Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 03 March-2025, Page No.- 4099-4111

DOI: 10.47191/etj/v10i03.20, I.F. – 8.482

Expanding the Horizons of Principal Component Analysis: Versatile Applications from Environmental Monitoring to Chemometrics

Firdaws Rizgar Tato[1], Hajar Maseeh Yasin[2]

[1,2] Akre University for Applied Sciences Technical College of Informatics Department of Information Technology

ABSTRACT: Principal Component Analysis (PCA) is a popular statistical method for large dataset analysis and has established itself in many fields, ranging from environmental modeling to spectroscopy. In this work, we bring to light its use in monitoring NO2 air pollution from satellite data during the lockdown periods of COVID-19, incorporating the strengths of Weighted PCA and Rescaled PCA to achieve improved predictability. Additionally, PCA has been applied in resonant ultrasound spectroscopy to optimize measurement points, especially in samples with complex geometries, demonstrating its effectiveness in reducing data points of collection while maintaining accuracy. Further, PCA's application coupled with classification methods like LDA and SVM has effectively determined the geographic origin of Indonesian coconuts, demonstrating its effectiveness in enhancing classification accuracy and chemometric analysis. The versatility of PCA is also evidenced in its use in clustering high-dimensional data through Adaptive Local PCA, which employs a neural network-based approach with adaptive learning rates to enhance clustering quality in dynamic data environments. These examples show the flexibility and utility of PCA in big data analysis across different fields, and a greater application of PCA in gaseous pollutant analysis and other complex data issues is needed.

KEYWORDS: Principal Component Analysis (PCA), Dimensionality Reduction, High-dimensional Data, Pattern Recognition, Big Data Analytics, Spectroscopy, Adaptive Learning.

1. INTRODUCTION

Principal Component Analysis (PCA) has been extensively utilized for analyzing big datasets to explain intricate patterns and gain useful information [1]. applied PCA for analyzing the NO2 air pollution changes during the COVID-19 lockdown periods based on satellite images. They proposed two novel PCA models: Weighted PCA (WPCA) and Rescaled PCA (RPCA), both of which were found to be efficient in forecasting changes in the level of air pollution. The research determined that Principal Component Analysis (PCA) serves as a dependable method for recognizing and forecasting air pollution trends, advocating for its utilization in analyzing other gaseous pollutants. Within the framework of resonant ultrasound spectroscopy, Beardslee[2]. investigated the implementation of PCA to enhance the selection of measurement points, especially in relation to specimens with intricate geometries. Their approach minimized the number of measurement points utilized with preserving accurate spectral data, thereby making data processing more efficient. This was one illustration of how PCA can be applied to improve data acquisition and processing in complex systems. PCA has also seen application in classification and chemometric analysis. [3] integrated PCA with LDA and SVM classifiers for determining the geographic origin of Indonesian coconuts. The model had high accuracy, thereby demonstrating the effectiveness of PCA in classification and its capability in maximizing visualization results in chemometric research.[4] presented Adaptive Local PCA for improving clustering of high-dimensional data. Their approach employed a neural network-based algorithm with adaptive learning rates and ranking metrics, which improved the quality of clustering, especially for non-stationary data distributions. This work highlighted PCA's capacity to adaptively process complex and dynamic data environments. These investigations illustrate the efficacy and versatility of PCA in big data analytics, ranging from environmental modeling and spectroscopy to classification and high-dimensional data clustering. PCA continues to be a powerful statistical method for uncovering informative patterns and interpreting large datasets[5].

2. BACKGROUND THEORY

2.1. Big Data Fundamentals

Big data is greatly important in today's digital age for revolutionizing industries and advancing new ideas in different industries. Big data, due to its enormous a
mount, different types, and high velocity, requires advanced methods for processing and handling it to extract valuable information. Methods like Principal Component Analysis (PCA) are greatly needed for reducing dimension and facilitating the understanding of big data. PCA is very useful for unsupervised data analysis. It determines principal factors and is used in areas like finance and biomedical research [6]

. Big data brings issues like privacy and security, especially in cloud and Internet of Things (IoT) environments. In such environments, techniques like encryption are used to improve the protection of data during the processing of data using techniques like PCA and LDA. This helps to protect the privacy of sensitive data [7] . Additionally, big data enables business model innovation by the capacity of businesses to reimagine processes and create new value propositions, founded on thoroughly analyzed and intelligent data [8] .

2.2. Statistical Foundations

In the context of data analysis, a grasp of fundamental statistical concepts such as variance, covariance, and the application of eigenvalues and eigenvectors is valuable. Variance quantifies the extent to which data points differ from their mean, offering insightful information regarding the spread of a dataset. Covariance builds upon this by quantifying the extent to which two variables change in tandem and thereby offers an understanding of the correlation between variables in a dataset. Variance and covariance together are the building blocks of covariance matrices, which capture the covariance of more than one variable and play a central role in multivariate statistical analysis [9] [10] . Aside from this, eigenvalues and eigenvectors also play a significant role in data transformation through methods such as Principal Component Analysis (PCA). PCA employs them for the purpose of decreasing the dimensionality of data sets without sacrificing their most important aspects. Eigenvalues tell us the significance or magnitude of every axis in the new feature space established, and eigenvectors specify their orientations. This conversion assists in the recovery of the highest amount of variance, or information, present in the data and hence provides a simpler and improved analysis. Moreover, this process not only makes large volumes of data easy but also makes data interpretation used in different areas, ranging from engineering to social sciences, more transparent and effective [9] [10].

2.3. Linear Algebra Concepts

Principal Component Analysis (PCA) uses basic ideas of linear algebra, i.e., diagonalization and orthogonality, to successfully deal with and reduce the complexities involved in high-dimensional datasets. Through the process of converting the covariance matrix to a diagonal matrix using eigenvectors, PCA guarantees that the resulting principal components have the highest possible variance. This step is crucial in dimensionality reduction while retaining valuable information and hence plays a fundamental role in various areas, including finance and biomedical research [2] [11] . The orthogonality of eigenvectors ensures the principal components are uncorrelated, a feature that plays a key role in ensuring no information is lost in the process of dimensionality reduction.

To improve the performance and robustness of Principal Component Analysis (PCA), particularly in the presence of outliers and noisy data, various robust PCA methods have been proposed. These methods modify the covariance matrix estimation to limit the effects of anomalies, thus leading to principal components that are more representative of the intrinsic data structure [12] . Furthermore, sparse PCA has been proposed to introduce sparsity in the components. This adjustment not only helps maintain the interpretability of the results but also highlights the most relevant features, thus enhancing the applicability of Principal Component Analysis (PCA) to real-life scenarios that require simple and accurate feature selection [13] .

Besides, newer approaches such as the utilization of 'data nuggets' have been suggested for effective management of very large datasets. This approach summarizes data into manageable chunks while maintaining the structural information of the dataset, thereby enabling the use of Principal Component Analysis (PCA) where the classical approach is computationally infeasible [11] [14] .

2.4. PCA Fundamentals

Principal Component Analysis (PCA) is a fundamental statistical method in data analysis and machine learning that is utilized to highlight variations and uncover significant patterns within a dataset. PCA reduces the complexity of high-dimensional data by projecting it onto a lower-dimensional space that retains most of the significant information. The method starts with normalizing the dataset such that each variable has zero mean and unit variance to provide equal weightage to them in analysis. The covariance matrix is then computed to identify correlations between variables. By computing the eigenvectors and eigenvalues of this matrix, PCA effectively reduces the dimensionality of the data without losing its most significant variability. This step involves choosing the most important features, i.e., the leading k eigenvectors that encompass the prevailing variance and hidden patterns within the data [15] [11].

Assumptions PCA makes are the assumption of linearity, which states that the principal components are linear combinations of the original variables. The assumption is crucial as it dictates that directions of maximum variance will be most informative regarding the data structure. Secondly, PCA also assumes that the most significant variances contain the most important information about the data, thus the transformation to a space where the maximum of these variances is obtained. Normalization or scaling of data is also important before applying PCA so that all the variables contribute equally to the analysis and remove any bias towards variables with inherently larger scales. Standardization is required for a fair comparison and appropriate contribution of features based on various scales of measurement [16] [17].

The significance of PCA is in its ability to reduce data dimensions without loss of required information, thus rendering it valuable for exploratory data analysis, signal denoising, and as a pre-processing step before applying other machine learning techniques. The reduction in dimensionality enables improved data visualization and promotes more efficient processing of the data in subsequent analysis, which in large datasets, where excessive dimensionality may render effective analysis and interpretation cumbersome [11] . The use of Principal Component Analysis (PCA) encompasses an extensive variety of areas, from genomics to finance, where the identification of the underlying structure of complicated data is crucial [18] [19] .

3. LITERATURE REVIEW

3.1. Historical Perspective and Evolution

Li et al. (2024) [16] discuss that Principal Component Analysis (PCA) was initially developed as a statistical method for dimensionality reduction and data interpretation. Its application has expanded to high-dimensional data in contemporary big data contexts, ranging from basic statistical tests to sophisticated machine learning and artificial intelligence models. Developments in scalability and noise removal have enabled PCA to become more effective and precise in data analysis.

Ikegwu et al. (2024) [6] say the field of big data analytics for climate change has developed from conventional data processing techniques to sophisticated machine learning and artificial intelligence techniques. This was on account of the increasing amount and sophistication of climate data, such as satellite images and sensor networks, hence the need for more effective data processing and predictive modeling techniques.

Wang et al. (2024) [20] explain that PCA was initially a basic statistical method but evolved into a complex method combined with machine learning and deep learning. PCA is now handling intricate high-dimensional data with greater effectiveness and accuracy in many fields, including wind power forecasting.

Perez and Toraman (2024) [21] explain how PCA evolved from simple statistical methods to advanced tools in combination with gas chromatography and spectrometry. The innovation increased accuracy in analyzing complex chemical reactions, especially polypropylene decomposition and pyrolysis.

Baidya et al. (2024) [22] explain that PCA and multivariate statistical analysis evolved from basic statistical methods to advanced techniques integrated with machine learning. What was originally used for covariance analysis is now used to process complex geochemical data, making mineral exploration even more efficient in identifying fluid sources and deposit types.

Christensen et al. (2023) [23] explain that PCA evolved from a statistical method to a sophisticated methodology coupled with Density Functional Theory (DFT) for improving CO and HCOOH selectivity classification in electrochemical CO2 reduction.

Mehmood et al. (2024) [24] discuss that the development of big data analytics in manufacturing SMEs originated from conventional data processing approaches, yet it has proceeded to incorporate dynamic capabilities such as green innovation. The development illustrates an increasing demand for sustainability and competitiveness, and thus enhanced efficiency in operation and environmental performance.

Faaique (2024) [12] explains that big data astronomy evolved from simple data processing to complex machine learning as a result of huge datasets from projects like SDSS and LSST.

Wu et al. (2024) [25] describe how tourism and hospitality forecasting developed from conventional statistical techniques to sophisticated machine learning and artificial intelligence techniques. This shift was driven by growing availability of real-time, high-frequency data that had been collected from online sources, thereby improving the accuracy of forecasts and facilitating informed strategic business decision-making in the industry.

Zhao and Liang (2024) [26] explain how the teaching of foreign languages evolved from traditional methods to using big data models like cross-lingual embeddings and self-encoders to improve bilingual translation and learning outcomes.

Sharma and Gurung (2024) [27] describe how predictive maintenance has developed from conventional preventive and reactive strategies to sophisticated systems involving big data analytics and machine learning technology. This development has been motivated by the demand for minimizing unexpected downtime and enhancing maintenance in highly automated production setups.

Jin et al. (2024) [18] outline how big data analytics in additive manufacturing (AM) evolved from basic data processing to advanced machine learning and digital twins, with improved process optimization, quality control, and predictive maintenance.

Migenda et al. (2024) [4] state that Principal Component Analysis (PCA) has evolved into adaptive local PCA to improve clustering performance for high-dimensional, dynamic data sets. This extension uses variable learning rates with possible functions to deal with non-stationary data more effectively.

Gandaglia et al. (2024) [28] explain that big data analytics in prostate cancer treatment evolved from traditional clinical decision-making to advanced real-world data analysis. The evolution was driven by the need to improve conservative management strategies using large international databases, with better patient characterization and outcome prediction.

Beardslee et al. (2024) [2] state that Principal Component Analysis (PCA) designed to optimize resonant ultrasound spectroscopy (RUS) by reducing arbitrary point choice and measurement time, while also boosting accuracy for samples with complex geometries.

Nyangon and Akintunde (2024) [29] explain how PCA evolved into an advanced electricity price forecasting technique, improving accuracy and grid control in renewable energy markets through the handling of heteroskedastic noise.

Xie et al. (2024) [13] explain that True Sparse PCA (TSPCA) was created in order to maximize the use of sensors in virtual metrology for cost savings and efficiency by reducing the number of sensors required.

Mukherjee et al. (2024) [30] explain that multi-omics data analysis evolved from classical biologic studies to advanced integration methodologies through the power of big data and machine learning, with augmented insights into disease mechanisms and precision therapeutics.

Ali et al. (2023) [15] explain that big data analytics in ICS IDS evolved from traditional security practices to AI-based models for dealing with advanced cyberattacks, enhancing threat detection and real-time analysis.

3.2. Empirical Studies and Applications

Li et al. (2024) [31] illustrate that PCA has wide-ranging applications in fields including image processing, genomics, financial modeling, and environmental science. PCA effectively reduces dimensionality, improves model accuracy, and reveals patterns in multidimensional data. In machine learning, PCA optimizes accuracy and reduces overfitting by preprocessing data.

Ikegwu et al. (2024) [6] observe that heightened big data analytics enhances climate change research through more accurate forecasts of season changes, weather extremes, and health risks. AI and machine learning facilitate real-time processing and accurate climate modeling, which enhances environmental monitoring and disaster management**.**

Wang et al. (2024) [10] note that PCA enhances the precision and effectiveness of offshore wind power forecasting by reducing dimensionality and eliminating noise. It optimizes prediction models like BiLSTM with excellent precision and stability for wind power systems.

Perez and Toraman (2024) [21] note that PCA is used to reduce dimensionality and identify patterns in chemical reaction data to enhance predictive modeling and understanding of reaction mechanisms. It finds wide use in environmental studies, chemical kinetics, and materials science.

Baidya et al. (2024) [22] highlight that PCA is used for analyzing hydrothermal biotite chemistry to distinguish between types of ore deposits and fluid sources. It enhances predictive modeling and facilitates greater precision of classification in geochemical data sets in mineral exploration and environmental studies.

Christensen et al. (2023) [23] point out how PCA classifies CO vs. HCOOH selectivity in examining reaction pathways to enhance the design and optimization of catalysts in CO2 reduction.

Mehmood et al. (2024) [24] point out how big data analytics improves economic and environmental performance in manufacturing SMEs through facilitation of green innovation. Empirical research proves its capability for enhancing decision-making, efficiency in operations, and sustainability practices, translating into improved economic performance and reduction of environmental impact**.**

Faaique (2024) [12]explains that big data analytics enhances pattern recognition, noise reduction, and predictive modeling in astronomy, accelerating celestial discovery and exploration.

Wu et al. (2024) [25] point out that big data analytics is heavily applied in the tourism and hospitality sectors for forecasting demand, enhancing operational effectiveness and customer satisfaction. Empirical studies indicate its suitability in utilizing web-based volume information, social media metrics, and online text information to improve accuracy in prediction and respond to fast-changing market developments.

Zhao and Liang (2024) [26] highlight that big data corpus analysis enhances vocabulary, sentence comprehension, and reading ability, leading to more effective foreign language teaching.

Sharma and Gurung (2024) [27] highlight how predictive maintenance increases equipment availability and reduces maintenance cost in manufacturing by using machine learning algorithms to foreca
st failures prior to their occurrence. South Korean applications illustrate how it can be used to increase operational efficiency and minimize disruption.

Jin et al. (2024) [18] note that big data analytics enhances material analysis, design optimization, defect detection, and sustainability in AM. Integration of digital twins enables real-time monitoring and predictive maintenance, which improves product quality and operational efficiency.

Migenda et al. (2024) [4] note that adaptive local PCA is better than traditional approaches at clustering high-dimensional data, especially real-time data streams. Its adaptive learning rates enhance accuracy and efficiency, with efficacy for dynamic data environments.

Kovács and Haidu (2023) [1] point out that WPCA and RPCA both successfully modeled the NO2 concentration variability with Sentinel-5P data, demonstrating long-term effects of COVID-19 on air quality with precise predictions reinforced by ground observations.

Gandaglia et al. (2024) [28] note that big data analysis effectively characterizes prostate cancer patients undergoing

conservative management, reporting patterns of comorbidity, hospitalization rates, and symptomatic progression patterns. It improves decision-making and personalized treatment through the use of real-world evidence in different international datasets.

Beardslee et al. (2024) [2] point out that PCA successfully selects the optimal RUS measurement points, reducing the number needed without loss of accuracy and improving elastic constant inversion for challenging geometries.

Nyangon and Akintunde (2024) [29] highlight that PCA enhances day-ahead price prediction by reducing skewness and streamlining regression models, making it easier for renewable energy to grow in integrated energy markets.

Xie et al. (2024) [13] highlight that TSPCA efficiently reduces the number of required sensors without sacrificing predictive performance in semiconductor manufacturing, thus cutting costs and ensuring accurate data acquisition.

Mukherjee et al. (2024) [30] note that multi-omics improves disease classification, diagnosis, and therapeutic targeting through the integration of high-throughput data, effectively modeling disease networks and validating drug targets.

Ali et al. (2023) [15] note that AI-based IDS achieve 99% detection accuracy for ICS cyberattacks, enhancing the security of critical infrastructure through real-time threat detection.

3.3. Comparative Analysis

Li et al. (2024) [31]compare PCA with other dimensionality reduction methods like LDA and t-SNE. PCA is favored because of variance preservation and simplicity, while LDA is better for supervised use. t-SNE is more appropriate for visualization but lacks interpretability. The choice is based on the data and analysis objectives.

Ikegwu et al. (2024) [6] compare analytics methods, pointing out that deep learning and machine learning are highly accurate but at the expense of high computing power, while traditional methods are quicker but less effective with complicated data. Where to apply is dictated by application needs, taking into account accuracy, speed, and data complexity.

Wang et al. (2024) [20] show that PCA combined with complex algorithms like SSA and VMD outperforms traditional methods. Hybrid models of PCA and BiLSTM yield improved prediction precision and efficiency for complex time-series data.

Perez and Toraman (2024) [21] illustrate that PCA in combination with two-dimensional gas chromatography achieves higher resolution and pattern recognition than is possible with standard methods. The combination enables better analytical accuracy in analyzing chemical mixtures and pyrolysis products.

Baidya et al. (2024) [22] contrast PCA with other multivariate techniques such as PLS-DA, demonstrating that PCA is efficient in dimensionality reduction and pattern identification. Nevertheless, PLS-DA provides superior classification accuracy in intricate mineralogical data sets. Integration of PCA with machine learning algorithms enhances predictive capabilities.

Christensen et al. (2023) [23] demonstrate that PCA is superior to conventional DFT methods through the application of multiple descriptors in predicting precise metal catalyst behavior.

Mehmood et al. (2024) [24] contrast big data analytics with the traditional decision-making process, stipulating those analytics provides more precise outputs and greater predictability. Analytics surpasses traditional approaches in competitiveness and sustainability through dynamic capabilities, including green innovation and proper resource allocation.

Faaique (2024) [12] illustrates how machine learning algorithms are more efficient and precise than traditional methods for analyzing astronomical data, enabling detailed cosmic research.

Wu et al. (2024) [25] compare conventional techniques of prediction with big data analytics, showing how machine learning and AI models produce better accuracy and versatility. These newer techniques surpass conventional models as they use unstructured data from social media, images, and videos to provide finer demand forecasting.

Zhao and Liang (2024) [26] illustrate that big data models outperform the traditional approaches to teaching by improving vocabulary memorization and sentence comprehension via cross-lingual embeddings.

Sharma and Gurung (2024) [27] contrast predictive maintenance with conventional methods, demonstrating data-driven techniques to be more precise and economical. In contrast to preventive maintenance, which adheres to fixed schedules, predictive maintenance employs real-time data to fine-tune timing and resource distribution.

Jin et al. (2024) [18] compare traditional manufacturing and big data-driven AM, stating that data analytics offers better accuracy, flexibility, and cost-effectiveness with adaptive optimization and automatic quality control.

Migenda et al. (2024) [4] show that adaptive local PCA is more effective compared to k-means and Gaussian Mixture Models for high-dimensional clustering. It is more computationally efficient as it skips full covariance matrix computation, thus is suitable for changing data distributions.

Gandaglia et al. (2024) [28] compare conventional clinical practices and big data analytics, illustrating that real-world data yields greater precision in patient characterization and outcome prediction. In contrast to conventional methodologies, big data analytics enables a broad range of comorbidity analysis and long-term outcomes in heterogeneous populations.

Beardslee et al. (2024) [2] illustrate that PCA outperforms traditional point selection by reducing time, removing

redundancy, and optimizing modal information extraction for accurate material property analysis.

Nyangon and Akintunde (2024) [29] demonstrate that PCA performs better than conventional approaches in noise filtering and identifying intricate patterns, thereby providing higher accuracy in volatile renewable energy markets.

Xie et al. (2024) [13] show that TSPCA outperforms classical sparse PCA in that it explains the same variance using fewer sensors, which enhances cost-effectiveness and computational efficiency.

Mukherjee et al. (2024) [30] illustrate how integration of multi-omics outperforms single-omics approaches by providing a comprehensive perspective on disease mechanisms and enhancing predictive performance via machine learning.

Ali et al. (2023) [15] state that AI-based IDS are more accurate, flexible, and capable of processing dynamic attack patterns than traditional models, providing improved security against sophisticated cyber threats.

3.4. Gaps and Future Directions

Li et al. (2024) [16] recognize shortcomings in PCA's non-linear data processing capability and sensitivity to outliers. Future directions include the creation of robust PCA variants and the incorporation of deep learning for improved pattern recognition. Quantum computing advancements and hybrid models may improve PCA's efficiency and scalability.

Ikegwu et al. (2024) [6] identify processing difficulties of unstructured data, computational costs, and adaptive, real-time models. Scalable, low-power algorithms, quantum computing, and hybrid models that combine machine learning, deep learning, and traditional approaches need to be taken into account in future research to enhance climate forecasting.

Wang et al. (2024) [20] identify potential for improvement of PCA in handling non-linear and unstable data patterns. Future research should integrate PCA with deep learning models, explore quantum computing, and develop adaptive hybrid models to enhance prediction precision.

Perez and Toraman (2024) [21] mention difficulty in dealing with non-linear data and high computational expenses. The future direction of research should be on combining PCA with machine learning for better pattern recognition and on using quantum computing for quicker processing.

Baidya et al. (2024) [22] recognize difficulty in processing non-linear data and high computational expense. Future studies need to combine PCA with sophisticated machine learning models and investigate quantum computing for quick data processing. Creating adaptive models for dynamic geochemical data would further improve the effectiveness of PCA.

Christensen et al. (2023) [23] acknowledge the limitations of modeling adsorbate interactions and non-linear mechanisms. Augment PCA with adaptive algorithms and quantum computing in future research.

Mehmood et al. (2024) [24] recognize obstacles to complete fusion of big data analytics and green innovation from both organizational and technological limitations. The future research direction should revolve around adaptive algorithms, improved data management processes, and innovative strategies for green innovation in a bid to achieve optimum economic and environmental returns.

Faaique (2024) [12] foresees difficulty in dealing with unstructured data and requires adaptive, scalable analytics models and quantum computing to process at higher speeds.

Wu et al. (2024) [25] acknowledge challenges in the integration of unstructured data and the need for dynamic and adaptive forecasting models. Future research must enhance real-time processing of data, develop scalable algorithms, and explore cloud computing for more efficient tourism demand forecasting.

Zhao and Liang (2024) [26] find difficulties in mapping cross-lingual embeddings to pedagogical models and call for adaptive learning approaches and more advanced unsupervised algorithms.

Sharma and Gurung (2024) [27] note issues in integrating data, expensive implementation, and requirements for highly technical expertise. The way forward for research lies in scalable, adaptive algorithms and better data management systems to make predictive accuracy cost-effective.

Jin et al. (2024) [18] identify challenges in big data, digital twin, and machine learning integration due to data complexity and high costs. Scalable algorithms, real-time processing, and adaptive hybrid models are what future research should focus on.

Migenda et al. (2024) [4] acknowledge problems with keeping all PCA units active during training. Future research is needed to improve adaptive learning, integrate PCA with neural networks, and enhance non-linear pattern processing and scalability.

Kovács and Haidu (2023) [1] identify problems with meteorological bias modeling and suggest the application of adaptive algorithms and real-time data processing for better prediction.

Gandaglia et al. (2024) [28] point out the difficulty of merging heterogeneous international datasets and the necessity for adaptive algorithms in order to improve predictive accuracy. Future research should be directed at the creation of data standardization, real-time analysis techniques, and individualized treatment plans for prostate cancer.

Beardslee et al. (2024) [2] refer to challenges in scaling PCA to complicated geometries and suggest blending adaptive algorithms and hybrid models with machine learning for improved precision.

Nyangon and Akintunde (2024) [29] identify issues in real-time processing and noise handling. The future research must integrate PCA with adaptive algorithms and machine learning for improved prediction.

Xie et al. (2024) [13] acknowledge challenges for using TSPCA on non-linear data and suggest integrating it with machine learning and exploring real-time manufacturing applications.

Mukherjee et al. (2024) [30] highlight integration challenges relating to data, computational complexity, and ethics. Future research is needed to enhance scalability, interpretability, and data privacy in multi-omics analysis.

Ali et al. (2023) [15]acknowledge data imbalance issues and suggest enhancing AI scalability and the integration of machine learning with deep learning to strengthen ICS cybersecurity.

4. DISCUSSION AND TABLE COMPARISON

Table 1: Summary About Literature Review on Details

Author(s)	Context of Evolution	Application Domain	Complexity Level	Remarks
Li et al. (2024)	PCA from statistical tool to complex ML and AI model	Dimensionality reduction, data interpretation	High	Handles high-dimensional data with scalability and noise reduction
Ikegwu et al. (2024)	Big data analytics evolution driven by climate data complexity	Climate change research, predictive modeling	Very High	Integrates satellite imagery and sensor networks for accurate climate predictions
Wang et al. (2024)	PCA integrated with ML and DL for complex data analysis	Wind power prediction	High	Enhances efficiency and accuracy in renewable energy systems
Perez and Toraman (2024)	PCA combined with gas chromatography for chemical analysis	Chemical reactions, pyrolysis studies	Medium	Improves accuracy in complex chemical decomposition analysis
Baidya et al. (2024)	PCA and multivariate analysis advanced for geochemical data	Mineral exploration, environmental studies	High	Enhances classification accuracy and predictive modeling
Christensen et al. (2023)	PCA integrated with Density Functional Theory (DFT)	Electrochemical CO2 reduction	High	Improves selectivity classification in complex chemical reactions
Mehmood et al. (2024)	Big data analytics for dynamic green innovation	Manufacturing SMEs	Medium	Supports sustainability and competitive advantage
Faaique (2024)	ML for astronomical data analysis due to massive datasets	Astronomy, celestial mapping	Very High	Advanced ML models handle complex multi-wavelength data

Wu et al. (2024)	Transition from traditional to ML/AI forecasting	Tourism and hospitality	Medium	Enhances demand forecasting accuracy using high-frequency web data
Zhao and Liang (2024)	Big data models for foreign language education	Language teaching, cross-lingual embeddings	Medium	Improves bilingual translation and learning outcomes
Sharma and Gurung (2024)	Evolution to ML-based predictive maintenance	Manufacturing, predictive maintenance	Medium	Reduces downtime and optimizes maintenance strategies
Jin et al. (2024)	Digital twins and ML for additive manufacturing	Additive manufacturing	High	Enhances process optimization, quality control, and predictive maintenance
Migenda et al. (2024)	Adaptive local PCA for high-dimensional clustering	Clustering in dynamic data	High	Uses variable learning rates and potential functions for non-stationary data
Kovács and Haidu (2023)	Advanced PCA models (WPCA, RPCA) for air pollution modeling	NO2 pollution modeling, satellite data	High	Accurate temporal-spatial NO2 predictions during COVID-19 lockdowns
Gandaglia et al. (2024)	Big data analytics for prostate cancer management	Prostate cancer, outcome prediction	Medium	Enhances patient characterization using international datasets
Beardslee et al. (2024)	PCA for optimizing RUS measurement in material analysis	Material property analysis	Medium	Reduces point selection and improves accuracy in complex geometries
Nyangon and Akintunde (2024)	PCA for electricity price forecasting	Renewable energy markets	High	Improves grid management and forecasting accuracy
Xie et al. (2024)	True Sparse PCA (TSPCA) for sensor optimization	Semiconductor manufacturing	High	Minimizes sensors while maintaining predictive accuracy

Mukherjee et al. (2024)	Multi-omics integration using big data and ML	Disease mechanisms, targeted therapies	Very High	Enhances predictive accuracy and therapeutic targeting
Ali et al. (2023)	AI-enhanced IDS for cybersecurity in ICS	Cybersecurity, ICS IDS	Very High	Achieves 99% accuracy in detecting complex cyber threats

The table 1: gives a detailed summary of the evolution, areas of application, levels of sophistication, and key contributions of Principal Component Analysis (PCA) and big data analysis in various fields, showing how it has progressed from a conventional statistical technique for dimensionality reduction to a sophisticated technique that is combined with machine learning (ML), deep learning (DL), and artificial intelligence (AI). It is a response to the growing demand for analyzing high-dimensional data, predictive accuracy, and operational efficiency. Works like Li et al. (2024)[16] showcase PCA's ability to scale and filter noise, and Ikegwu et al. (2024) [6] show how it is applied with satellite images for climate modeling. Specialized applications comprise chemical analysis (Perez and Toraman, 2024) [21] , electrochemical CO2 reduction (Christensen et al., 2023) [23] , and semiconductor manufacturing sensor optimization (Xie et al., 2024) [13] , all indicating PCA's adaptability to domain-specific challenges. Complexity varies between Medium and Very High with the greater complexity observed in dynamic high-dimensional data applications such as climate research (Ikegwu et al., 2024) [6] , astronomical data analysis (Faaique, 2024) [12] , multi-omics data integration (Mukherjee et al., 2024) [30] , and cybersecurity applications (Ali et al., 2023) [15] . Comparative studies highlight PCA's superiority in dimensionality reduction, pattern recognition, and predictive capabilities, particularly in hybrid models (Wang et al., 2024; Xie et al., 2024) [20][13] , and new combinations with scientific techniques (Perez and Toraman, 2024; et al., 2023) [21] [23]. However, there are issues with handling non-linear data, dynamic system precision, and computational efficiency, demanding hybrid models, adaptive algorithms, and real-time processing, especially in renewable energy markets (Nyangon and Akintunde, 2024) [29] and cybersecurity (Ali et al., 2023)[15] . Research in the future needs to focus on integrating PCA with ML, DL, and AI for improved scalability and interpretability, and expanding its applications in quantum computing, precision medicine, and smart manufacturing. In summary, PCA's history highlights its stability, scalability, and applicability in contemporary data analytics, the way forward to innovative solutions in new fields, with its application as a core dimensionality reduction and pattern identification tool becoming increasingly vital as data complexity increases.

5. APPLICATIONS OF PCA IN BIG DATA

PCA plays a key role in big data analytics in numerous industries for enhancing operational and strategic decision-making. PCA has been utilized to reduce intricate data structures, for example, by identifying the subspace where the majority of variance in big datasets resides, thus serving as a requisite tool in contemporary analytics [32]. Additionally, Principal Component Analysis (PCA) is used for electricity price forecasting in some of the California Independent System Operator (CAISO), enhancing accuracy through the management of heteroskedastic noise, thereby benefiting grid operations and integration of renewable sources [29]. Also, in the field of manufacturing, True Sparse PCA (TSPCA) has been employed to minimize the number of sensors required in virtual metrology, thereby reducing equipment expenses without sacrificing data analysis quality [13]. PCA plays a key role in big data analytics in numerous industries for enhancing operational and strategic decision-making. PCA has been utilized to reduce intricate data structures, for example, by identifying the subspace where the majority of variance in big datasets resides, thus serving as a requisite tool in contemporary analytics [6]. Additionally, Principal Component Analysis (PCA) is used for electricity price forecasting in some of the California Independent System Operator (CAISO), enhancing accuracy through the management of heteroskedastic noise, thereby benefiting grid operations and integration of renewable sources [31]. Also, in the field of manufacturing, True Sparse PCA (TSPCA) has been employed to minimize the number of sensors required in virtual metrology, thereby reducing equipment expenses without sacrificing data analysis quality [13].

6. EXTRACT STATISTICS

The figure 1 highlights effectively displays the prevalence of complexity levels in a dataset with categories well separated by color for readability. The most prevalent is high complexity with nine cases, suggesting a dataset with the majority of challenging tasks needing high skill or knowledge. Medium complexity has the second highest with seven cases, suggesting a high volume of moderately challenging content. The extremely high complexity category, although occurring only four times, represents the most challenging tasks, perhaps emphasizing specialized or unusual challenges within the research's subject area. This

kind of pattern demonstrates a skew toward high complexity levels, which may influence resource planning and strategic direction, such as giving priority to enhancement in areas demanding greater expertise or more rigorous training programs.

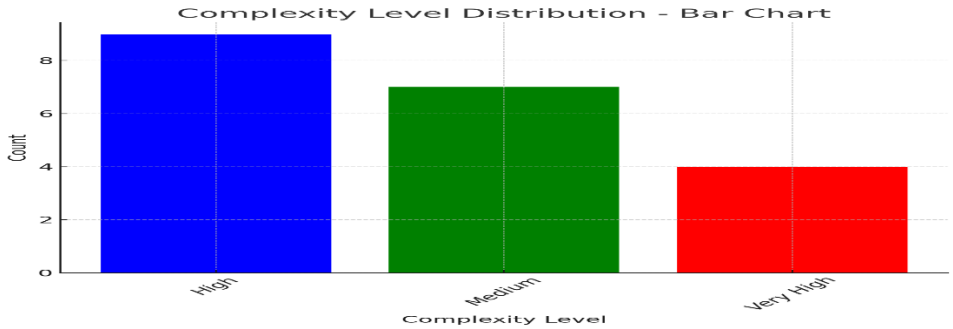

Figure 1: Complexity Level Frequency Distribution

The figure 2 below is a clear, colorful illustration of the spread of various research fields, with each field representing an equal proportion of 5% of the dataset. The even distribution indicates that every research field is represented singly, implying a wide range of both applied and theoretical fields. The distribution spans across many sectors, from energy, manufacturing, health, environmental studies, and cybersecurity, just to name a few. This depiction not only stresses the interdisciplinary character of contemporary studies but also brings to the fore the equitable representation of each discipline in this specific dataset. Such distribution could be suggestive of the presence of an equilibrated research agenda or a collection to encompass a diversified range of interests and issues intrinsic to contemporary science and technology. The provided chart elegantly portrays the wealth of research objects and their possible interconnections that can be instrumental in advancing interdisciplinary collaboration and novel ideas.

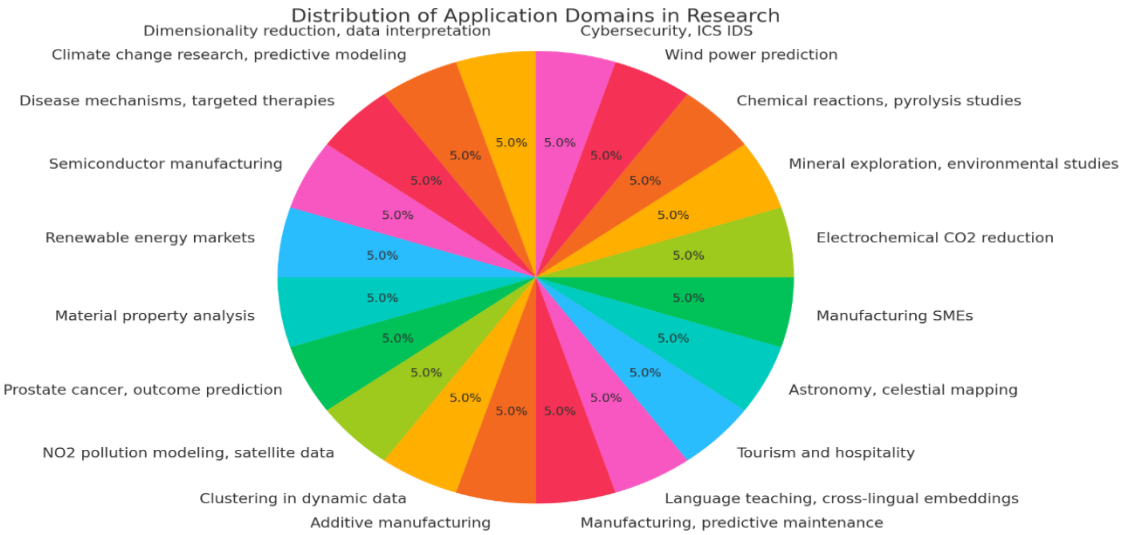

Figure 2: Distribution of Research Topics across Diverse Domains

The figure 3 shows how many words there are in comments on different research improvements. The x-axis shows the number of words per comment, and the y-axis shows how often these numbers occur. The data peaks at 7 words, which shows that most comments are short, tending to be about 7 words. There are a number of comments that contain 8 words. There are fewer comments that contain 5, 6, 9, and 10 words. This spread suggests a tendency for brief descriptions of research improvements. The preference for brevity could be the result of the need to convey the gist of the improvements with speed in situations like executive summaries, presentations, or reports where there is limited space and reader attention. The focus on 7-8 words shows a nice way of giving enough details without being too wordy. The mix of brief and longer comments shows different levels of detail, which may be in line with how complex or important the changes are.

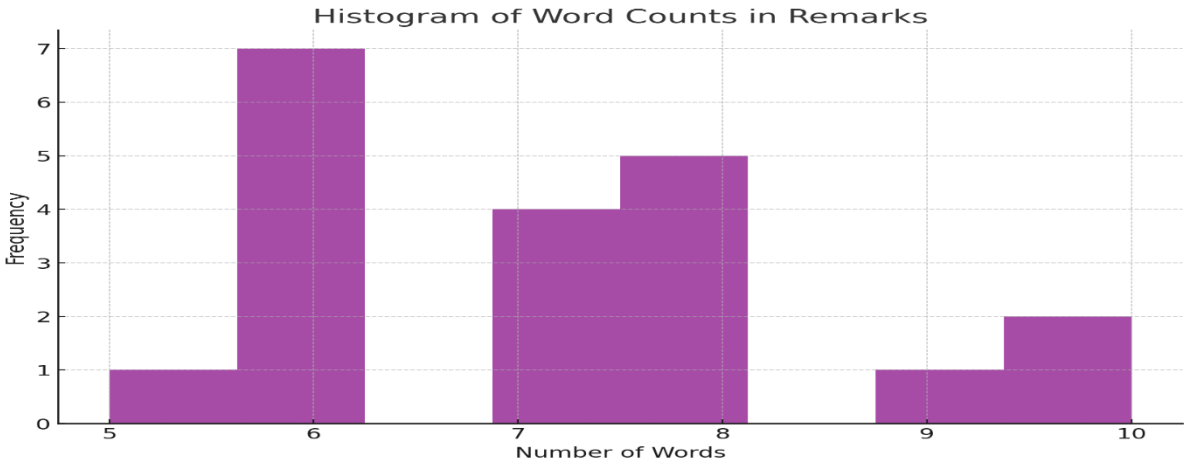

figure 3: Distribution in Research Improvement Remarks

7. RECOMMENDATIONS

1. Enhance Non-Linear Data Handling: Integrate Kernel PCA or deep learning to manage non-linear data more effectively.
2. Adaptive Learning Techniques: Use adaptive algorithms and neural networks for dynamic data environments.
3. Quantum Computing Integration: Explore quantum computing for improved scalability and computational efficiency.
4. Hybrid Models: Combine PCA with other dimensionality reduction methods (e.g., t-SNE, LDA) for better interpretability and accuracy.
5. Real-Time Data Processing: Develop real-time algorithms with big data frameworks like Apache Spark.
6. Emerging Technologies: Integrate PCA with AI and deep learning for advanced predictive analytics.
7. Enhanced Data Preprocessing: Implement rigorous data preprocessing to improve PCA outcomes.
8. New Application Domains: Expand PCA applications to emerging fields such as genomics, finance, and autonomous systems.

8. CONCLUSION

Principal Component Analysis (PCA) across diverse applications underscores its pivotal role in handling and interpreting complex datasets. By showcasing PCA's effectiveness in environments ranging from environmental monitoring to intricate geometrical data analyses and high-dimensional clustering, the study highlights the adaptability and precision of PCA in extracting meaningful patterns and reducing dimensionality. Furthermore, the integration of PCA with advanced classification techniques and its enhancement through neural network-based algorithms emphasize its potential to evolve alongside emerging data analysis technologies. As we continue to navigate the expanse of big data, PCA remains a fundamental tool, demonstrating not only the capacity to improve current methodologies but also to innovate new applications that respond to the evolving demands of data science. The ongoing refinement and adaptation of PCA methodologies will be crucial in harnessing the full potential of big data across various scientific and commercial fields, ensuring that data analysis remains both manageable and insightful.

REFERENCES

1. ", Ionel Haidu," no. 2, 2023.
2. L. Beardslee, P. Shokouhi, and T. J. Ulrich, "Optimal measurement point selection for resonant ultrasound spectroscopy of complex-shaped specimens using principal component analysis," NDT and E International, vol. 141, no. October 2023, p. 103000, 2024, doi: 10.1016/j.ndteint.2023.103000.
3. R. Hayati, A. A. Munawar, E. Lukitaningsih, N. Earlia, T. Karma, and R. Idroes, "Combination of PCA with LDA and SVM classifiers: A model for determining the geographical origin of coconut in the coastal plantation, Aceh Province, Indonesia," Case Studies in Chemical and Environmental Engineering, vol. 9, no. August 2023, p. 100552, 2024, doi: 10.1016/j.cscee.2023.100552.
4. N. Migenda, R. Möller, and W. Schenck, "Adaptive local Principal Component Analysis improves the clustering of high-dimensional data," Pattern Recognit, vol. 146, no. October 2023, p. 110030, 2024, doi: 10.1016/j.patcog.2023.110030.
5. F. R. Tato and I. M. Ibrahim, "Bio-Inspired Algorithms in Healthcare," vol. 07, no. 02, pp. 233–239, 2024.
6. A. C. Ikegwu, H. F. Nweke, E. Mkpojiogu, C. V. Anikwe, S. A. Igwe, and U. R. Alo, "Recently emerging trends in big data analytic methods for modeling and combating climate change effects," Energy Informatics, vol. 7, no. 1, 2024, doi: 10.1186/s42162-024-00307-5.

7. L. Jiasen, W. X. An, L. Guofeng, Y. Dan, and Z. Jindan, "Improved secure PCA and LDA algorithms for intelligent computing in IoT-to-cloud setting," Comput Intell, vol. 40, no. 1, 2024, doi: 10.1111/coin.12613.
8. A. Ciacci and L. Penco, "Business model innovation: harnessing big data analytics and digital transformation in hostile environments," Journal of Small Business and Enterprise Development, vol. 31, no. 8, pp. 22–46, 2023, doi: 10.1108/JSBED-10-2022-0424.
9. G. T. Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," IEEE Access, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.
10. T. Zhang, J. Wang, Q. Ma, and L. Fu, "Improving the Detection Effect of Long-Baseline Lightning Location Networks Using PCA and Waveform Cross-Correlation Methods," Remote Sens (Basel), vol. 16, no. 5, 2024, doi: 10.3390/rs16050885.
11. T. E. Beavers et al., "Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure," Journal of Computational and Graphical Statistics, pp. 1–31, 2024, doi: 10.1080/10618600.2024.2341896.
12. M. Faaique, "Overview of Big Data Analytics in Modern Astronomy," International Journal of Mathematics, Statistics, and Computer Science, vol. 2, pp. 96–113, 2023, doi: 10.59543/ijmscs.v2i.8561.
13. Y. Xie, T. Wang, Y. S. Jeong, A. Tosyali, and M. K. Jeong, "True sparse PCA for reducing the number of essential sensors in virtual metrology," Int J Prod Res, vol. 62, no. 6, pp. 2142–2157, 2024, doi: 10.1080/00207543.2023.2217282.
14. "Eido1822025AJRCOS129749 نسخة النشر.pdf."
15. B. S. Ali et al., ICS-IDS: application of big data analysis in AI-based intrusion detection systems to identify cyberattacks in ICS networks, vol. 80, no. 6. Springer US, 2024. doi: 10.1007/s11227-023-05764-5.
16. X. Li and Y. S. Lee, "Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm," Journal of Cases on Information Technology, vol. 26, no. 1, pp. 1–16, 2024, doi: 10.4018/JCIT.336916.
17. R. Avdal and H. Maseeh, "Advancing Cybersecurity through Machine Learning : Bridging Gaps , Overcoming Challenges , and Enhancing Protection," vol. 18, no. 2, pp. 206–217, 2025.
18. L. Jin et al., "Big data, machine learning, and digital twin assisted additive manufacturing: A review," Mater Des, vol. 244, no. June, p. 113086, 2024, doi: 10.1016/j.matdes.2024.113086.
19. P. Manoharan and K. D. Kokkotas, "Finding universal relations using statistical data analysis," Physical Review D, vol. 109, no. 10, pp. 1–20, 2024, doi: 10.1103/PhysRevD.109.103033.
20. Zhang. Wang, C., Zhang, Y., Ding, H., "Applied Mathematics and Nonlinear Sciences," Applied Mathematics and Nonlinear Sciences, vol. 8, no. 2, pp. 3383–3392, 2023.
21. Pereza, "Pr ep rin t n ot pe er r ev Pr ep ot pe er".
22. Baidya, "Pr ep rin t n ot pe er r iew Pr ep t n ot pe er ed".
23. O. Christensen, A. Bagger, and J. Rossmeisl, "The Missing Link for Electrochemical CO2 Reduction: Classification of CO vs HCOOH Selectivity via PCA, Reaction Pathways, and Coverage Analysis," ACS Catal, vol. 14, no. 4, pp. 2151–2161, 2024, doi: 10.1021/acscatal.3c04851.
24. K. Mehmood, F. Jabeen, M. Rashid, S. M. Alshibani, A. Lanteri, and G. Santoro, "Unraveling the transformation: the three-wave time-lagged study on big data analytics, green innovation and their impact on economic and environmental performance in manufacturing SMEs," European Journal of Innovation Management, 2024, doi: 10.1108/EJIM-10-2023-0903.
25. D. C. Wu, S. Zhong, J. Wu, and H. Song, "Tourism and Hospitality Forecasting With Big Data: A Systematic Review of the Literature," Journal of Hospitality and Tourism Research, 2024, doi: 10.1177/10963480231223151.
26. Y. Zhao and G. Liang, "A study on the innovative model of foreign language teaching in universities using big data corpus," Journal of Computational Methods in Sciences and Engineering, vol. 24, no. 1, pp. 87–103, 2024, doi: 10.3233/JCM-237113.
27. R. Sharma and S. Gurung, "Implementing Big Data Analytics and Machine Learning for Predictive Maintenance in Manufacturing Facilities in South Korea," AI, IoT and the Fourth Industrial Revolution …, vol. 14, no. 1, 2024, [Online]. Available: https://scicadence.com/index.php/AI-IoT-REVIEW/article/view/36%0Ahttps://scicadence.com/index.php/AI-IoT-REVIEW/article/download/36/34
28. G. Gandaglia et al., "Clinical Characterization of Patients Diagnosed with Prostate Cancer and Undergoing Conservative Management: A PIONEER Analysis Based on Big Data," Eur Urol, vol. 85, no. 5, pp. 457–465, 2024, doi: 10.1016/j.eururo.2023.06.012.
29. J. Nyangon and R. Akintunde, "Principal component analysis of day-ahead electricity price forecasting in

CAISO and its implications for highly integrated renewable energy markets," Wiley Interdiscip Rev Energy Environ, vol. 13, no. 1, 2024, doi: 10.1002/wene.504.

30. A. Mukherjee, S. Abraham, A. Singh, S. Balaji, and K. S. Mukunthan, "From Data to Cure: A Comprehensive Exploration of Multi-omics Data Analysis for Targeted Therapies," Mol Biotechnol, no. 0123456789, 2024, doi: 10.1007/s12033-024-01133-6.
31. J. Li, M. S. Othman, H. Chen, and L. M. Yusuf, "Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning," J Big Data, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00892-y.
32. Z. Wang et al., "Ultra-Short-Term Offshore Wind Power Prediction Based on PCA-SSA-VMD and BiLSTM," Sensors, vol. 24, no. 2, 2024, doi: 10.3390/s24020444.