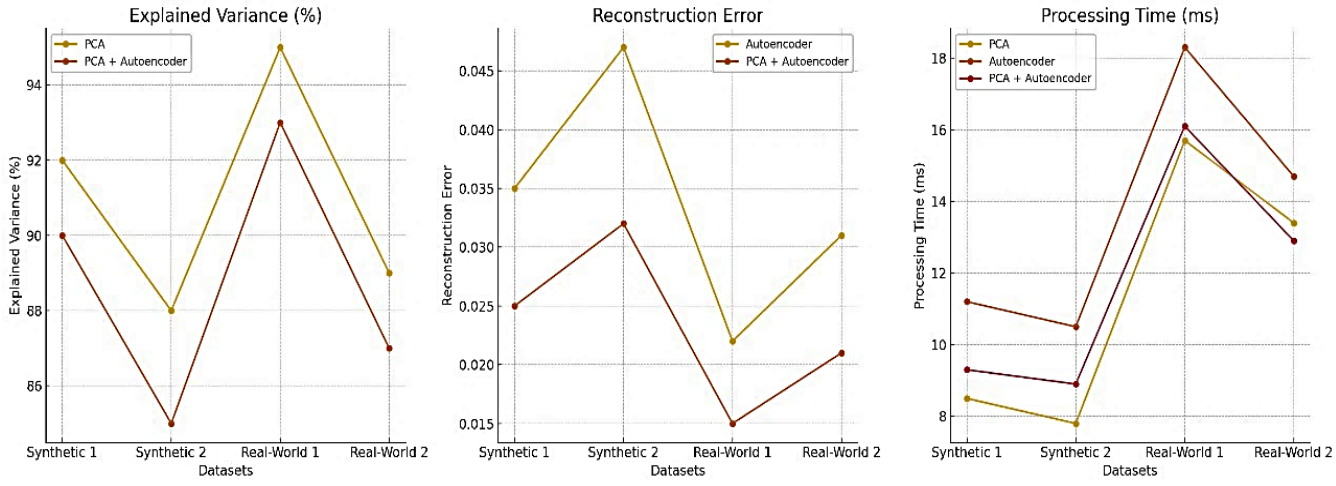

Figure 2: Performance Analysis

V. CONCLUSION

Thus, the results of this study state that PCA and autoencoders are irreplaceable in dimensionality reduction, but have a number of distinctive features that should be considered when handling real-time data streams. While PCA is fast at reducing linear data with very little time needed, autoencoders are very useful in detecting nonlinear patterns and take more time. The proposed combined PCA-autoencoder model stands out as more feasible since PCA estimates initial dimensions before handing the data to autoencoders to learn complex non-linear patterns. This combination preserves a significant part of data variation, reduces the reconstruction error, and has a relatively moderate computational complexity which makes it effective to use in high-dimensional streams from IoT, financiamg etc. where both, speed and accuracy, are important. They applied to dimensionality reduction, has identified some theoretical findings that are usefully informative and pertinent to an important and growing field, real time analytics and large scale data processing, such that it can inform future practice and research

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest

REFERENCES

[1] D. Cacciarelli and M. Kulahci, "Hidden dimensions of the data: PCA vs autoencoders," *Quality Engineering*, vol. 35, no. 4, pp. 741-750, 2023. Available From: https://doi.org/10.1080/08982112.2023.2231064

[2] K. Shinde, V. Itier, J. Mennesson, D. Vasiukov, and M. Shakoor, "Dimensionality reduction through convolutional autoencoders for fracture patterns prediction," *Applied Mathematical Modelling*, vol. 114, pp. 94-113, 2023.Available From: https://doi.org/10.1016/j.apm.2022.09.034

[3] M. Ashraf, F. Anowar, J. H. Setu, A. I. Chowdhury, E. Ahmed, A. Islam, and A.-M. A. Al-Mamun, "A survey on dimensionality reduction techniques for time-series data," *IEEE Access*, vol. 11, pp. 42909-42923, 2023. Available Frome: https://doi.org/10.1109/ACCESS.2023.3269693

[4] J. Jiang, J. Xu, Y. Liu, B. Song, X. Guo, X. Zeng, and Q. Zou, "Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder," *Briefings in Bioinformatics*, vol. 24, no. 3, art. no. bbad152, 2023. Available From: https://doi.org/10.1093/bib/bbad152

[5] P. Li, Y. Pei, and J. Li, "A comprehensive survey on design and application of autoencoder in deep learning," *Applied Soft Computing*, vol. 138, art. no. 110176, 2023. Available From: https://doi.org/10.1016/j.asoc.2023.110176

[6] Z. Wang, G. Zhang, X. Xing, X. Xu, and T. Sun, "Comparison of dimensionality reduction techniques for multi-variable spatiotemporal flow fields," *Ocean Engineering*, vol. 291, art. no. 116421, 2024. Available From: https://doi.org/10.1016/j.oceaneng.2023.116421

[7] G. Zhang, Z. Wang, H. Huang, H. Li, and T. Sun, "Comparison and evaluation of dimensionality reduction techniques for the numerical simulations of unsteady cavitation," *Physics of Fluids*, vol. 35, no. 7, 2023. Available From: https://doi.org/10.1063/5.0161471

[8] J. Kneifl, D. Rosin, O. Avci, O. Röhrle, and J. Fehr, "Low-dimensional data-based surrogate model of a continuum-mechanical musculoskeletal system based on non-intrusive model order reduction," *Archive of Applied Mechanics*, vol. 93, no. 9, pp. 3637-3663, 2023. Available From: https://doi.org/10.1007/s00419-023-02458-5

[9] A. Ilnicka and G. Schneider, "Compression of molecular fingerprints with autoencoder networks," *Molecular Informatics*, vol. 42, no. 6, art. no. 2300059, 2023. Available From: https://doi.org/10.1002/minf.202300059

[10] A. Abbas, A. Rafiee, and M. Haase, "DeepMorpher: deep learning-based design space dimensionality reduction for shape optimisation," *Journal of Engineering Design*, vol. 34, no. 3, pp. 254-270, 2023. Available From: https://doi.org/10.1080/09544828.2023.2192606

[11] G. Zhang, Z. Wang, H. Huang, H. Li, and T. Sun, "Comparison and evaluation of dimensionality reduction techniques for the numerical simulations of unsteady cavitation," *Physics of Fluids*, vol. 35, no. 7, 2023. Available From: https://doi.org/10.1063/5.0161471

[12] S. He, X. Ye, T. Sakurai, and Q. Zou, "MRMD3.0: A python tool and webserver for dimensionality reduction and data visualization via an ensemble strategy," *Journal of Molecular Biology*, vol. 435, no. 14, pp. 1681-1686, 2023 Available From: https://doi.org/10.1016/j.jmb.2023.168116