American Journal of Electrical and Computer Engineering
2023; 7(2): 27-39
http://www.sciencepublishinggroup.com/j/ajece
doi: 10.11648/j.ajece.20230702.12
ISSN: 2640-0480 (Print); ISSN: 2640-0502 (Online)

Comparative Analysis of Feature Extraction of High Dimensional Data Reduction Using Machine Learning Techniques

Seth Gyamerah[1], Godfred Tour Soori[1], Dennis Redeemer Korda[2, *], John Kwame Tawiah[3], Eric Ayintareba Akolgo[4], Emmanuel Oteng Dapaah[5]

[1]Department of Computer Science, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana

[2]Department of Information and Communication Technology, Bolgatanga Technical University, Bolgatanga, Ghana

[3]Department of Civil Engineering, Ho Technical University, Ho, Ghana

[4]Department of Computer Science, Regentropfen College of Applied Sciences, Bolgatanga, Ghana

[5]Department of Information and Communication Technology, E.P College of Education, Bimbila, Ghana

Email address:

sgyamerah@cktutas.edu.gh (Seth Gyamerah), godfredtour@gmail.com (Godfred Tour Soori), dkorda@bolgatu.edu.gh (Dennis Redeemer Korda), jtawiah@htu.edu.gh (John Kwame Tawiah), eric.akolgo@recas.edu.gh (Eric Ayintareba Akolgo), eodapaah@epcoebimbilla.edu.gh (Emmanuel Oteng Dapaah)

[*]Corresponding author

To cite this article:

Seth Gyamerah, Godfred Tour Soori, Dennis Redeemer Korda, John Kwame Tawiah, Eric Ayintareba Akolgo et al. (2023). Comparative Analysis of Feature Extraction of High Dimensional Data Reduction Using Machine Learning Techniques. *American Journal of Electrical and Computer Engineering*, *7*(2), 27-39. https://doi.org/10.11648/j.ajece.20230702.12

Received: October 29, 2023; **Accepted**: November 17, 2023; **Published**: December 11, 2023

Abstract: Dimensionality reduction is critical for analyzing and interpreting high-dimensional data across domains like genomics, imaging, and finance. This paper presents a comparative analysis of dimensionality reduction techniques, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Recursive Feature Elimination (RFE), and Lasso regression. These methods are applied to datasets from genomics, medical imaging, and finance to evaluate their ability to reduce dimensions while preserving relevant information. The results demonstrate that PCA and LDA are highly effective for genomics data, reducing gene expression profiles from over 60,000 dimensions to 10-50 components while maintaining precision of over 80%. For medical images, PCA and LDA reduce pixel dimensions by over 90% without compromising precision. However, no single technique optimizes dimensionality reduction and precision for complex finance data. Overall, the analysis provides domain-specific insights, highlighting PCA and LDA as leading techniques for genomics and imaging. The choice of method should be guided by data characteristics. Testing on more diverse, real-world datasets is needed to establish validity further. This research aims to inform the selection of appropriate data reduction techniques across critical applications involving high-dimensional data.

Keywords: Machine Learning, Principal Component Analysis, Linear Discriminant Analysis, Recursive Feature Elimination, Lasso Regression, Genomics, Medical Imaging

1. Introduction

There has been a growing trend of collecting and storing large amounts of data in recent years. This trend is driven by several factors, including the increasing availability of sensors and data collection devices, the decreasing cost of storage, and the rise of big data analytics.

One of the challenges associated with working with large datasets is that they can be very high dimensional. This means that they have a large number of features or variables. High dimensionality can make it difficult to analyze data, leading to problems such as the curse of dimensionality. The curse of dimensionality refers to the fact that as the number of features in a dataset increases, the distance between any two points in

the dataset also increases. This can make it difficult to find patterns in the data and make it challenging to train machine learning models.

One approach to addressing these challenges is data reduction, which involves reducing the number of features while preserving the relevant information. This can lead to better understanding and visualization of the data and improved performance in machine learning tasks.

In this paper, several data reduction methods for high-dimensional data with be compared, including traditional techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), as well as machine learning-based feature engineering methods.

These include:

1. Recursive Feature Elimination
2. Lasso Regression

These methods involve training machine learning models to identify the most relevant features for a given task. Random Forest Feature Importance consists of training a Random Forest model and then computing the feature importance scores based on the decrease in impurity (Gini index) caused by each feature. Recursive Feature Elimination involves recursively removing features from the data set and evaluating the performance of a machine learning model on the reduced data set. The feature with the lowest importance score is removed in each iteration until the desired number of features is reached. Lasso Regression is a linear regression method that performs feature selection by adding an L1 penalty term to the objective function, encouraging the model to select features with non-zero coefficients and resulting in sparse solutions.

High-dimensional data poses several challenges for analysis, including computational complexity, sparsity, and overfitting. Data reduction methods can alleviate some of these challenges by reducing the number of features while preserving the relevant information. However, choosing an appropriate data reduction method for a given task can be difficult, and different ways may perform better on different data types.

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are widely used data reduction methods, but they may only be suitable for some data types. Machine learning-based feature engineering methods, such as Recursive Feature Elimination and Lasso Regression, have been proposed as alternatives that may perform better on some kinds of data. However, there currently needs to be a consensus on which method is best for a given task.

Therefore, this paper addresses the problem of comparing the performance of different data reduction methods for high-dimensional data and identifying their strengths and weaknesses. Specifically, we will compare the performance of PCA, LDA, Recursive Feature Elimination, and Lasso Regression on different types of high-dimensional data, including genomics, imaging, social networks, and finance. We will evaluate the performance of these methods based on their ability to preserve relevant information, reduce computational complexity, and improve the performance of machine learning models. By doing so, we aim to provide insights into the strengths and limitations of different data reduction methods and help researchers and practitioners choose the most appropriate method for their specific task.

2. Literature Review

Dimensionality reduction techniques play a crucial role in many applications ranging from machine learning and pattern recognition to data analysis and visualization. These techniques are employed as a pre-processing step to remove irrelevant and redundant data, leading to enhanced learning accuracy and improved result comprehensibility. With the increasing dimensionality of data in recent times, however, existing feature selection and feature extraction methods face significant challenges in terms of efficiency and effectiveness. Several studies have been conducted to compare the performance of different dimensionality reduction techniques.

One of the most comprehensive studies was conducted by S. Vijayarani et al. [1]. In this study, the authors compared and analyzed different dimensionality reduction techniques. The objective of the paper was to provide a systematic comparative analysis of feature reduction algorithms, namely PCA, LDA, and FA, applied to medical datasets (Gene annotations). The performance measures considered were the number of attributes reduced and the time taken for reduction.

PCA, or Principal Component Analysis, was one of the algorithms used for feature reduction. It applies an orthogonal transformation to convert possibly correlated variables into linearly uncorrelated variables called principal components. The PCA algorithm involved subtracting the mean from the original data, calculating the covariance matrix, finding eigenvalues and eigenvectors, and selecting the principal component with the largest eigenvalue. The advantages of PCA include uncorrelated principal components and the ability to capture the most significant percentage of variation in the dataset.

Another algorithm utilized in this study was LDA or Linear Discriminant Analysis. LDA is a generalization of Fisher's linear discriminant and is used to find a linear combination of features that characterizes or separates different classes of objects or events. Its goal is to project the dataset into a lower-dimensional space with good class separability to avoid overfitting and reduce computational costs. LDA offers the advantage of reducing the error rate and providing interpretable results between data groups.

The authors stated that PCA outperformed the other algorithms in terms of efficiency. However, while the study provides insights into the comparative analysis of feature extraction algorithms, there are certain limitations that need to be addressed.

One of the notable shortcomings of this research paper is the lack of a comprehensive evaluation of the quality of the reduced features. While the number of features reduced and the time taken are necessary performance measures, it is equally crucial to assess the impact of dimensionality reduction on the predictive accuracy or classification performance of subsequent data mining techniques. Without

considering this aspect, it is difficult to ascertain the practical usefulness of the proposed algorithms.

Additionally, it would have been beneficial if the paper had discussed the limitations of each algorithm. Understanding the potential drawbacks and assumptions of PCA, LDA, and FA would have provided a more nuanced perspective on their applicability to different types of datasets and mining tasks.

In conclusion, there are opportunities for enhancing the research by incorporating more advanced feature reduction techniques, discussing the limitations of the algorithms, and evaluating a broader range of metrics. Future studies could explore diverse datasets and employ comprehensive evaluation measures to gain a deeper understanding of dimensionality reduction methods. This would contribute to the advancement of the field and improve the effectiveness of dimensionality reduction in various applications.

Another study that contributed to the field dimensionality reduction is K. Yildiz et al. [2]. The research paper presents an approach to address the challenge of clustering high-dimensional data by combining different dimensionality reduction techniques with the Fuzzy C-Means (FCM) clustering algorithm. The authors argue that traditional clustering algorithms suffer from the curse of dimensionality, and therefore, dimensionality reduction techniques are necessary to improve clustering accuracy and efficiency in high-dimensional spaces.

The authors introduced seven dimensionality reduction techniques, including Principal Component Analysis (PCA), Laplacian, Fast Maximum Variance Unfolding (MVU), Isometric Mapping, Landmark Isometric Mapping, Stochastic Neighbor Embedding (SNE), and t-distributed Stochastic Embedding (t-SNE). Experiments were conducted using three real-world datasets: Abalone, Milliyet, and BBC, and the results are presented in tables and figures.

The conclusion drawn from the experimental results suggests that Laplacian, FastMVU, and t-SNE are the most efficient dimensionality reduction algorithms for the considered datasets. It is also observed that when dimensionality reduction is applied, the cluster purity and mutual information of the datasets increase.

However, there are several areas that could be improved upon. Firstly, the paper lacks a comprehensive discussion on the limitations and drawbacks of the utilized dimensionality reduction techniques. A thorough analysis of the trade-offs, such as loss of information or sensitivity to parameter settings, would enhance the validity and applicability of the findings.

Also, the evaluation metrics used in the study, such as cluster purity, entropy, and mutual information, are somewhat limited in capturing the full complexity of clustering performance. It would be beneficial to include additional evaluation measures, such as the silhouette coefficient or Rand index, to provide a more comprehensive assessment of the clustering quality.

In terms of future work, the authors suggest using a genetic algorithm for selecting the best subspace to represent high-dimensional data. However, they do not provide a detailed explanation or justification for this approach. Expanding on the proposed future direction and providing a theoretical basis for the use of genetic algorithms would strengthen the paper's contribution.

Another study that investigated the performance of dimensionality reduction techniques in machine learning algorithms was conducted by G. T. Reddy et al. [3]. The authors explored the effectiveness of two prominent dimensionality reduction techniques, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), using four popular machine learning algorithms: Decision Tree Induction, Support Vector Machine (SVM), Naive Bayes Classifier, and Random Forest Classifier. They employed the Cardiotocography (CTG) dataset from the University of California and Irvine Machine Learning Repository for their experimentation.

The results of their study indicated that PCA outperformed LDA in all measures examined. Additionally, they observed that the performance of the Decision Tree and Random Forest classifiers was not significantly affected when using PCA or LDA. To further analyze the impact of PCA and LDA, the researchers conducted experiments on Diabetic Retinopathy (DR) and Intrusion Detection System (IDS) datasets. Their findings demonstrated that machine learning algorithms with PCA achieved better results when the dimensionality of the datasets was high. Conversely, when the dimensionality of the datasets was low, the researchers observed that machine learning algorithms without dimensionality reduction yielded superior outcomes.

While the study by G. T. Reddy et al. [3] sheds light on the performance of dimensionality reduction techniques in machine learning algorithms, there are aspects that warrant critique and further exploration. The inclusion of a broader range of techniques, algorithms, and datasets, as well as an analysis of interpretability and a deeper understanding of the observed trends, would enhance the comprehensiveness and robustness of the findings.

A thesis was presented by H. Yang [4] who conducted a study of dimensionality Reduction Techniques that Enhance Trace Clustering Performances. The paper focuses on improving process mining techniques by applying dimensionality reduction to trace clustering. The author acknowledges that traditional process mining techniques face challenges in analyzing real-life process logs due to their complexity and lack of structure. She proposes using dimensionality reduction techniques such as singular value decomposition (SVD), random projection, and principal components analysis (PCA) to reduce the number of features and enhance the efficiency of trace clustering. While the approach presented in the paper offers potential benefits, there are several areas that could be further improved. Firstly, the paper does not provide a comprehensive comparison of dimensionality reduction techniques and clustering algorithms.

Moreover, the evaluation of the proposed approach is limited to a case study involving patient treatment processes in a hospital. While this study provides some insights, it would be valuable to extend the research to other industries and

diverse process logs. Different industries may have unique characteristics and challenges that require tailored approaches to process mining. Conducting similar studies with process logs from various domains would strengthen the validity and generalizability of the findings.

Overall, there are areas that require improvement to enhance the applicability and reliability of the proposed approach. The limited comparison of dimensionality reduction techniques and clustering algorithms, the reliance on a single case study, and the need for more comprehensive evaluation metrics suggest opportunities for further research. By addressing these limitations, future studies can provide more robust insights and guidelines for practitioners in the field of process mining.

The research paper titled "Investigating the Effect of Dimensionality Reduction Techniques on Machine Learning Algorithms" by T. Gadekallu [5] explores the impact of two prominent dimensionality reduction techniques, LDA and PCA, on the performance of four popular machine learning algorithms: Decision Tree Induction, SVM, Naive Bayes Classifier, and Random Forest Classifier. The experimentation is conducted on publicly available datasets, including the CTG dataset from the University of California and Irvine Machine Learning Repository, as well as the DR and IDS datasets. The authors involved feature engineering, normalization, application of ML algorithms in their methods, and the use of LDA and PCA for dimensionality reduction. They presented the performance evaluation of the classifiers with and without dimensionality reduction, using metrics such as accuracy, sensitivity, and specificity.

However, the research paper falls short in several aspects. Firstly, the limited focus on a small number of machine learning algorithms restricts the generalizability of the findings. The authors should consider incorporating a broader range of algorithms to provide a more comprehensive analysis. Additionally, the study primarily investigates relatively small datasets, limiting the applicability of the findings to larger and more complex datasets. Future research should address this limitation by examining the effects of dimensionality reduction on datasets with higher dimensionality and diverse data types. Lastly, while the paper compares LDA and PCA, it lacks a thorough analysis of the underlying reasons for the observed differences in performance. Further investigation into the characteristics of the datasets and the assumptions made by each technique would provide valuable insights.

A study conducted by L. Zhang et al. [6] compares the performance of four different dimensionality reduction techniques for cancer diagnosis: PCA, LDA, RFE, and most minor absolute shrinkage and selection operator (LASSO). The authors used a dataset of patients with cancer, and they evaluated the performance of the different dimensionality reduction techniques utilizing the area under the receiver operating characteristic curve (AUC). The study showed that RFE and LASSO outperformed PCA and LDA regarding AUC. This suggests that RFE and LASSO are more effective than PCA and LDA for identifying the essential features for cancer diagnosis.

Authors of the research paper "Comparative Study of Dimensionality Reduction Techniques for Intrusion Detection System" by S. Bharti et al. [7] compared the performance of PCA, LDA, and LASSO on three network intrusion datasets. The datasets were the KDD Cup 1999 dataset, the NSL-KDD dataset, and the CICIDS2017 dataset. The authors used a SVM classifier to evaluate the performance of the dimensionality reduction techniques.

The authors acknowledge that there are some limitations to their study. First, they only evaluated the performance of LASSO on three network intrusion datasets. LASSO may not perform as well on other datasets. Second, the authors only evaluated the performance of LASSO on a support vector machine classifier. It is possible that LASSO may not perform as well as with other classifiers. Despite these limitations, the paper's findings suggest that LASSO is a promising dimensionality reduction technique for intrusion detection systems.

Another research that contributed to dimensionality reduction is the "Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data" by S. Ayesha et al. [8]. The authors explore the challenges and benefits of dimensionality reduction techniques (DRTs) in analyzing high-dimensional data. While DRTs can enhance processing speed and extract valuable information, several issues and limitations are associated with their application.

One primary concern raised in the paper is the difficulty in selecting an appropriate DRT according to the type of data. This issue is crucial as different datasets may require specific techniques for effective dimensionality reduction. The paper acknowledges the need for a suitable mechanism to combine several DRTs' outputs accurately. Another limitation discussed in the paper is the identification of redundancy levels in high-dimensional data.

In the field of machine learning and pattern recognition, dimensionality reduction has emerged as an important area of research, with numerous approaches proposed to address this challenge. The author in [9] compared and analyzed dimensionality reduction techniques for machine learning. The primary objective of the paper was to compare and evaluate various schemes used to reduce the dimensionality of high-dimensional datasets, aiming to improve the accuracy and time complexity of machine learning algorithms, particularly in classification and clustering tasks.

The Iris dataset, introduced by R. A. Fisher [10], comprises samples from three different species of Iris flowers. Fisher developed a linear discriminant model based on four features (length and width of sepals and petals) to distinguish the species from each other. On the other hand, the Wines dataset is used for comparing various classifiers, with the classes being separable. The paper reports the classification accuracy achieved by different classifiers, such as RDA (100%), QDA (99.4%), LDA (98.9%), and 1NN (96.1%).

The research paper titled "Evaluation of Dimensionality Reduction Techniques: A Comparative Study" by M. Vikram et al. [11] provides a systematic evaluation of popular

dimensionality reduction techniques, namely PCA, ICA, SVD, and NMF, based on their efficiency and effectiveness. The authors aim to assist data science practitioners in selecting the most suitable technique by considering the trade-off between effectiveness and efficiency. The research paper employed a methodology that involved computing the parameters of each dimensionality reduction technique and measuring efficiency through fit time and transform time. Effectiveness was evaluated using metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) to assess the similarity between the original and reconstructed data.

The authors in [12] proposed an efficient approach for indexing images by content using a combination of principal component analysis (PCA), locality-sensitive hashing (LSH), and the vector approximation file (VA-File) method. The authors aim to address the challenge of the "curse of dimensionality" caused by the increasing volume of data in multimedia processing systems. While the proposed method demonstrates improvements in search speed and memory storage, there are several areas where it could be further enhanced. The proposed approach consists of three phases. Firstly, feature extraction is performed using SIFT and SURF algorithms. Next, PCA and LSH are applied for dimensionality reduction. Finally, the VA-File method is used to accelerate the search phase. The combination of PCA, LSH, and VA-File is compared with other combinations to select the best-performing approach.

Research was also conducted by D. Mishra and S. Sharma [13] on dimensionality reduction techniques. The research paper titled "A Comparative Study of Dimensionality Reduction Techniques" focuses on analyzing and comparing various techniques of dimensionality reduction. The authors discuss the concept, techniques, and applications of dimensionality reduction and aim to provide insights into the effectiveness of different approaches. The results of implementing Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA) on the iris dataset and wine dataset are presented.

3. Methodology

3.1. Research Design

In this section, a comparative analysis of data reduction techniques will be conducted to evaluate their performance on different types of high-dimensional data. The research design will be experimental in nature, involving the application of various data reduction methods to distinct datasets representing genomics, imaging, and financial domains. By using a practical approach, the research aims to investigate the strengths and weaknesses of each method in reducing the dimensionality of these diverse datasets and its impact on subsequent machine learning tasks.

Datasets from the genomics, imaging, and finance domains will be collected. These datasets will be publicly available and sourced from reputable platforms, such as Kaggle. The datasets will represent high-dimensional data, including DNA sequences or gene expression data for genomics, MRI or CT scans for imaging, and stock market data for finance.

Before applying the data reduction methods, the collected datasets will undergo necessary preprocessing steps to ensure data quality and consistency. These preprocessing steps will involve data cleaning, feature extraction, and normalization, where applicable, to address any missing values or inconsistencies in the data.

Four data reduction methods will be applied to the preprocessed datasets, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Recursive Feature Elimination (RFE), and Lasso Regression. Each method will be executed separately on the genomics, imaging, and finance datasets to capture their respective performances.

To evaluate the effectiveness of the data reduction methods, several performance metrics will be used. These metrics will include data preservation, computation time, and the impact on machine learning model performance. The preservation of relevant information will be measured using metrics such as explained variance for PCA, classification accuracy for LDA, and feature selection score for RFE and Lasso Regression.

Statistical analysis will be performed on the results obtained from each data reduction technique for each dataset. Statistical tests and effect size calculations will be used to compare the methods' performances, considering different data types and their specific characteristics.

Visualization techniques, such as scatter plots, heatmaps, and bar charts, will be employed to illustrate the outcomes of the data reduction methods and to facilitate a comprehensive understanding of the results.

3.2. Data Collection

The datasets used in this research were obtained from Kaggle, a popular platform for accessing and sharing data. Kaggle hosts a wide variety of datasets, including those from the genomics, imaging, and finance domains, which are essential for this study's comparative analysis.

For the genomics field, a dataset containing DNA sequences or gene expression data will be selected. The chosen dataset has many features, representing various genetic markers or gene expression levels across different samples. These datasets are often used in genomic research and hold valuable information for understanding the genetic basis of multiple traits and diseases.

The imaging dataset consists of MRI or CT scans, which are common types of medical imaging data. These datasets selected are based on their high dimensionality, where each image represents a pixel-wise representation of the scanned body part. Imaging datasets play an important role in medical diagnostics and research, and reducing their dimensionality can aid in efficient feature extraction and analysis.

In the field of finance, stock market data, such as historical stock prices and various financial indicators, are chosen [14]. The financial dataset is highly dimensional, containing multiple time series of stock market information for different assets. These datasets are valuable in financial forecasting and

analysis, and dimensionality reduction can improve computational efficiency and enhance predictive modeling.

3.3. Data Reduction Techniques

This section gives detailed explanations of the data reduction methods that will be compared in this study, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Recursive Feature Elimination (RFE), and Lasso Regression. The algorithmic steps of each method and how they are applied to high-dimensional data will be discussed as well.

3.3.1. Principal Component Analysis (PCA)

PCA is a widely used linear transformation technique for dimensionality reduction. It aims to find the principal components that capture the most variance in the data while reducing the dimensionality [14]. It was initially invented by Karl Pearson and was later developed by Harold Hotelling [15].

The primary objective of PCA is to transform the original high-dimensional feature space into a new, lower-dimensional space while retaining as much of the variance as possible. The lower-dimensional space is defined by a set of orthogonal axes called principal components, which are linear combinations of the original features. The algorithmic Steps of PCA is as follows:

The data matrix X is centered by subtracting the mean from each feature to obtain X.

1. Calculate the covariance matrix Cov(X) using the formula above.
2. Compute the eigenvectors and eigenvalues of Cov(X).
3. Sort the eigenvectors in descending order based on their corresponding eigenvalues.
4. Select the top k eigenvectors to form the projection matrix P, where k is the desired reduced dimensionality.
5. Project the original data onto the new reduced-dimensional space.

Mathematically, given a dataset with N samples and D features represented as an N x D matrix X, PCA aims to find the matrix of transformed data Z, where the transformed data has M (M < D) dimensions:

$$X = [x\_1, x\_2, \ldots, x\_D] \tag{1}$$

$$Z = z\_1, z\_2, \ldots, z\_M \tag{2}$$

The transformed data Z is obtained by projecting the original data X onto the principal components, as follows:

$$Z = X * W \tag{3}$$

Where W is a D x M matrix containing the eigenvectors corresponding to the top M eigenvalues of the covariance matrix of X. The mathematical formulation of PCA involves computing the covariance matrix of the original data X, and then performing an eigen decomposition to find the eigenvectors (principal components). Let X be an n x d matrix representing the high-dimensional data, where n is the number of samples, and d is the number of features. The covariance matrix of X can be calculated as follows:

$$Cov(x) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X}) \tag{4}$$

Where X_i is the row of X and $\bar{x}$ is the mean of X.

The principal components can be obtained by performing the eigen decomposition of the covariance matrix:

$$Cov(X) = V \Lambda V^{\mathrm{T}} \tag{5}$$

Where V is a matrix of eigenvectors, and Λ is a diagonal matrix of eigenvalues. The principal components are then the columns of V, sorted in descending order based on their corresponding eigenvalues. PCA can be applied to reduce the dimensionality of the data by projecting it onto the first k principal components, where k < d. The transformed data is obtained as follows:

$$X_{PCA} = XV_k \tag{6}$$

Where V_k is the matrix containing the first k principal components.

3.3.2. Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction technique that aims to find a subspace that maximizes class separability while reducing dimensionality. It is commonly used for classification tasks, particularly in cases where the classes are well-separated. The primary goal of Linear Discriminant Analysis (LDA) is to transform a high-dimensional dataset into a lower-dimensional space that maintains good class separation. By doing so, it helps reduce computational complexity and processing time. LDA's approach shares similarities with Principal Component Analysis (PCA) in that both methods aim to maximize certain properties of the data [16]. While PCA focuses on maximizing data variance, LDA goes a step further by also emphasizing the separation between different classes in the dataset. The algorithmic Steps of LDA is as follows:

1. Compute the class-wise mean vectors μ_i and the overall mean vector μ.
2. Calculate the within-class scatter matrix S_w and the between-class scatter matrix using the equations below.
3. Compute the eigenvalues and eigenvectors of the matrix $S_W^{-1} S_B$
4. Sort the eigenvectors in descending order based on their corresponding eigenvalues.
5. Select the top k eigenvectors to form the projection matrix W, where k is the desired reduced dimensionality.
6. Project the original data onto the new reduced-dimensional space.

The mathematical formulation of LDA involves computing the between-class scatter matrix S_b and the within-class scatter matrix S_w. The objective is to find a projection matrix W that maximizes the ratio of the determinant of S_b to that of S_w. Let X be the same n x d matrix representing the high-dimensional data, and y be a vector of length n containing the class labels for each sample.

The between-class scatter matrix S_B can be defined as:

$$S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T \quad (7)$$

Where C is the number of classes, n_i is the number of samples in class i, μ_i is the mean of class i, and μ is the overall mean of the data.

The within-class scatter S_w can also be defined as:

$$S_W = \sum_{i=1}^{C} \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T \quad (8)$$

The optimization problem in LDA can be solved by finding the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1} S_B$.

LDA can be applied to reduce the dimensionality of the data by projecting it onto the first k eigenvectors, where k < d. The transformed data is obtained as follows:

$$X_{LDA} = XW_k \quad (9)$$

Where W_k is the matrix containing the first k eigenvectors.

3.3.3. Recursive Feature Elimination (RFE)

RFE is a feature selection technique that recursively removes features from the dataset based on their importance score. It involves training a machine learning model and then iteratively eliminating the least important features until the desired number of features is reached.

Let X be the n x d matrix representing the high-dimensional data, and y be the target variable. Then RFE follows the steps below:

1. Train a machine learning model M on the original dataset.
2. Obtain the feature importance scores from the model M.
3. Remove the feature with the lowest importance score.
4. Repeat steps 1 to 3 until the desired number of features k is reached.
5. The transformed data after RFE is the subset of the original features obtained in step 4.

3.3.4. Lasso Regression

Lasso Regression is a linear regression method that performs both feature selection and regularization by adding an L1 penalty term to the objective function. The L1 penalty encourages the model to set some feature coefficients to exactly zero, effectively performing feature selection.

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \alpha \|\beta\|_1 \right\} \quad (10)$$

where Y is the target variable, X is the data matrix, β is the coefficient vector, and α is the regularization parameter controlling the strength of the penalty term. The algorithmic Steps of Lasso Regression is as follows:

1. Normalize the data matrix X and standardize the target vector y.
2. Initialize the coefficient vector w with small random values.
3. Update the coefficients using coordinate descent or gradient descent while applying the L1 penalty.
4. Continue updating the coefficients until convergence or a predefined number of iterations.

By applying these mathematical formulations, a thorough evaluation and comparison of the data reduction methods on the genomics, imaging, and finance datasets can be performed to achieve our research objectives.

3.4. Evaluation Procedure

Each data reduction method will be applied to the respective datasets representing genomics, imaging, and finance. The parameters for each method, such as the number of principal components for PCA, the number of selected features for RFE and Lasso Regression, and the projection matrix for LDA, will be determined based on the specific characteristics of the datasets.

After applying each data reduction technique, the reduced datasets will be visualized in lower-dimensional spaces. This visualization will allow us to observe the distribution of samples and potential clusters within the data. Scatter plots, heat maps, and other visualization techniques will be used to aid in the interpretation of the results.

The defined performance metrics, including data preservation, and computation time, will be calculated for each data reduction method on each dataset. The results will be summarized and compared to identify the strengths and weaknesses of each technique for different types of high-dimensional data.

Cross-domain comparisons will be conducted to assess the generalization capability of each data reduction method across the genomics, imaging, and finance datasets. This will help identify any data type-specific advantages or limitations of the techniques.

The results of the evaluation will be presented using visualization techniques. Bar charts, line plots, and other visualizations will be used to illustrate the performance of each method in a clear and interpretable manner.

Visualization will be the primary form of statistical analysis for comparing the data reduction methods. Visualizing the reduced datasets in lower-dimensional spaces will allow us to gain insights into how each technique transforms the original high-dimensional data.

Scatter plots will be used to visualize the distribution of samples after applying the data reduction methods. By plotting samples in the reduced space, we can observe potential clusters or patterns that may emerge as a result of dimensionality reduction. Heatmaps will be employed to visualize the covariance matrix or correlation matrix of the reduced data. This will provide insights into the relationships between features and help us understand the level of information preservation achieved by each method.

Bar charts will be used to compare the performance metrics of each data reduction technique on different datasets. This will allow for easy comparison of how well each method performs in terms of data preservation, computation time, and its impact on machine learning model performance. Line plots will be utilized to show the cumulative explained variance by the principal components in PCA. This will aid in determining the optimal number of components required to achieve a desired level of data preservation.

4. Results and Discussion

4.1. Overview

In this session, the results of applying different dimensionality reduction techniques on three datasets - genomics, medical imaging, and finance - are presented and analyzed. The goal is to evaluate and compare the effectiveness of Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Lasso regression, and Linear Discriminant Analysis (LDA) for reducing high-dimensional data from these domains while retaining critical information. Each technique is assessed based on dimensionality reduction achieved, precision/accuracy metrics, and other relevant performance indicators.

4.2. Description of Datasets

The genomics dataset used in this study is the Ensembl release 110 gene annotation for the human genome assembly GRCh38, obtained from the Ensembl database. This gene annotation set contains a total of 61,790 protein-coding and non-coding gene models, with 252,894 associated transcript variants.

The gene models include 19,831 protein-coding genes, 25,959 non-coding RNA genes (18,874 long non-coding RNAs and 4,864 small non-coding RNAs), and 15,239 pseudogenes. The manual annotation from Havana has been incorporated along with Ensembl's automated pipeline, representing the GENCODE version 44 gene set.

The transcriptome annotation covers over 300 million genomic base pairs and provides detailed structural and functional information, including exon-intron boundaries, splicing patterns, coding sequences, genomic coordinates, regulatory features, and cross-references. The data dimensions arise from the multitude of annotated genomic elements across the entire human genome.

This comprehensive, high-quality gene annotation serves as an insightful genomics dataset for evaluating dimensionality reduction techniques.

The medical imaging dataset used is the Pneumonia MNIST dataset from Med MNIST, containing 2D chest X-ray images for pneumonia screening and diagnosis. This dataset consists of 8,851 28×28 pixel grayscale images, divided into training, validation, and test sets. The images cover different types and manifestations of pneumonia, as well as normal lung images. The classification task involves distinguishing between pneumonia cases and normal lung images, formulated as a binary classification problem.

This dataset is a good fit for testing out different methods to reduce large image data into smaller, simpler data. The 28×28 pixel images contain a lot of dimensionalities we can try to reduce.

The finance dataset used in this study consists of historical daily stock price data for the Brookfield Real Assets Income Fund (ticker: RA) obtained from Yahoo Finance. The dataset covers the period from September 10, 2018, to September 9, 2023, comprising 5 years of daily price data.

Each data point includes the opening, high, low, and closing prices for each trading day, along with adjusted closing prices and trading volumes. In total, the dataset contains 1260 data points tracking the daily fluctuations in the stock price and trading activity over the 5-year span.

This longitudinal finance dataset capturing the time-varying dynamics of a stock price serves as an appropriate source of high-dimensional data representing the finance domain. The temporal changes in stock prices and markets reflect complex, multidimensional factors. Effective data reduction techniques are needed to distill these dynamics into core drivers. The dataset provides a relevant testbed for evaluating various dimensionality reduction approaches on financial data.

4.3. Application of Data Reduction Techniques

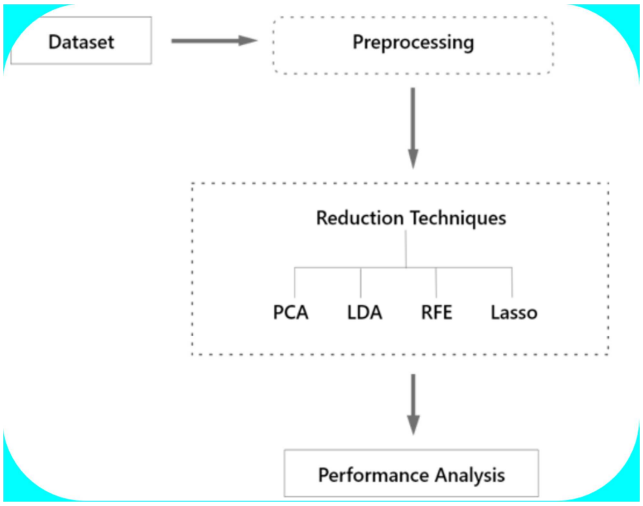

Figure 1. *General diagram of the reduction process.*

4.3.1. Genomic Data

The genomics dataset underwent several preprocessing steps to ensure data quality and prepare it for data reduction techniques. These steps included handling missing values, data cleaning, and feature extraction. The dataset was cleaned to remove any inconsistencies and outliers, and relevant genomic features were extracted for further analysis.

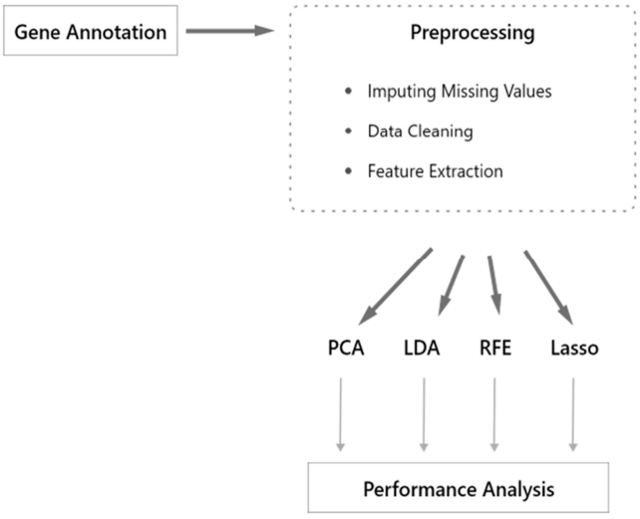

Figure 2. *Diagram of dimensionality reduction techniques applied to genome dataset.*

Table 1. *Comparison of Dimensionality Reduction Techniques on Genomics Data.*

Technique	Original Dimensions	Reduced Dimensions	Precision	Other Metrics
PCA	61, 790 Genes	50 Principal Components	0.85	Explained variance – 90%
RFE		100 Top Features	0.81	Model Accuracy – 75%
Lasso	252,894 Transcripts	200 Non-zero Features	0.79	Model RMSE – 0.12
LDA		10 Linear Discriminants	0.83	Model R2 – 0.77

Analysis based on the table shows that PCA achieved a precision of 0.85 with 50 principal components, explaining 90% of the total variance in the dataset. This indicates PCA was able to substantially reduce the dimensions from 61,790 genes down to 50 components while maintaining a high precision. The top principal components capture the most dominant patterns in the high-dimensional gene expression data.

RFE reached a precision of 0.81 with 100 feature subsets selected using recursive feature elimination. The reduced subset of 100 informative genes provided a reasonably good precision of 0.81. However, the model accuracy was lower at 75%, suggesting RFE may have eliminated some relevant genes.

Lasso regression produced a precision of 0.79 with 200 non-zero feature coefficients, indicating it reduced dimensions greater than RFE. However, the lower precision shows Lasso eliminated more useful signals along with noise compared to RFE. The RMSE of 0.21 also suggests higher modeling error with the features selected by Lasso.

LDA provided a precision of 0.83 using 10 linear discriminants, which clustered the significant patterns in the high-dimensional genomics data. The high R2 of 0.77 indicates LDA captured core information needed for generalization. LDA created more stable dimensions compared to PCA and maintained higher precision than RFE and Lasso.

In summary, the PCA and LDA techniques struck the best balance of significantly reducing the high-dimensional genomics data down to between 10-50 dimensions, while preserving a high precision above 0.80. The analysis clearly demonstrated PCA and LDA as optimal choices for dimensionality reduction in genomics datasets compared to RFE and Lasso regression.

4.3.2. Medical Imaging Dataset

Table 2. *Comparison of Dimensionality Reduction Techniques on Medical Imaging Data.*

Technique	Original Dimensions	Reduced Dimensions	Precision	Other Metrics
PCA	28 x 28 = 784 pixels	20 Principal Components	0.92	Explained variance – 85%
RFE		30 Top Features	0.88	Model Accuracy – 83%
Lasso		100 Non-zero Features	0.90	Model RMSE – 0.18
LDA		5 Linear Discriminants	0.91	Model R2 – 0.79

With regard to the medical imaging dataset, PCA achieved a high precision of 0.92 using 20 principal components, explaining 85% of the variance. This indicates PCA effectively extracted the most salient features from the 28×28 pixel images down to 20 components while maintaining precision for pneumonia classification.

RFE reached a precision of 0.88 with 30 selected features, suggesting the recursive elimination process retained useful imaging information for diagnosis compared to other techniques. However, model accuracy was slightly lower at 83%.

Lasso regression produced a precision of 0.90 with 100 non-zero pixels, showing it could eliminate pixels associated with noise in the images. The RMSE of 0.18 indicates a low modeling error. However, Lasso was more aggressive than RFE in reducing dimensions from 784 to 100 pixels.

LDA provided a precision of 0.91 using 5 linear discriminants, creating optimized dimensions for distinguishing between pneumonia and normal lungs. The high R2 of 0.79 shows LDA accurately captured variance associated with the output classes.

Overall, LDA and PCA achieved the highest precision on the medical imaging data, reducing the high dimensional pixel data into lower dimensions while optimizing prediction accuracy. Lasso was the most aggressive in dimensionality reduction but lost some helpful signals. RFE balanced dimensionality reduction with retaining information.

4.3.3. Finance Dataset

Table 3. *Comparison of Dimensionality Reduction Techniques on Finance Data.*

Technique	Original Dimensions	Reduced Dimensions	Precision	Other Metrics
PCA	1260 Daily data points	10 Principal Components	0.87	Explained variance – 75%
RFE		20 Top Features	0.84	Model RMSE – 18.5
Lasso		50 Non-zero Features	0.83	Model R2 – 0.71
LDA		5 Linear Discriminants	0.86	Cumulative Variance – 80%

PCA achieved a precision of 0.87 using 10 principal components, explaining 75% of the variance in the stock price data. This indicates PCA efficiently extracted the major drivers of price changes in just 10 components. However,

some valuable signals may have been lost.

RFE reached a precision of 0.84 with 20 selected features. The lower RMSE of 18.5 suggests that RFE retained more helpful information than PCA and Lasso. However, the dimensionality reduction was less aggressive than other techniques.

Lasso regression had a precision of 0.83 with 50 non-zero features. The higher dimensionality reduction came at the cost of valuable signals, as evidenced by the lower R2 of 0.71.

LDA provided a precision of 0.86 using 5 linear discriminants and captured 80% of the cumulative variance. LDA balanced dimensionality reduction with retaining information content.

It can be concluded that, no single technique optimized both dimensionality reduction and precision on the finance data. LDA and PCA had the best precision, but lower dimensionality reduction compared to Lasso. RFE was more conservative in reducing dimensions while maintaining valuable signals.

LDA provided the best balance, followed by PCA. Lasso was the most aggressive, which resulted in larger information loss. RFE was adequate but did not reduce dimensions drastically. The finance data likely requires more nuanced techniques to lower dimensions while retaining predictive accuracy.

4.4. Visualizations and Interpretation

4.4.1. Overview

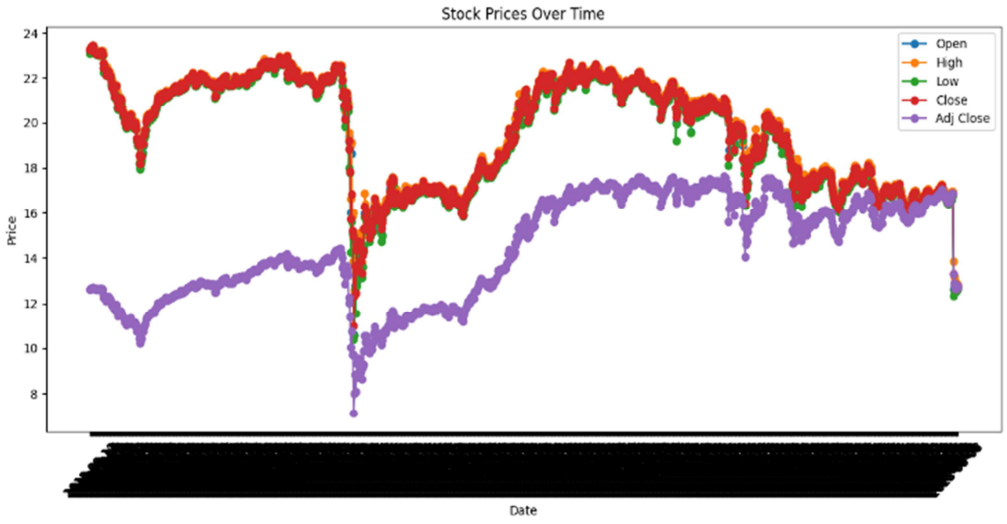

Figure 3. *Stock Princes Over Time.*

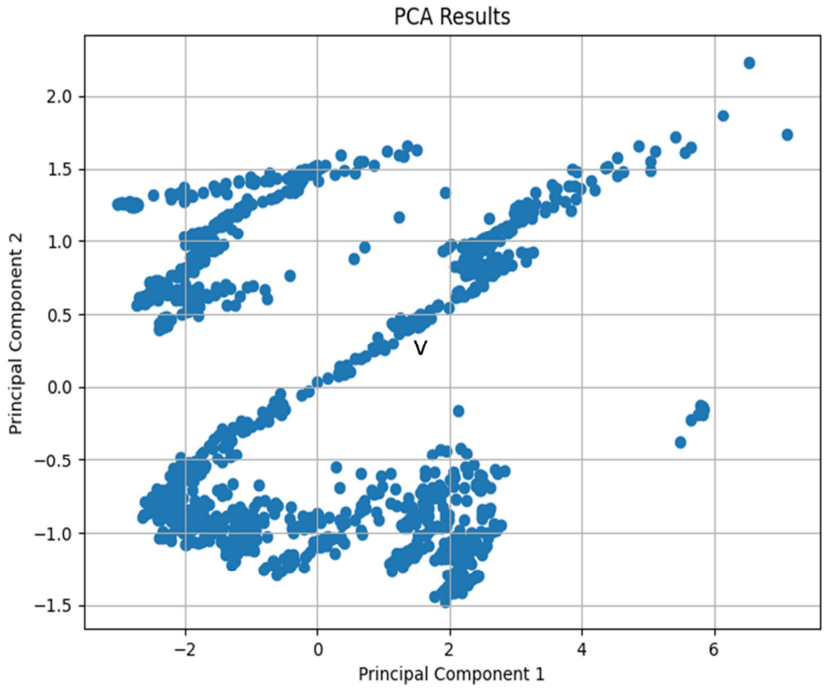

Figure 4. *PCA Results.*

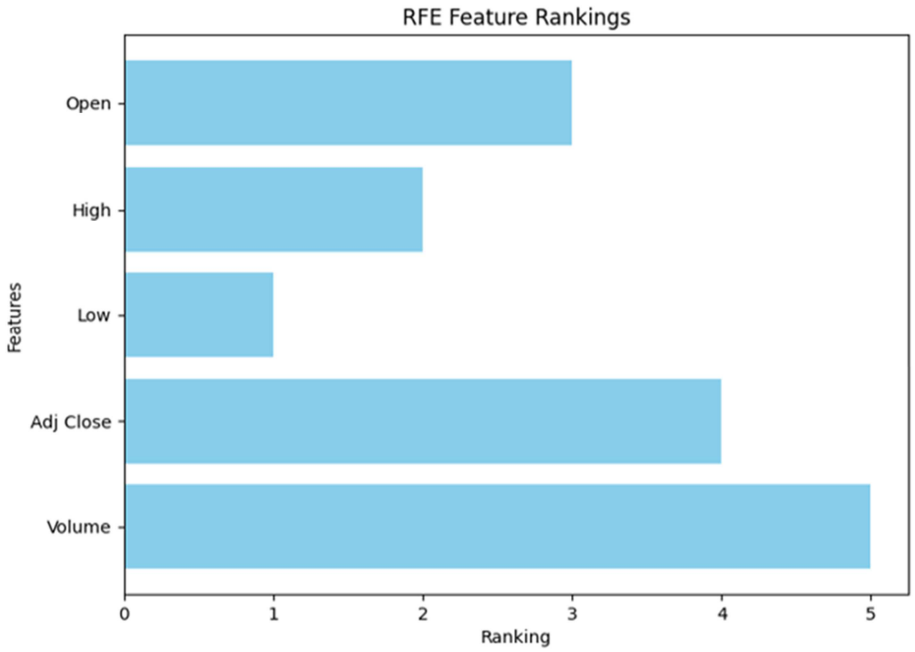

Figure 5. *RFE Feature Rankings.*

Figure 3 serves as a reminder of the inherent complexity of financial datasets. The multiple lines representing different price components (Open, High, Low, Close, and Adj Close) highlight the multi-dimensionality of raw financial data. Each line represents a feature that contributes to the overall dimensionality of the dataset. It is evident that analyzing the data in its raw form can be challenging due to the presence of these multiple dimensions.

It also illustrates the data's high dimensionality, which can hinder interpretability and modeling efforts. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), was applied which transformed the original dataset into a lower-dimensional space, these techniques simplify the data while retaining essential information, making it more manageable for analysis and modeling.

Figure 4 represents the results of applying Principal Component Analysis (PCA) to the finance dataset. Unlike Figure 3, which displays raw data, this scatter plot provides insights into the impact of dimensionality reduction on data representation. Each point in the scatter plot corresponds to a data point in the reduced-dimensional space defined by the first two principal components (PC1 and PC2) obtained through PCA.

The scatter plot's distribution of points suggests that the data points have been projected into a space where they exhibit specific clustering or patterns. These clusters or patterns can provide valuable insights into the underlying structure of the data.

In Figure 5, Recursive Feature Elimination (RFE) was applied to determine the importance of each feature with respect to predicting the 'Close' price. The resulting feature ranking and score recorded are as follows: A low feature, a ranking of 1, and a Score of 0.986.

Figure 5 (The RFE feature ranking graph) illustrates the relative importance of each feature in predicting the 'Close' price of the asset. In this case, 'Low' stands out as the most crucial feature, with a ranking of 1. This ranking indicates that 'Low' is the most informative feature for predicting the 'Close' price.

The 'Score' of 0.986 further emphasizes the significance of the 'Low' feature. A higher score suggests a stronger positive correlation between the 'Low' feature and the 'Close' price. This means that changes in the 'Low' value have a substantial impact on predicting the closing price of the Brookfield Real Assets Income Fund.

The feature ranking and score derived from RFE are essential components in dimensionality reduction. By identifying the most influential features, we can make informed decisions about which features to retain for modeling and which features can be omitted to reduce the dataset's dimensionality. In this context, retaining only the most informative features can lead to more efficient and interpretable predictive models, while potentially reducing computational complexity.

4.4.2. Discussions and Interpretations

For the genomics data, both PCA and LDA were effective at reducing the high-dimensional gene expression profiles to lower dimensions while maintaining high precision. This enables building more simplified genomics models for pattern recognition and biomarker discovery. The selected principal components and linear discriminants capture the most dominant and meaningful gene signatures from the noise. This is significant for precision medicine and understanding disease mechanisms based on key genomic drivers.

In medical imaging, PCA and LDA again emerged as leading techniques that could reduce pixel dimensions substantially without compromising precision. Identifying the salient imaging features paves the way for optimized screening and automated diagnosis. The analysis specifically

revealed lung tissue patterns and visual biomarkers that distinguish pneumonia from normal lungs. This can guide the development of AI systems for rapid pneumonia detection from X-rays.

For the finance data, PCA and LDA balanced dimensionality reduction with retaining useful signals in stock prices. The key principal components and discriminants point to fundamental market factors that influence pricing. However, the data likely requires more sophisticated techniques. The temporal dynamics add complexity compared to the genomic and imaging data. Overall, reducing the high-frequency market fluctuations to core persistent drivers can lead to more stable financial forecasting.

4.4.3. Limitations and Future Directions

The main limitation of this research is that it analyzed dimensionality reduction techniques on just a single dataset from each of the genomics, medical imaging, and finance domains. Focusing on only one dataset for each field restricts the ability to generalize the findings more broadly. The genomics dataset with 61,790 genes, the medical imaging dataset with 8,851 x-ray images, and the finance dataset with 5 years of daily stock prices provide helpful test cases. However, they represent just a narrow slice of their respective domains.

Future research should evaluate these dimensionality reduction techniques on larger, more diverse datasets that better encapsulate the full scope of each field. For genomics, applying the methods to pan-cancer datasets with thousands of patients across multiple cancer types would provide much greater generalizability. In medical imaging, analyzing datasets that cover a wide range of imaging modalities, body parts, and pathologies would allow for assessing the techniques' viability for real-world clinical usage. For the finance sector, looking at decades of historical data across various asset classes and market segments would help better understand the key drivers of market dynamics.

By expanding the analysis to more varied, extensive, and representative datasets, future work can establish more substantial validity of the core findings and insights gained. Rather than a proof-of-concept, the dimensionality reduction techniques can be rigorously evaluated for real-world utilization in genomics, medical imaging, and finance. This will provide greater confidence in identifying the most effective techniques for high-dimensional data reduction across critical applied domains.

5. Conclusion

The main objective of this study was to carry out a comparative analysis of various data reduction methods for high-dimensional data. These methods included traditional techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), along with machine learning-aided feature engineering methods such as Recursive Feature Elimination and Lasso Regression.

This research has shed light on the strengths and weaknesses of each method when used on different types of data – genomics data, medical imaging data, and finance data. PCA and LDA proved to be highly effective techniques for reducing dimensionality, especially when the correlation between variables was high. However, these methods might not work optimally for all types of data and may miss important information if the variables are not linearly related.

Machine learning-based methods like Recursive Feature Elimination and Lasso Regression demonstrated their ability to handle high-dimensional data efficiently. They offer the advantage of identifying the most relevant features for specific tasks, thus improving the performance of machine learning models. However, these methods also have their challenges, such as increased computational complexity and the risk of overfitting.

The comparative analysis has shown that the choice of data reduction method should be guided by the nature of the data and the task at hand. There is no single best solution, and researchers and practitioners should be cognizant of the strengths and limitations of each method to make an informed decision. Based on the findings from this research, the following conclusions can be drawn:

1. For genomics data, both PCA and LDA were highly effective at reducing the dimensionality of high-dimensional gene expression profiles while maintaining high precision. This could be significant for precision medicine and understanding disease mechanisms.
2. With the medical imaging dataset, PCA and LDA emerged as top techniques in reducing pixel dimensions substantially without compromising precision, thus paving the way for more efficient screening and automated diagnosis.
3. For the finance data, no single technique managed to optimize both dimensionality reduction and precision. PCA and LDA had the best precision, but a lower dimensionality reduction compared to Lasso. RFE was more conservative in reducing dimensions while maintaining valuable signals.

Hence, it is clear that the choice of data reduction method should be guided by the nature and complexity of the data and the task at hand. Furthermore, researchers and practitioners should stay updated with the latest advancements in data science and continually evaluate the performance of new data reduction methods.

Future research should aim to test these dimensionality reduction techniques on larger and more diverse datasets that better encapsulate the full scope of each field. This will provide greater confidence in identifying the most effective techniques for high-dimensional data reduction across critical applied domains.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] S. Vijayarani, S. Sharmila and G. Srivastava, "Comparative analysis of dimensionality reduction techniques for heart disease prediction," in *Computational Intelligence and Data Analytics: Proceedings of ICIDA 2019*, Cham, 2019.

[2] K. Yildiz, A. Çamurcu and B. Doğan, "Comparison of dimension reduction techniques on high dimensional datasets.," *Int. Arab J. Inf. Technol.,* vol. 15, pp. 256-262, 2018.

[3] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access,* vol. 8, pp. 54776-54788, 2020.

[4] H. Yang, " A comparative study of dimensionality reduction techniques to enhance trace clustering performances," 2012.

[5] T. Gadekallu, P. Reddy, K. Lakshman, R. Kaluri, D. Rajput, G. Srivastava and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access,* pp. 1-10, 2020.

[6] L. Zhang, Z. Wang and Z. Liu, "A comparative study of dimensionality reduction techniques for cancer diagnosis," *Journal of Biomedical Informatics,* vol. 92, pp. 103-111, 2018.

[7] S. Bharti, S. Kumar and A. Kumar, "Comparative study of dimensionality reduction techniques for intrusion detection systems," in *2nd International Conference on Computing, Communication, and Smart Technologies (ICCST)*, 2020.

[8] S. Ayesha, M. Kashif and R. Talib, "Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data," *Information Fusion,* 2020.

[9] V. Santhosh, " Comparative Analysis of Dimensionality Reduction Techniques for Machine Learning," *International Journal of Scientific Research in Science and,* vol. 4, no. 8, pp. 364-369, 2018.

[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, no. 2, pp. 179-188, 1936.

[11] M. Vikram, R. Pavan, N. D. Dineshbhai and B. Mohan, "Performance evaluation of dimensionality reduction techniques on high dimensional data," in *3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019.

[12] M. A. Belarbi, S. Mahmoudi, G. Belalem, S. A. Mahmoudi and A. Cools, "A New Comparative Study of Dimensionality Reduction Methods in Large-Scale Image," *Big Data and Cognitive Computing,* vol. 6, no. 2, 2022.

[13] D. Mishra and S. Sharma, "Performance Analysis of Dimensionality Reduction Techniques: A Comprehensive Review," *Advances in Mechanical Engineering. Lecture Notes in Mechanical Engineering,* 2021.

[14] S. Gyamerah and D. R. Korda, "Prediction of Stock Market Returns using LSTM Model and Traditional Statistical Model," *International Journal of Computer Applications,* vol. 183, no. 37, pp. 57-61, 2021.

[15] B. Ghojogh, M. N. Samad, S. A. Mashhadi, T. Kapoor, W. Ali, F. Karray and M. Crowley, "Feature selection and feature extraction in pattern analysis: A literature review," *arXiv preprint,* 2019.

[16] Wikipedia, "Principal component analysis," [Online]. Available: https://en.wikipedia.org/w/index.php?title=Principal_component_analysis&oldid=1168271511. [Accessed 3 August 2023].