INVESTIGATION INTO THE EFFECT OF DIMENSIONALITY REDUCTION TECHNIQUES IN MACHINE LEARNING ALGORITHMS

By
OLAWALE OMOTOSHO

A thesis submitted to the Faculty of Science & Engineering of the University of Hull in partial fulfilment of the requirement for the degree of Master in Artificial Intelligence and Data Science.

January 2024

Abstract: Machine learning algorithms are increasingly relied upon to extract meaningful insights from high-dimensional data. However, as data complexity grows, the computational demands of these algorithms can become overwhelming. Dimensionality reduction techniques offer a viable solution to this challenge, by transforming high-dimensional data into a lower-dimensional representation while preserving essential information. This research investigated the intricate relationship between dimensionality reduction (DR) techniques and machine learning algorithms for classification tasks. Its goal was to optimize model performance by understanding how DR methods impact different algorithms. Three DR techniques (PCA, LDA, UMAP) were applied to four machine learning algorithms (SVM, Gradient Boosting, Random Forest, and Logistic Regression) across imbalanced and balanced datasets. Performance was evaluated using accuracy, precision, F1 score, and generalization capabilities. DR technique selection significantly influenced model performance. LDA-SVM pairing achieved remarkable accuracy (0.99) and robustness on imbalanced data, showcasing their complementary strengths in maximizing class separation. UMAP emerged as a viable alternative to linear DR methods. Despite similar average accuracy to PCA, UMAP demonstrated superior convergence and mitigated overfitting, aligning with its envisioned potential for optimizing performance. Choosing the right DR technique for specific algorithms and data scenarios is crucial for optimizing performance and generalization. Future research should explore ensembles of DR techniques and address UMAP's interpretability trade-off.

Keywords: Dimensionality reduction, machine learning, feature selection, LDA, PCA, UMAP

1. INTRODUCTION

The ever-increasing dimensionality of datasets presents a significant challenge in the field of machine learning, where the overwhelming abundance of features can impede model training and lead to overfitting. One of the more pressing issues in data-based modelling is the presence of high dimensionality (Wan et al., 2021). This study contributes to this ongoing discourse by comprehensively examining the impact of various dimensionality reduction methods on the performance of established machine learning algorithms.

The realm of dimensionality reduction is characterised by two primary approaches: feature selection and feature extraction (Lan, 2011; Mendez, 2022). Feature extraction, as defined by Khalid et al. (2020), involves transforming the original features into a new set that encapsulates the essence of the data's patterns and characteristics. In contrast, feature selection identifies and selects a subset of the most crucial features that best represent the data (Zebari et al., 2020).

Previous studies have laid the groundwork for understanding the intricate relationship between dimensionality reduction and the efficacy of machine learning models.

The work of Fournier and Aloise (2019) delves into the application of Autoencoders for dimensionality reduction, particularly in comparison to traditional methods such as PCA (Principal Component Analysis) and Isomap (Isometric Feature Mapping).

Sharma and Dey (2012) conducted a comprehensive study exploring various methods and their impact on classification performance. Their findings resonate with those of Saeys et al. (2008), who demonstrated the strength of the SVM-RFE combination for cancer data, highlighting the potential for specific algorithms and feature selection to work in synergy.

Data scientists often face a challenging dilemma when it comes to dimensionality reduction: selecting the most appropriate techniques to optimize model performance. Pham et al. (2022) tackled this challenge by identifying critical features for crop yield forecasting models. Their study meticulously compared the performance of feature selection (FS), feature extraction (FX), and their combination (FSX). The results demonstrated that the FSX approach outperformed both FS and FX individually, highlighting the synergistic benefits of integrating these methods. This observation aligns with the findings of Kabir et al. (2023), who applied feature selection in conjunction with PCA to showcase the positive impact of this combined approach on various machine learning algorithms in building predictive models for prostate cancer.

Shaik (2019), demonstrated the effectiveness of Linear Discriminant Analysis (LDA) in reducing noise and capturing the essential features of high-dimensional data, leading to improved model interpretability. Abdul Salam et al. (2021), comprehensive study undertakes a detailed comparative analysis of nine dimensionality reduction methods, including Factor Analysis (FA), Linear Discriminant Analysis (LDA), and random forest. Their investigation, encompassing Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest Classifier (RFC), reveals that RFC, in particular, demonstrates proficiency in achieving dimensionality reduction, effectively mitigating overfitting while maintaining or even improving classification accuracy.

Similarly, Wan et al. (2021) delve into the influence of dimensionality reduction on the prediction of concrete compressive strength using Support Vector Regression (SVR), XGBoost, and Artificial Neural Network (ANN). Their findings showcase that, depending on the model and selected features, dimensionality reduction positively impacts prediction accuracy and training speed. Notably, for SVR, the

model with manually selected or PCA-selected features outperforms the original feature set, further underlining the nuanced effects of dimensionality reduction on different machine-learning models.

In the study carried out by Kasun et al. (2017), the efficacy of linear and non-linear dimensionality reduction methods are put to the test as they compared PCA to Autoencoders (AE). The study showed that Autoencoders outperformed PCA at capturing specific and distinct features. However, a seminal review by van der Maaten et al. (2009) surveys and compares various dimensionality reduction techniques, concluding that nonlinear methods, while promising, are yet to surpass the performance of Linear methods like PCA.

Motivated by previous research, this work aims to systematically assess the impact of dimensionality reduction (DR) techniques on the classification performance of diverse machine learning algorithms including RF, LR GB and SVM classifiers. It investigates both traditional methods like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) alongside the newer non-linear approach, Uniform Manifold Approximation and Projection (UMAP). This analysis will illuminate the suitability and effectiveness of these DR techniques for feature reduction in classification tasks across various machine learning models

Section 2 details the data acquisition, processing and preparation for classification tasks, following this;

Section 3 provides a comprehensive explanation of the chosen DR techniques and an overview of the modelling architecture

Section 4, presents the results and analysis obtained from the models.

Section 5, discusses the key findings and conclusions drawn from the analysis

2. Data

2. 1 Data collection, cleaning and pre-processing

A lung cancer dataset was employed for this project obtained from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial [PLCO - The Cancer Data Access System]. The PLCO dataset is a large-scale, prospective study that enrolled over 150,000 participants and collected data on their health, lifestyle, and cancer history. The dataset further includes information on lung cancer incidence, as well as data on other potential risk factors for lung cancer, such as smoking history, family history of cancer, and environmental exposures. Ultimately, a classification task was carried out on this dataset and the results with and without the application of dimensionality reduction techniques are evaluated for the objective of this research.

The patients were de-identified to ensure their privacy and confidentiality in adhering to ethical guidelines.

The PLCO dataset was pre-processed before use in this study. This involved cleaning the data for missing values, removing outliers, and normalizing variables. To fortify the integrity of the dataset, an in-depth exploratory data analysis (EDA) was conducted. The initial step involved identifying and addressing missing or NaN values through strategic imputation using column medians. This process ensured data completeness and laid the groundwork for subsequent analyses.

A pivotal aspect of the EDA included extracting the top 30 correlated features. This step, vital to the study's objective, ensured that the subsequent analysis considered the most influential variables.

Data Balancing

Initial exploratory data analysis (EDA) revealed a significant imbalance in the dataset, with an overrepresentation of negative lung cancer cases compared to positive ones. This imbalance posed a potential risk of bias in the performance of machine learning models. Notably, the target variable (Y), representing positive lung cancer cases, constituted the minority class.

To address this issue, a hybrid balancing technique was employed. It combined oversampling of the minority class using the Synthetic Minority Oversampling Technique (SMOTE) to increase its representation, and undersampling of the majority class using the NearMiss algorithm to maintain a manageable dataset size. This approach created a more balanced dataset and mitigated potential bias in the subsequent modelling process.

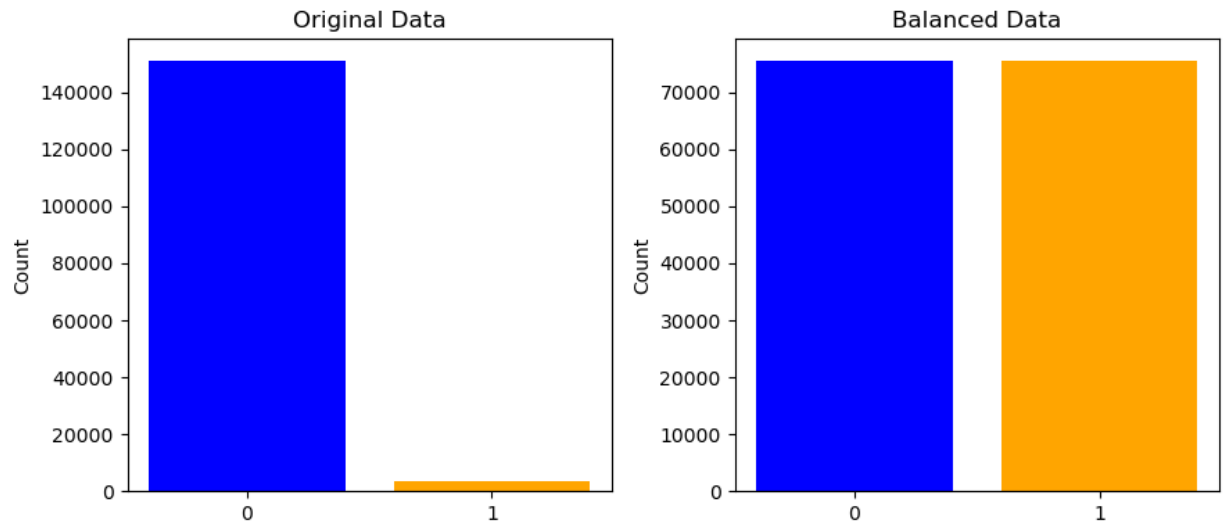

Figure 1: Two bar plots visualizing the distribution of classes (0 and 1) in the lung cancer dataset before and after applying a balancing technique. The left plot is the original data distribution, right plot is balanced data.

Class Labels

For the classification task, the primary focus is predicting the presence of positive lung cancer cases. The target variable is denoted as “lung_cancer” in the data. The class labels are 0 and 1. Where 1 denotes the presence of lung cancer and 0 signifies the absence.

3. Methodology

3.1 Feature Selection

Feature selection played a crucial role in the experimental set-up, ensuring that only the most relevant and informative features were utilised for subsequent analyses.

Variables introduced after a person's final "lung_cancer" diagnosis and factors associated with "death" were staunchly excluded from consideration of my selections to maintain the study's focus.

Feature selection was initiated by applying two techniques to generate feature importance: Random Forest Classifier and SelectKBest.

Random Forest Classifier, a robust ensemble learning method, constructs multiple decision trees to generate predictions (Akhiat et al., 2021). This algorithm yields feature importance scores, which quantify the relative impact of each feature on model predictions. The top 40 features were extracted based on their importance, ensuring the retention of the most influential factors.

Additionally, SelectKBest, a filter-based technique, was employed to assess each feature's ability to discriminate between the two classes (Desyani et al., 2020). This method identifies features exhibiting statistically significant differences in their distribution between the lung cancer and non-lung cancer classes. The top 40 features were again extracted based on their F-scores.

By strategically fusing the insights from both methods, a comprehensive set of 19 highly relevant features was created. This provided a focused set for further analysis, boosted model performances, and facilitated a clearer understanding of the feature-target variable relationships.

3.2 Dimensionality Reduction Techniques

The crux of this research lies in investigating the impact of DR techniques. Three reduction techniques were systematically applied, and their impacts on model performances were rigorously assessed

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the data while preserving the most significant variance. This linear technique, as described by Kabir et al. (2023) and Abdul Salam et al. (2021), projects the original data onto a new coordinate system where the principal components capture the directions of maximum variance. The first principal component captures the highest variance in the data, followed by the second component capturing the maximum variance in the residual data, and so on. This transformation effectively reduces the number of features while retaining as much information as possible.

Component Selection: The number of principal components to retain was carefully selected based on the cumulative variance ratio. Initially, the n_components was set to 8 for the first phase of the research on the imbalance data to capture the maximum variance ratio as possible. For the balanced dataset, the first two components captured 98% of the variance, indicating their ability to represent the majority of the data's information. Therefore, PCA was applied with 2 components for the latter phase.

Projecting the data onto the first two principal components yielded an informative 2D scatter plot which presented a crucial observation in the data. Significant overlap in the two classes. This insight reveals the similarities the both classes share in feature space.

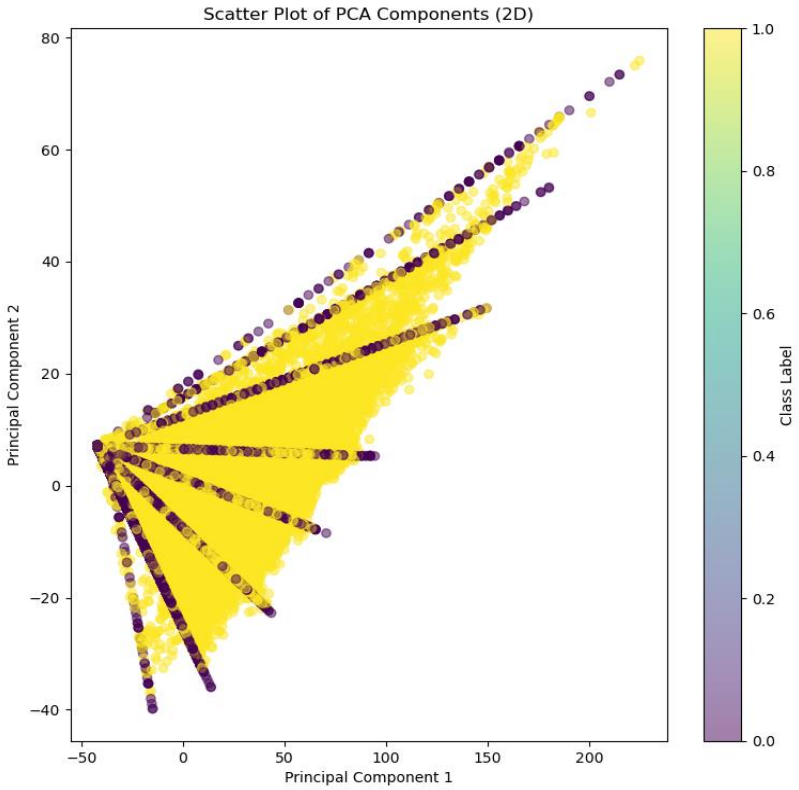

Figure 2: 2D scatter plot visualizing the distribution of balanced training data samples after dimensionality reduction using Principal Component Analysis (PCA). Each point represents a data sample, coloured according to its class label. Visual overlap between colours indicates difficulty in separating classes using linear models.

Previous works that have employed autoencoders for dimensionality reduction on similar cancer datasets (Biggs et al, 2023; Kabir et al, 2023), this project explores the efficacy of **Linear Discriminant Analysis (LDA)**. LDA stands out as a supervised dimensionality reduction technique for its ability to enhance classification performance (Tang et al, 2005).

LDA's core strength is in its ability to make different classes more distinguishable making it simpler for machine learning models to recognize unique patterns between classes (Tang et al, 2005).

Table 1: Hyperparameters employed across ML Algorithms

Classifier	Hyperparameter	Values	Dimensionality Reduction Techniques	n_components aplied
RFC	n_estimators	100	PCA	n=8 n=2
	max_depth	None		
	min_samples_split	2		
	min_samples_leaf	1		
	random_state	None		
LR	C	1.0		
	solver	Lbfgs	LDA	n=1
	random_state	0		
GB	learning_rate	0.1		
	n_estimators	100		
	max_depth	3		
	random_state	0		
SVM	kernel	Rbf	UMAP	n=10
	C	1.0		
	gamma	'auto'		
	random_state	0		

In a deeper exploration of this analysis, a non-linear dimension reduction technique, **Uniform Manifold Approximation and Projection (UMAP)** was employed. This choice was motivated to potentially address the ongoing discussions surrounding the efficacy of linear versus non-linear as part of my findings. UMAP's unique approach seeks to find a low-dimensional representation of the data that closely resembles the original high-dimensional space (McInnes et al., 2020). Unlike linear techniques like PCA, which can introduce distortions to the data's intricate structure, UMAP effectively preserves the local relationships between data points. By projecting the high-dimensional data onto a lower-dimensional subspace that maintains this local structure, UMAP enables the identification of patterns and insights within the data that might otherwise remain hidden in the original high-dimensional space (Allaoui et al., 2020; McInnes et al., 2020). Moreover, UMAP's computational efficiency makes it suitable for handling even large-scale datasets. After careful tuning, the number of components in the UMAP projection was set to 10 to balance information retention with computational efficiency and time limitations.

3.3 Proposed Methodology

Four machine learning algorithms (Logistic Regression, Random Forest Classifier, Gradient Boosting, and Support Vector Machine) were selected for assessing the impact of dimensionality reduction. Each algorithm was trained and evaluated on both the original high-dimensional data and versions thereof processed with different dimensionality reduction techniques.

A total of 28 models were trained and evaluated. This included variations in classifiers and dimensionality reduction methods applied to both balanced and imbalanced datasets. The purpose of this design was to investigate the effects of dimensionality reduction on performance, both with and without balancing the data beforehand.

Evaluation methods included learning curves for visualizing training dynamics (convergence, overfitting, underfitting), Receiver Operating Characteristic (ROC) curves for analyzing classification capabilities across different dimensionality reduction scenarios, and classification report scores and model training times. This comprehensive approach aimed to provide a deeper understanding of how the chosen algorithms handle reduced-dimensional data

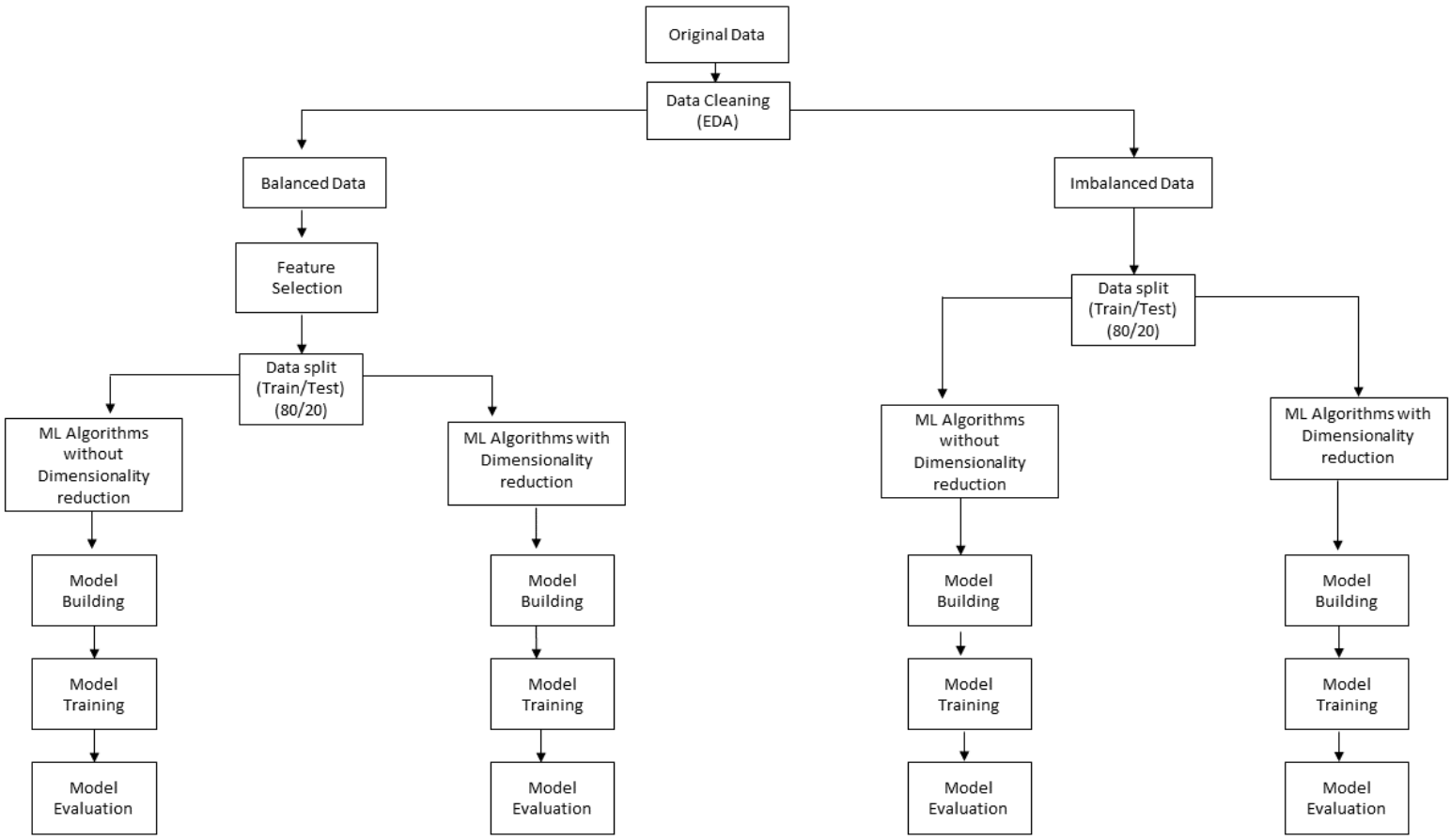

Figure 3: Methodology Flowchart

4. Results

This experiment explored the intricate interplay between three dimensionality reduction (DR) techniques and four machine learning algorithms. Its objective was to uncover hidden patterns and relationships, ultimately optimizing the performance of predictive classification models. The results highlighted the crucial influence of DR selection on specific algorithms.

The initial phase of the dimensionality reduction (DR) investigation focused on established techniques identified through a comprehensive literature review. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were both applied to the imbalanced dataset to evaluate their effectiveness in feature extraction. As shown in Table 2, the Support Vector Machine (SVM) initially struggled to accurately identify positive cases, resulting in low precision and F1 scores. However, the application of LDA in conjunction with SVM dramatically improved performance, achieving an impressive accuracy of 0.99, perfect precision of 1.0, and a strong F1 score of 0.88. This significant improvement can be attributed to LDA's focus on maximizing separation between classes (Tang et al., 2005) while PCA focuses more on capturing variance (Sachin, 2015). SVM took advantage of the enhanced separability when coupled with LDA.

Interestingly, Random Forest Classifier presented a contrasting behaviour. While its base model achieved perfect accuracy (1.0), both RF-PCA and RF-LDA models showcased longer training times compared to the original model. This observation diverged from the general trend observed in most other algorithms, which typically is a result of the ensemble nature of Random Forest which builds multiple decision trees, each trained on a random subset of features and samples. This inherent randomness can make them less sensitive to dimensionality reduction techniques like PCA, which aim to uncover linear relationships between features.

It is also worth noting that PCA-reduced data consistently exhibited challenges in achieving high Precision and F1 scores across all classifiers.

Table 2: Model evaluation metrics on the original dataset

Classifier Model	**Accuracy**	**Precision**	**F1**	**Training Time(s)**
RF	1.00	1.00	1.00	11.04
LR	0.99	0.92	0.74	4.58
GB	1.00	1.00	1.00	67.74
SVM	0.98	0.00	0.00	1475.95
RF-PCA	0.98	0.58	0.22	16.11
LR-PCA	0.98	0.00	0.00	1.00
GB-PCA	0.98	0.65	0.11	20.87
SVM-PCA	0.98	0.00	0.00	60.95
RF-LDA	0.99	0.90	0.83	13.30
LR-LDA	0.99	0.99	0.87	0.20
GB-LDA	0.99	0.98	0.87	7.48
SVM-LDA	0.99	1.00	0.88	14.04

While all the models attained high accuracy with and without DR, their corresponding training and validation curves reveal potential overfitting caused by the dataset's imbalance. This suggests that the models might struggle to generalize to new, unseen data. The sole exception was the combination of LDA-reduced data and the Gradient Boosting (GB) base model, as seen in Figure 4. While this model experienced slight decreases in accuracy, F1 score, and precision, it demonstrated a corrected curve, indicating its ability to learn new data effectively. Overall, dimensionality reduction using LDA on the GB classifier emerged as the most robust approach for handling imbalanced data, achieving a remarkable

accuracy of 0.99 and a fitting curve with potential for improvement with hyperparameter tuning. This performance aligns with real-world scenarios where class imbalances are common.

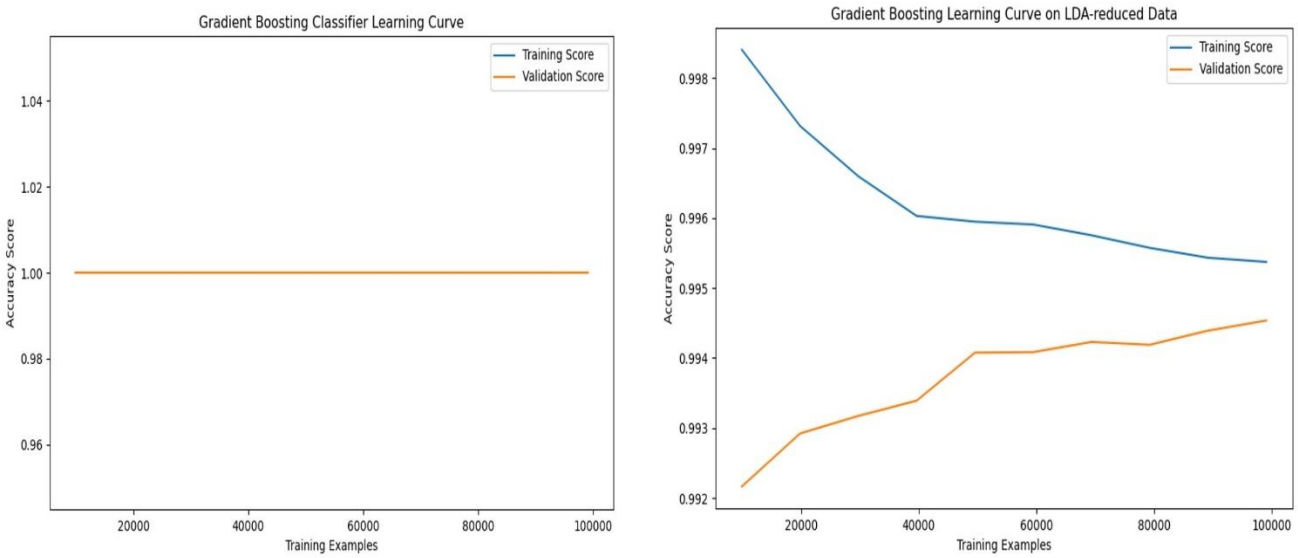

Figure 4: Gradient Boosting Learning Curves Comparing Generalization Before and After Dimensionality Reduction with LDA on Imbalanced Dataset. Left plot: Learning curve before dimensionality reduction exhibits overfitting and poor generalization to unseen data. Right plot: Learning curve after dimensionality reduction demonstrates a reduced gap between curves, suggesting improved generalization and reduced overfitting.

Table 3: Model evaluation metrics on Balanced-dataset

Classifier Model	**Accuracy**	**Precision**	**F-1**	**Training Time(s)**	AUC
RF	0.99	0.99	0.99	6.76	1.00
LR	0.94	0.98	0.94	0.82	0.98
GB	0.99	1.00	0.99	20.63	1.00
SVM	0.92	0.95	0.91	686.75	0.98
RF-PCA	0.96	0.98	0.96	12.90	1.00
LR- PCA	0.74	0.73	0.75	0.16	0.93
GB- PCA	0.78	0.76	0.80	8.31	0.98
SVM- PCA	0.77	0.72	0.79	3932.25	0.93
RF-LDA	0.92	0.93	0.92	12.18	0.96
LR-LDA	0.92	0.95	0.92	0.07	0.98
GB-LDA	0.93	0.98	0.92	6.42	0.98
SVM-LDA	0.93	0.99	0.92	170.21	0.96
RF-UMAP	0.91	0.93	0.91	42.90	0.97
LR-UMAP	0.70	0.68	0.72	0.28	0.73
GB-UMAP	0.83	0.80	0.84	43.51	0.90
SVM-UMAP	0.81	0.78	0.82	466.13	0.89

The second phase of this research saw the introduction of the third DR technique, UMAP, a non-linear method, subsequently having a similar average accuracy of 81% to PCA across all four algorithms setting the base for the analysis. Generally, DR decreases the general performance in model accuracy, precision, and f1 scores across most classifiers. However, the extent of this impact varied depending on the specific technique and classifier (see Table 3).

Accuracy vs Generalization: Examining the learning curves revealed a pivotal role of dimensionality reduction in addressing overfitting, particularly evident in the Gradient Boosting (GB) classifier. Despite the overall accuracy dip, the reduction in overfitting implies a potential enhancement in the model's ability to generalize to new, unseen data.

Furthermore, all three dimensionality reduction techniques (PCA, LDA, UMAP) contributed to a significant reduction in training times. This highlights the trade-off between model accuracy and training efficiency, where dimensionality reduction facilitates faster model training at the expense of a slight decrease in accuracy

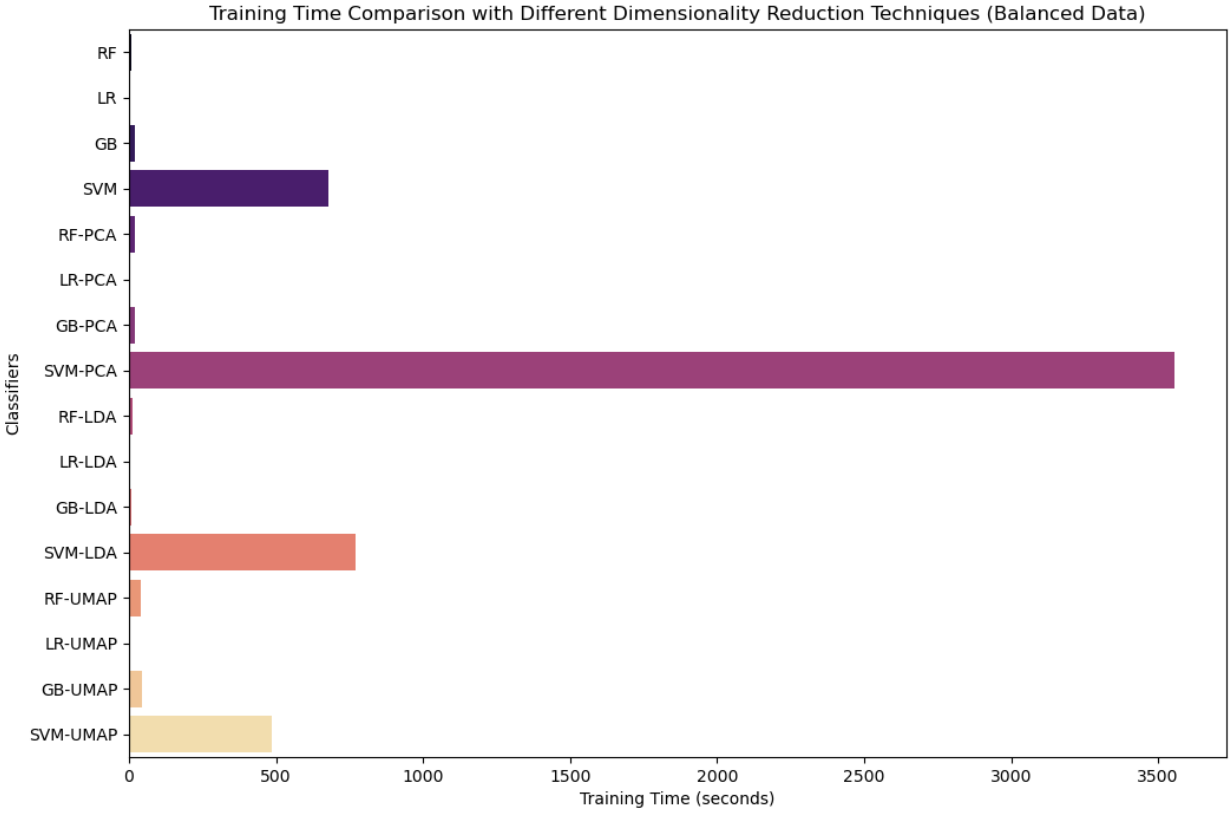

Figure 5: Bar plot visualizing the training time in seconds for different machine learning classifiers, both with and without applying dimensionality reduction techniques, on a balanced dataset.

Intriguingly, SVM-PCA stands out from other DR techniques by exhibiting a significant increase in computational time. This suggests a unique, potentially unfruitful interaction between the PCA-transformed data and the default SVM's linear kernel. Indeed, transformed features generated by PCA can sometimes be incompatible with certain kernel choices, leading to computational inefficiencies.

As discussed earlier in Section 3.2 and depicted in Figure 2, the similarity of data properties in the PCA-reduced feature space may impede SVM's ability to extract meaningful patterns, resulting in inefficient learning. This hypothesis is further supported by the model's modest accuracy of 77%.

While the base SVM model achieved an impressive 92% accuracy, 95% precision, and 92% F1 score, demonstrating strong predictive power, the UMAP-SVM combination emerged as the champion in terms of generalization. Despite a slightly lower current performance (82% accuracy, 79% precision, 82% F1 score), UMAP exhibited remarkable potential for exceptional generalization. This potential is evidenced by its remarkably consistent learning curve, closely mirroring the SVM's performance on unseen data. This suggests UMAP's ability to effectively capture underlying data structures and transfer knowledge to unseen data with greater efficacy than other DR techniques.

It is important, however, to acknowledge the trade-off involved: initial accuracy may be slightly sacrificed for superior long-term generalization. Therefore, UMAP stands out as a compelling option for applications where prioritizing future adaptation and robustness to new data outweighs slightly lower initial performance.

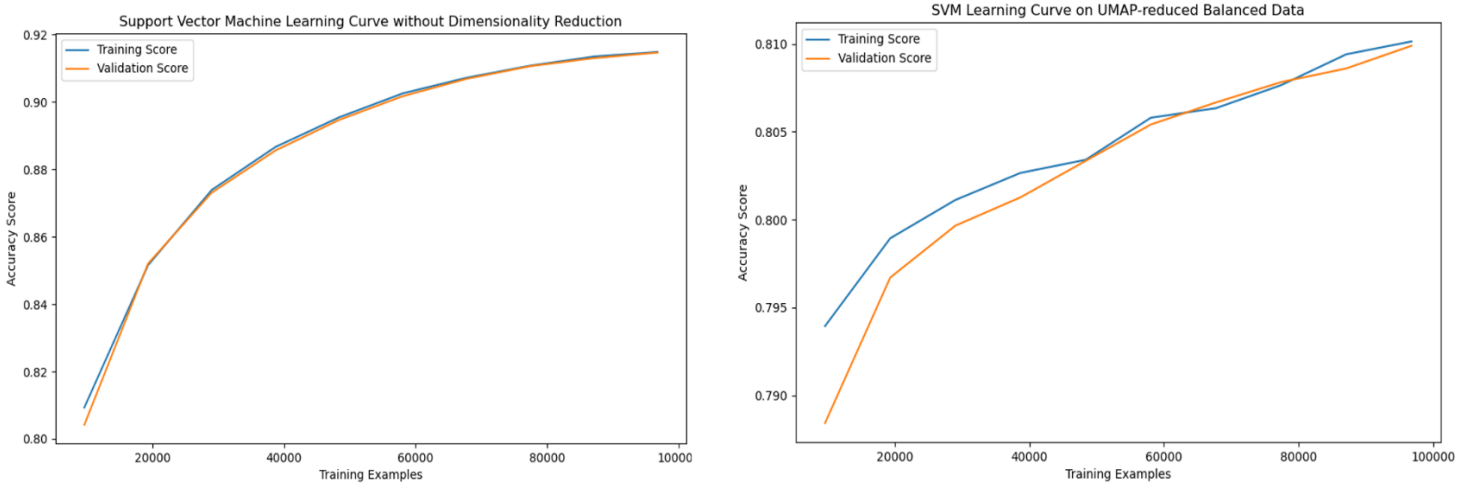

Figure 6: SVM Learning Curve before(left) and after Dimensionality Reduction with UMAP (right). Training(blue) and validation (Yellow) curves show stable convergence and generalization to unseen data

In the original set of classifiers on the balanced data, LR achieved an accuracy of 0.94, which is already a strong performance. However, when LDA is applied, LR maintains this high accuracy, with a score of 0.92. This suggests that LDA, in the context of LR, is effective in preserving the discriminative information needed for accurate predictions.

On the contrary, when LR is combined with PCA or UMAP, there is a more significant drop in accuracy. LR-PCA has an accuracy of 0.74, and LR-UMAP further decreases to 0.70. This indicates that the

reduction in dimensionality through PCA or UMAP may not be as well-suited for LR, possibly leading to loss of crucial information.

The notable performance of LR on LDA-reduced data could be attributed to the nature of LDA, which maximizes class separability. LR, being a linear model, benefits more from the transformed space created by LDA.

The stark decrease in performance for LR, GB, and SVM when combined with PCA implies that the information retained after dimensionality reduction might not be conducive to the decision boundaries or characteristics of these classifiers. Random Forest, being an ensemble method, seems less affected, possibly due to its inherent ability to handle diverse features.

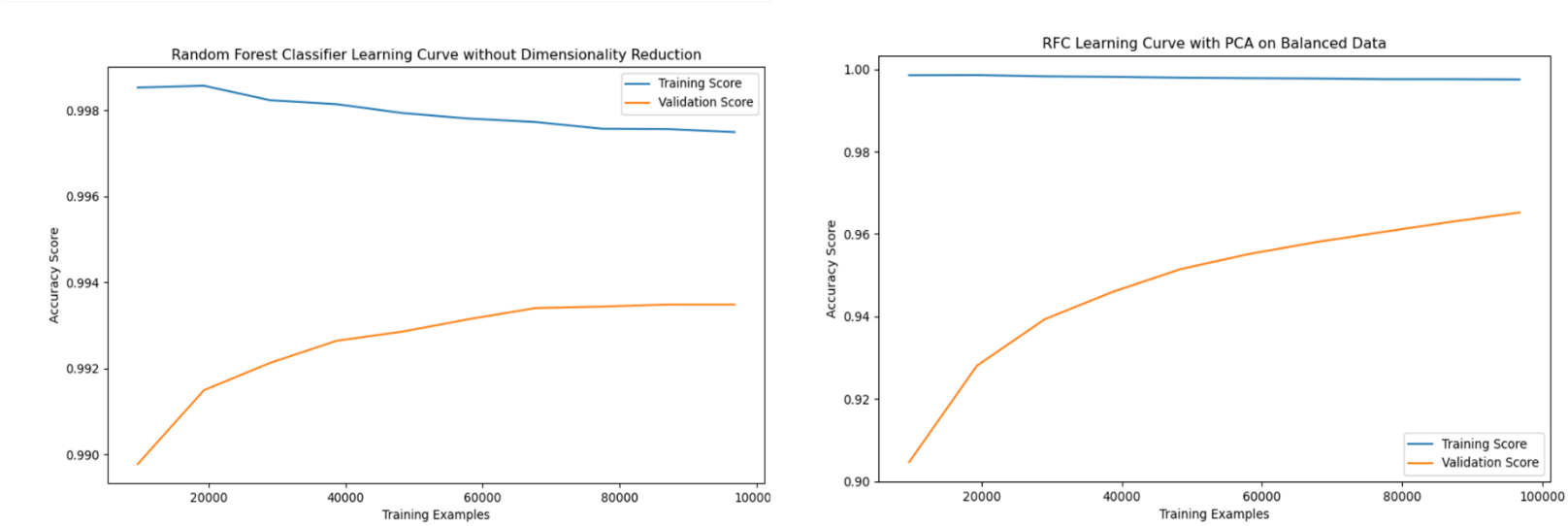

Figure 7: Random forest learning curves before (left) and after DR with PCA (right). With similar high accuracy performance, curves show similar convergence as the validation curve (yellow) mitigates initial overfitting indicating PCA's compatibility with RFC

ROC/AUC score analysis

- Overall strong AUC scores: Most models achieved AUC values above 0.90, indicating good discrimination ability between classes.
- In most cases, models with high accuracy also had high AUC scores, suggesting a strong correlation between both metrics.
- LDA as a generally consistent DR technique: It maintained or slightly improved AUC scores for most algorithms, except for RF.
- PCA's mixed impact on AUC: Its effects on AUC varied across algorithms, with notable drops for LR and SVM.

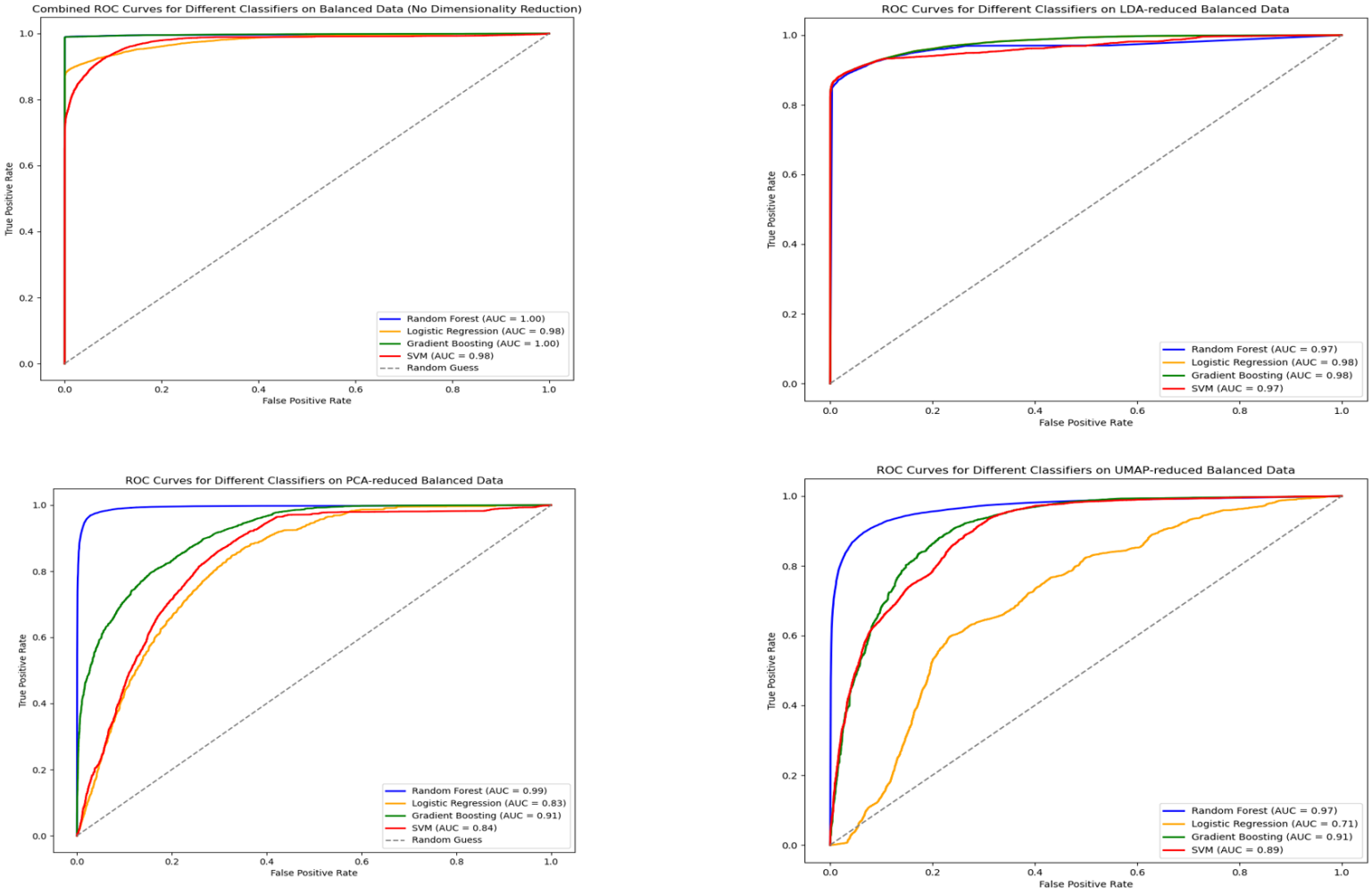

Figure 8: Multiple ROC curves visualize the performance of different machine learning classifiers (indicated by different colours), both in their original form and after applying various dimensionality reduction techniques (PCA, LDA, and UMAP). Each plot features a dashed grey line representing random guessing for reference. The Area Under the Curve (AUC) scores, quantifying model discrimination ability

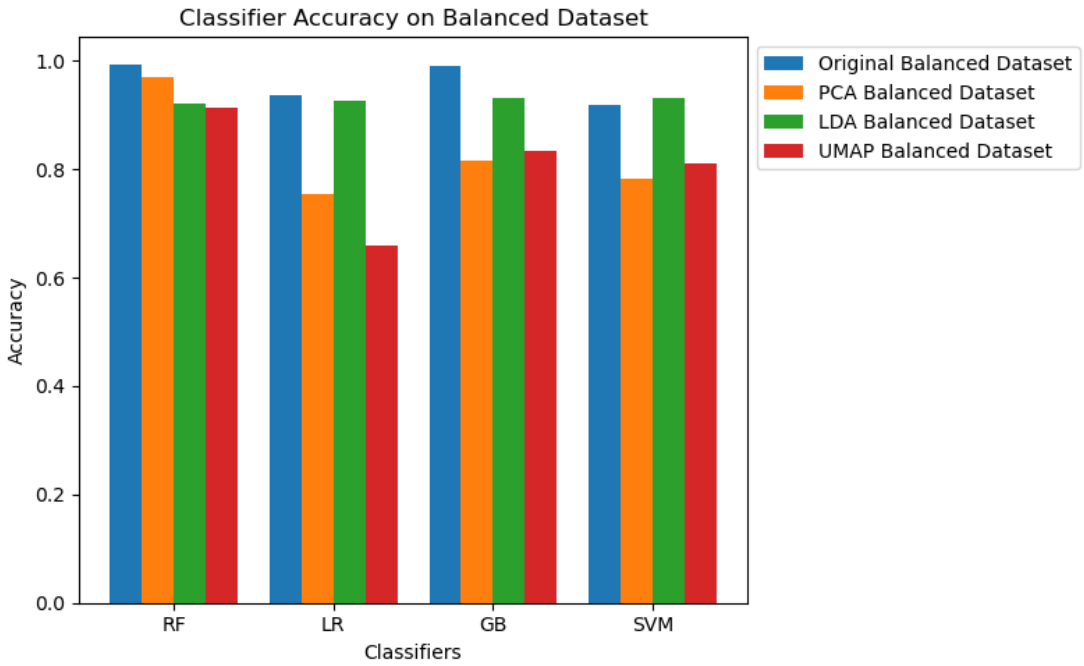

Figure 9: Bar plot comparing the classification accuracy of different machine learning algorithms, both in their original form and after applying various dimensionality reduction techniques (PCA, LDA, and UMAP).

5. Discussion

The influence of dimensionality reduction (DR) techniques on the performance of various machine learning algorithms for classification tasks was comprehensively investigated in this research. Both traditional techniques (PCA, LDA) and a non-linear approach (UMAP) were employed, providing valuable insights into their suitability and effectiveness across different models and data scenarios.

The choice of dimensionality reduction (DR) technique proved to be a crucial factor in optimizing model performance, particularly when paired with specific algorithms. This aligns with the findings of Xie and Zhang (2015), who demonstrated the effectiveness of LDA and SVM in fault diagnosis of rolling bearing systems. In this study, LDA's ability to maximize class separation in the reduced-dimensional space exhibited a distinct advantage for SVM when handling imbalanced data. This resulted in remarkable accuracy (0.98), precision (1.0), and F1 score (0.88), showcasing LDA's capacity to enhance SVM's ability to capture underlying data structures and learn distinctive characteristics in lower-dimensional representations.

The suitability of LDA and SVM stems from their complementary strengths. LDA's focus on maximizing class separation aligns well with SVM's goal of constructing a hyperplane that effectively separates classes (George. et al, 2012). This synergy enables the model to better identify and distinguish between classes, even in imbalanced datasets where one class is underrepresented.

Although dimensionality reduction (DR) techniques generally led to a slight decrease in accuracy, they often demonstrated a significant improvement in model generalization, particularly for the Gradient Boosting classifier on balanced datasets. This aligns with observations by Friedman (2002), who noted

Gradient Boosting's tendency to prioritize model complexity over raw accuracy, potentially enhancing generalization. The research findings suggest a delicate trade-off between predictive power on known data and the ability to generalize to unseen data, highlighting the importance of considering both aspects when evaluating model performance.

Prior work by Maaten et al. (2007) concluded that non-linear DR techniques like ISOMAP weren't readily outperforming PCA, our findings suggest UMAP may now present a viable alternative. Although the average accuracy between UMAP and PCA was similar (81%), UMAP demonstrated superior convergence and mitigated overfitting in Gradient Boosting and SVM models. This aligns with UMAP's ability to capture complex non-linear data structures, potentially fulfilling the future potential envisioned by Maaten et al. for techniques that optimize well and outperform traditional linear methods.

However, it is important to acknowledge some key limitations that influence the generalizability and applicability of these findings;

Exploring ensembles of different DR techniques tailored to specific algorithms and data types could potentially leverage the strengths of each method while mitigating their weaknesses. This could enhance overall performance, generalization, and interpretability.

UMAP's superior generalization with SVM, however impressive, comes at the cost of accuracy and interpretability. Understanding the reasons behind model predictions becomes much more complex with non-linear dimensionality reduction techniques, which could pose challenges in real-world applications where transparency is crucial.

Limited scope of data: Exploring diverse datasets across different domains and sizes can reveal how DR techniques and their interactions with algorithms vary under different conditions. Secondly, from these findings, if the original feature space is already small, the benefits of dimensionality reduction might be minimal, and the overhead could outweigh any gains.

Addressing these limitations through future research will be crucial for solidifying the practical use of DR techniques in various machine learning applications.

6. Conclusion

This research has comprehensively explored the interplay between dimensionality reduction (DR) techniques and diverse machine learning algorithms for classification tasks, unlocking valuable insights into their compatibility and effectiveness across different models and data scenarios. Notably, the study reveals a symbiotic relationship between LDA and SVM, highlighting how LDA's ability to maximize class separation empowers SVM to excel in handling imbalanced data. Moreover, it challenges previous assumptions by suggesting UMAP as a viable alternative to traditional non-linear DR methods. By addressing the limitations outlined in the discussion, such as exploring ensemble DR techniques and expanding the data scope, we can pave the way for further exploration and harness the transformative potential of DR techniques in various machine learning applications

References

Abdul Salam, M., Taher, A., Elgendy, M., & Mohamed, K. (2021). The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. International Journal of Advanced Computer Science and Applications, 12. https://doi.org/10.14569/IJACSA.2021.0120480.

Akhiat, Y., Manzali, Y., Chahhou, M., and Zinedine, A. (2021). A new noisy random forest based method for feature selection. Cybernetics and Information Technologies, 21(2), pp.10-28.

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (Eds.), Image and Signal Processing. ICISP 2020. (Vol. 12119, Lecture Notes in Computer Science). Springer, Cham. https://doi.org/10.1007/978-3-030-51935-3_34.

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, Isomap, LE, ICA, t-SNE). Computer Science Review, 40, 100378.

Anzanello, M., & Fogliatto, F. (2011). Learning curve models and applications: literature review and research directions. International Journal of Industrial Ergonomics, 41, 573-583. https://doi.org/10.1016/j.ergon.2011.05.001.

Biggs, M., Wang, Y., Soni, N., Priya, S., Bathla, G., & Canahuate, G. (2023). Evaluating Autoencoders for Dimensionality Reduction of MRI-derived Radiomics and Classification of Malignant Brain Tumors. In Proceedings of the 35th International Conference on Scientific and Statistical Database Management (SSDBM '23). Association for Computing Machinery, New York, NY, USA, Article 1, 1–11. https://doi.org/10.1145/3603719.3603737.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., & Song, M. (2003). Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research, 3(Mar), 1229-1243.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321-357.

Chen, H., Liu, H., Feng, H., Fu, H., Cai, W., Shao, X., & Chipot, C. (2022). MLCV: Bridging Machine-Learning-Based Dimensionality Reduction and Free-Energy Calculation. Journal of Chemical Information and Modeling, 62(1), 1-8. https://doi.org/10.1021/acs.jcim.1c01010.

Desyani, T., Saifudin, A., & Yulianti, Y. (2020). Feature Selection Based on Naive Bayes for Caesarean Section Prediction. IOP Conference Series: Materials Science and Engineering, 879, 012091. https://doi.org/10.1088/1757-899X/879/1/012091.

Fourrier, Q., & Aloise, D. (2019) "Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods," in 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, pp. 211-214. DOI: 10.1109/AIKE.2019.00044.

George, A., & Vidyapeetham, A. (2012). Anomaly detection based on machine learning: Dimensionality reduction using PCA and classification using SVM. International Journal of Computer Applications, 47(21), 5-8.

Jamal, A., Handayani, A., Septiandri, A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 9(3), 192. https://doi.org/10.24843/LKJITI.2018.v09.i03.p08.

Kabir, M. F., Chen, T., Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthcare Analytics, 3(100125). https://doi.org/10.1016/j.health.2022.10012.

Kasun, L. L. C., Yang, Y., Huang, G. B., & Zhang, Z. (2016). Dimension reduction with extreme learning machine. IEEE Transactions on Image Processing, 25(8), 3906-3918.

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In 2014 Science and Information Conference, (pp. 372-378). IEEE.

Kicska, G., & Kiss, A. (2021). Comparing swarm intelligence algorithms for dimension reduction in machine learning. Big Data and Cognitive Computing, 5(3), 36.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In Data Democracy (pp. 83-106). Academic Press.

Lan, T. (2011). Feature extraction, feature selection, and dimensionality reduction techniques for brain-computer interface. Doctor of Philosophy in Electrical Engineering examined and approved thesis. Oregon Health & Science University, OHSU Digital Commons, Scholar Archive, Paper 706.

Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426. Retrieved from https://doi.org/10.48550/arXiv.1802.03426.

Mendez, M. A. (2023). Linear and nonlinear dimensionality reduction from fluid mechanics to machine learning. Measurement Science and Technology, 34(4), 042001. https://doi.org/10.1088/1361-6501/acaffe.

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA) (pp. 18-20). IEEE.

Perlich, C. (2011). Learning Curves in Machine Learning. https://doi.org/10.1007/978-0-387-30164-8_452.

Pham, H. T., Awange, J., & Kuhn, M. (2022). Evaluation of three feature dimension reduction techniques for machine learning-based crop yield prediction models. Sensors, 22(17), 6609. https://doi.org/10.3390/s22176609.

Q. Fournier & D. Aloise (2019) "Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods," in 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, pp. 211-214. DOI: 10.1109/AIKE.2019.00044.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access, 8, 54776-54788.

Sachin, D. (2015). Dimensionality reduction and classification through PCA and LDA. International Journal of Computer Applications, 122(17).

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II (pp. 313-325).

Sharma, A., & Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In RACS '12: Proceedings of the 2012 ACM Research in Applied Computation Symposium. https://doi.org/10.1145/2401603.2401605.

Tang, E. K., Suganthan, P. N., Yao, X., & Qin, A. K. (2005). Linear dimensionality reduction using relevance weighted LDA. Pattern Recognition, 38(4), 485–493. https://doi.org/10.1016/j.patcog.2004.09.005.

van der Maaten, L., Postma, E., & Herik, H. (2007). Dimensionality Reduction: A Comparative Review. Journal of Machine Learning Research - JMLR, 10.

Wan, Y., Li, T., Wang, P., Duan, S., Zhang, C., & Li, N. (2021). Robust and efficient classification for underground metal target using dimensionality reduction and machine learning. IEEE Access, 9, 7384-7401.

Wan, Z., Xu, Y., & Šavija, B. (2021). On the Use of Machine Learning Models for Prediction of Compressive Strength of Concrete: Influence of Dimensionality Reduction on the Model Performance. Materials, 14(4), 713. https://doi.org/10.3390/ma14040713.

Wang, W., Huang, Y., Wang, Y., and Wang, L. (2014) "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2014.

Yang, Y., Sun, H., Zhang, Y., Zhang, T., Gong, J., Wei, Y., Duan, Y. G., Shu, M., Yang, Y., Wu, D., & Yu, D. (2021). Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Reports, 36(4).

Zhao, M., Tang, Y., Kim, H., & Hasegawa, K. (2018). Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. Cancer Informatics, 17. https://doi.org/10.1177/1176935118810215.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561.