Methodology for the comparative analysis of PCA and Autoencoders in dimensionality reduction: impact on classification accuracy and computational efficiency

Saša Mitrović, Neven Vrček
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{smitrovic, nvrcek}@foi.hr

Abstract. *This paper discusses an adapted CRISP-DM methodological framework for an automated comparison of the performance of dimensionality reduction methods — PCA, LDA, Factor Analysis and AE, t-SNE, Kernel PCA — in the context of multiclass and binary classification tasks using different datasets. The use of the adapted CRISP-DM methodology ensures an automated and systematic analysis in all phases of the CRISP-DM process, including data preparation, modelling and evaluation. The automation is achieved by dynamically determining the number of components according to the dimensional structure of the datasets by integrating dimensionality reduction methods and machine learning classification methods. The results within this study indicate that PCA and Autoencoder are effective dimensionality reduction techniques in linear and nonlinear space, respectively. Furthermore, the results show that, on average, PCA achieves higher classification accuracy with significantly lower computational effort, while AE shows advantages in higher classification accuracy of the nonlinear domain with the trade-off in low computational efficiency. The proposed approach enables a replicable, modular and computationally efficient evaluation of algorithms for analysing high-dimensional data and thus contributes to the improvement of automated systems in the field of intelligent data analysis. A Python implementation of the adapted CRISP-DM methodology for the automated comparison of the performance of two dimensionality reduction methods can be found at the following GitHub link: https://github.com/smitroviefos/amf_dim_red_paper2025* .

Keywords. CRISP-DM, PCA, AutoEncoders, Dimensionality reduction, classification performance, computational efficiency

1 Introduction

Given the growing importance of data-intensive intelligent systems in today's environment, especially in the field of analysing and processing big data, the high dimensionality of input features poses a significant challenge to the efficiency and interpretability of machine learning algorithms.(Baratchi et al., 2024) Dimensionality reduction as a key stage of data processing reduces model complexity, speeds up execution and improves the robustness of classification and regression systems. The most common techniques include principal component analysis (PCA)(Bartal et al., 2019; Cohen, 2017; Geron, 2019; Jolliffe, 2002; Van Der Maaten et al., 2009) and autoencoders (AE)(Chandra, 2024; Cohen, 2017; Goodfellow et al., 2016), which approach the problem from two different perspectives — linear and nonlinear.

The PCA method represents a projection of the data into a space with lower dimensionality, preserving the highest possible variance, while autoencoders, which are based on multilayer neural networks, allow the modelling of non-linear relationships between features, thus ensuring greater informativeness, but require more computational resources. Although there is scientific literature comparing these approaches applied to different tasks (Fournier & Aloise, 2019; Mayur Prakashrao Gore, 2024), a systematic analysis combining quantitative and qualitative metrics within a single methodological framework is still relatively limited in the SCOPUS database to the best of the authors' knowledge based on the literature review presented below. Therefore, this paper aims to explore the advantages and disadvantages of PCA and AE as representative techniques of linear and nonlinear dimensionality reduction, respectively, within the context of common machine learning classification methods. The primary objective is to analyse the strengths, limitations, and applicability of these methods across different data domains by embedding them into a fully automated analytical pipeline based on an adapted CRISP-DM (Cross-Industry Standard Process for Data Mining) methodological framework. Additionally, research paper compares linear methods (PCA, Linear Discriminant Analysis – LDA, and Factor Analysis) separately from nonlinear methods (Autoencoders, t-Distributed Stochastic Neighbour Embedding – t-SNE, and Kernel PCA), the study incorporates an extended evaluation that organises

dimensionality reduction algorithms into these two conceptual categories – linear and nonlinear. While the primary analytical emphasis remains on PCA and AE due to their widespread use, business domain-independent datasets and widely adopted classifiers are utilised to evaluate model performance through multiple quantitative metrics, including accuracy, precision, recall, F1 score, and ROC/AUC. This approach enables the assessment of generalisation capability, information preservation, and execution efficiency of each method, with the ultimate goal of contributing to a more structured and comparative understanding of dimensionality reduction techniques in automated machine learning pipelines.

2 Literature review

In analysing the existing scientific literature, the present paper focuses on the research gap identified by a number of authors discussing dimensionality reduction, PCA and AE methods. Specifically, Mendes Junior et al. (2020) point out that in previous research they found that the comparison of feature selection and dimensionality reduction methods was not sufficiently systematic and quantitatively sound. In particular, Mendes Junior et al. (2020) found that the application of dimensionality reduction techniques after feature selection often did not lead to statistically significant differences in the accuracy of machine learning classification methods. Furthermore, Mendes Junior et al. (2020) point out that the research gap also includes a limited number of comparative analyses of dimensionality reduction techniques and the lack of comprehensive evaluations in different problem domains, indicating the need for more methodologically rigorous approaches that cover a broader range of data, models and performance metrics.(Mendes Junior et al., 2020) Vantuch et al. (2016) point out that future research should focus on the problem of reconstruction methods that inadequately reconstruct the original datasets after applying dimensionality reduction, which can lead to information loss. Similarly, Vantuch et al. (2016) emphasise that the analysis of information loss and uncertainties that occur during transformation is insufficiently discussed in the literature, although these aspects can have a significant impact on the reliability of the model. Vantuch et al. (2016) therefore point out that there is a need for empirical comparisons that quantitatively evaluate the advantages of different dimensionality reduction algorithms in terms of maintaining data representativeness and the interpretability of the results.(Vantuch et al., 2016) In their research, Ghobadi et al. (2023) point out the need to adjust hyperparameters in dimensionality reduction methods to achieve optimal results in the unique data domain contexts. Since different datasets can have markedly heterogenous features — in terms of noise levels, correlations between features or structural complexity — fixed and manual approaches without automation of parameter adaptation often lead to suboptimal transformations. Ghobadi et al. (2023) therefore emphasise the importance of dynamic and context-sensitive parameter adjustments to increase model accuracy and maintain the semantic representation of the data in low-dimensional space.(Ghobadi et al., 2023)

In the present work, the above shortcoming is addressed using a systematic and methodologically structured approach that integrates a dynamic selection of the number of components, automated model evaluation and statistical validation of performance differences, thus contributing to the understanding of the effectiveness and applicability of modern dimensionality reduction techniques in scientific and applied contexts.

3 Methodology

In order to conduct a systematic analysis, the present authors used an approach based on the adaptation of the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining), which provides a comprehensive framework for conducting data analyses in six phases: problem understanding, data understanding, data preparation, modelling, evaluation, and implementation.(Bratkovsky, 2024; Chapman, 2000; Costa, 2022; Shearer, 2000)

3.1 Understanding the research problem

The main objective of this study is to compare two dimensionality reduction methods: as a linear method and autoencoders (AE) as a nonlinear, neural method. Dimensionality reduction is a crucial step in the processing and analysis of high-dimensional data, i.e. in the context of this work for classification tasks where the high complexity of the input space can negatively affect the generalisation capability of machine learning models. The comparison of PCA and AE is made in terms of classification accuracy, execution time, computational efficiency and preservation of the semantic structure of the data in the latent space. In addition, other dimensionality reduction methods are used, structured into two conceptual groups - linear and nonlinear. The linear methods include PCA, Linear Discriminant Analysis (LDA), and Factor Analysis (FA), all of which operate under the assumption of linear transformations and rely on variance maximisation or class separability criteria (Pedregosa et al., 2011). In contrast, the nonlinear methods comprise Autoencoders, t-Distributed Stochastic Neighbor Embedding (t-SNE), and Kernel PCA, each capable of modelling non-linear manifolds and capturing more complex intrinsic data structures through techniques such as deep representation learning or kernel-based transformations(Chandra, 2024; Cohen, 2017; Schölkopf et al., 1997). While the

core analytical focus remains on PCA and AE due to their conceptual contrast and practical relevance in both academic research papers and various business domain settings, the inclusion of other methods enables a more comprehensive and structured benchmarking process. The comparison is made in terms of classification accuracy, execution time, computational efficiency and preservation of the semantic structure of the data in the latent space.

3.2 Data understanding

The Phoneme (OpenML ID: 1489), Vehicle Silhouettes (OpenML ID: 54), Blood Transfusion Service Center (OpenML ID: 1464) and the HighDim (Gina agnostic, OpenML ID: 40978) datasets from the OpenML repository were used in this study as they represent different application domains and different feature structures, thus enabling an evaluation of the generalisation capability of dimensionality reduction algorithms. The Phoneme dataset consists of 5,404 instances and 5 acoustic features, containing the sound features of speech phonemes and is primarily used for speech recognition tasks; Vehicle dataset includes 846 instances with 18 geometrical features, containing geometric attributes of vehicles classified by type; Blood Transfusion Service Center dataset consists of 748 instances and 4 features containing demographic and behavioural data of blood donors with the aim of predicting their propensity to re-donate, while HighDim dataset is a high-dimensional binary classification handwritten digit as image recognition dataset with 1,400 instances and 970 features, constructed to evaluate algorithmic scalability and behaviour in high-dimensional spaces. This diversity provides a relevant experimental context to test the effectiveness of the methods within an automated, adapted CRISP-DM methodological framework.

3.3 Data preparation

The data were first normalised to the interval 0 to 1 to reduce numerical differences between features and ensure model stability. The data were stratified to divide them into training and test datasets in a ratio of 80:20. As dimensionality reduction methods can be sensitive to the size and distribution of the data, special care is taken to maintain class balance when splitting a dataset with an extremely high-class imbalance. In all cases - PCA, LDA, Factor Analysis and AE, t-SNE, Kernel PCA - the number of components and the dimension of the layer are dynamically determined as a minimum between the preset threshold (in the Python code the threshold is set to 10), the number of features and the number of instances in the dataset. Such an approach ensures the methodological validity and flexibility of the algorithms when applied to heterogenous datasets, while maintaining the structural consistency of the model. This enables the automated application of dimensionality reduction techniques in the CRISP-DM modelling and evaluation phase, regardless of the unique data domain context.

3.4 Modelling

Following dimensionality reduction methods were used in the modelling phase: (1) PCA, LDA, and Factor Analysis are the linear techniques methods. Methods were implemented from the scikit-learn library (Pedregosa et al., 2011).(Bartal et al., 2019; Cohen, 2017; Geron, 2019; Jolliffe, 2002; Van Der Maaten et al., 2009) (2) The autoencoder is an neural network with a hidden layer of dimension 64 implemented in the TensorFlow (Abadi et al., 2016) environment. The mean squared error (MSE)(Santosh et al., 2022), is used for the error function, while the Adam algorithm with an initial learning parameter of 0.001 is used for optimisation.(Chandra, 2024; Cohen, 2017; Goodfellow et al., 2016). Other nonlinear techniques such as t-SNE(van der Maaten & Hinton, 2008) and Kernel PCA (Schölkopf et al., 1997) are also used in the modelling phase.

After reduction, identical machine learning classification methods LogisticRegression (s.-l. developers, 2024a), RandomForest(s.-l. developers, 2024b), SVM(s.-l. developers, 2024c) and XGBoost(x. developers, 2024), are trained on the obtained latent representations to isolate the effects of the reduction method on the classification performance.

3.5 Evaluation

The evaluation of the model was based on several criteria: classification accuracy, execution time and visualisation of the reduced dimensionality using bar charts. The qualitative analysis of the visual representations provides information about the separability of the classes in the latent space. The quantitative results were analysed using McNemar's statistical test for dependent samples of binary classification to determine the significance of the differences between the accuracies of the machine learning methods (Perktold et al., 2025b). The quantitative results for dependent samples of multiclassification were analysed using Cochran's Q test (Perktold et al., 2025a). In addition, the data reconstruction ability of the autoencoder was compared to evaluate the information loss.

3.6 Implementation

The entire experiment was implemented in the Python programming language. The code is available at the GitHub link and uses the externally referenced libraries scikit-learn [17], tensorflow (Abadi et al., 2016), matplotlib(Barrett et al., 2005) and numpy(Harris et al., 2020). The programme code is modular and reproducible, and it is possible to extend the analysis to additional datasets and dimensionality reduction methods such as UMAP (Healy & McInnes, 2024).

The results show that PCA retains its advantage in terms of computational efficiency, while AE shows superiority in preserving complex structures and higher classification accuracy at lower dimensions. The conclusions obtained make a scientific contribution to the understanding of the trade-off between linearity and non-linearity in dimensionality reduction in the context of application to different domains of business data.

4 Results and discussion

The experiment uses an adapted CRISP-DM methodology to automate the process of intelligent data analysis and compares the effects of linear and nonlinear dimensionality reduction methods – PCA, LDA, Factor Analysis and AE, t-SNE, Kernel PCA – in combination with four classification methods LogisticRegression (s.-l. developers, 2024a), RandomForest(s.-l. developers, 2024b), SVM(s.-l. developers, 2024c) and XGBoost(x. developers, 2024) in the modelling phase. The evaluation was carried out with heterogenous datasets that differ in the number of classes, the distribution of features and the semantic domain. The aim was to assess how the choice of dimensionality reduction method affects the classification efficiency and computational complexity of the model.

Table 1. Accuracy per Datasets

DR Method	LogReg	RF	SVM	XGBoost
Dataset **Blood**				
AE	0.7467	**0.7667**	**0.7667**	0.74
FA	0.7467	0.7533	**0.76**	0.7333
KPCA	0.76	0.72	**0.7667**	0.7333
LDA	**0.7533**	0.6533	0.7467	0.74
PCA	0.7467	0.7467	**0.7667**	0.7267
t-SNE	**0.7667**	0.7333	0.7733	0.6867
Dataset **HighDim**				
AE	0.7522	0.7594	**0.7695**	0.7392
FA	0.7954	0.7896	**0.8112**	0.7954
KPCA	0.7896	0.7997	**0.8228**	0.7925
LDA	0.8228	0.8069	**0.8228**	0.8156
PCA	0.8055	**0.8228**	**0.8228**	0.8084
t-SNE	0.6643	**0.7752**	0.745	0.7233
Dataset **Phoneme**				
AE	0.7336	0.8705	0.8437	**0.8751**
FA	0.741	**0.8668**	0.7974	0.8548
KPCA	0.7512	**0.8742**	0.8039	0.8511
LDA	0.7373	0.7391	**0.7623**	0.7586
PCA	0.7364	**0.8788**	0.8326	0.864
t-SNE	0.7188	**0.8696**	0.778	0.8686
Dataset **Vehicle**				
AE	0.6588	0.6059	0.6765	**0.7059**
FA	0.7765	0.7588	**0.7882**	0.7765
KPCA	0.6412	0.6647	**0.6765**	0.6706
LDA	**0.5941**	0.5118	0.5882	0.5529
PCA	0.6765	**0.7353**	0.7176	0.7059
t-SNE	0.4882	**0.6882**	0.6	0.6353

The results, as shown in Table 1, show several important findings. Within the group of linear dimensionality reduction methods, accuracy results have shown that Principal Component Analysis (PCA) is the most consistent and effective dimensionality reduction technique. In datasets Phoneme and HighDim, PCA yielded very high classification accuracy, particularly when used in combination with RandomForest and XGBoost classifiers. Furthermore, PCA demonstrated extremely low computational complexity, making it highly suitable for integration into automated analytical pipelines within the CRISP-DM methodological framework. In contrast, Linear Discriminant Analysis (LDA) and Factor Analysis (FA) exhibited lower robustness across most experimental scenarios. Although they achieved satisfactory results on the HighDim dataset, their average classification accuracy was inferior to that achieved by PCA. Method LDA has limitations due to its restrictions in handling multiclass problems and the requirement for class labels during the reduction phase, which constrains its flexibility of application in automated systems aligned with the CRISP-DM methodology.

Within the group of nonlinear methods, accuracy results have shown that Autoencoder is the most consistent and effective technique to dimensionality reduction, achieving high classification accuracy and consistent performance across all datasets. While the Autoencoder provides high predictive accuracy, such performance is accompanied by longer reduction times compared to linear methods. This trade-off is justified by its ability to preserve nonlinear relationships in the data, thus enabling a better representation of complex structures in the latent space (Chandra, 2024). Given its accuracy, the Autoencoder positions itself as an exceptional tool for automating data processing pipeline within the CRISP-DM cycle, particularly when dealing with datasets exhibiting nonlinear distributions. The t-SNE method showed acceptable accuracy on the Phoneme dataset but also demonstrated high variability and substantial computational overhead. It is important to emphasize that t-SNE was limited to three components due to scalability constraints, attempts with higher dimensions failed to complete even after several hours of execution. Despite its usefulness for visualization, the application of t-SNE in classification-oriented workflows, such as those structured by CRISP-DM, remains methodologically limited. KernelPCA, as a nonlinear extension of PCA, produced solid results, particularly on the HighDim dataset. Its capacity to preserve complex relationships in the data without excessive computational burden makes it a

methodologically balanced choice, although it does not surpass the Autoencoder in overall effectiveness.

The accuracy results within this study indicate that PCA and Autoencoder are effective strategies for dimensionality reduction in linear and nonlinear space, respectively. Method PCA offers an efficient and fast linear transformation suitable for a wide range of problems, while the Autoencoder provides superior quality for complex and nonlinear datasets. Based on the experimental findings and within the context of developing automated systems aligned with the CRISP-DM methodology, the authors of this study recommend the combined use of PCA and Autoencoders, taking into consideration the trade-off between computational efficiency and the quality of data representation in lower-dimensional space.

Table 2. Reduction execution time per Datasets

DR Method	LogReg	RF	SVM	XGBoost
Dataset **Blood**				
AE	4.4284	4.4284	4.4284	4.4284
FA	0.004	0.004	0.004	0.004
KPCA	0.0198	0.0198	0.0198	0.0198
LDA	0.001	0.001	0.001	0.001
PCA	0.0011	0.0011	0.0011	0.0011
t-SNE	5.8834	5.8834	5.8834	5.8834
Dataset **HighDim**				
AE	9.1207	9.1207	9.1207	9.1207
FA	0.1198	0.1198	0.1198	0.1198
KPCA	0.9979	0.9979	0.9979	0.9979
LDA	0.1953	0.1953	0.1953	0.1953
PCA	0.0346	0.0346	0.0346	0.0346
t-SNE	58.4865	58.487	58.487	58.4865
Dataset **Phoneme**				
AE	10.0798	10.08	10.08	10.0798
FA	0.0057	0.0057	0.0057	0.0057
KPCA	1.1999	1.1999	1.1999	1.1999
LDA	0.002	0.002	0.002	0.002
PCA	0.0303	0.0303	0.0303	0.0303
t-SNE	67.0619	67.062	67.062	67.0619
Dataset **Vehicle**				
AE	4.6146	4.6146	4.6146	4.6146
FA	0.2431	0.2431	0.2431	0.2431
KPCA	0.0439	0.0439	0.0439	0.0439
LDA	0.002	0.002	0.002	0.002
PCA	0.002	0.002	0.002	0.002
t-SNE	9.3425	9.3425	9.3425	9.3425

The reduction execution time results, as shown in Table 2, for various dimensionality reduction techniques across four experimental datasets (Blood, HighDim, Phoneme, and Vehicle) provide insight into the computational complexity of each method, which is a crucial consideration when designing automated systems based on the CRISP-DM methodological framework. The linear method PCA (Principal Component Analysis) demonstrates consistently low computational complexity across all datasets (with execution times ranging from approximately 0.001 to 0.03 seconds), making it a methodologically and computationally optimal option for dimensionality reduction in scenarios involving large data volumes and the need for rapid processing. Similarly as PCA, LDA (Linear Discriminant Analysis) and FA (Factor Analysis) also achieve very low reduction execution times, however, it should be noted that their applicability depends on the characteristics of the dataset, such as the classification type problem number of classes and the linearity of relationships among features.

Within the group of nonlinear methods the method Kernel PCA (KPCA) requires slightly more computational time (from 0.04 seconds on the Vehicle dataset to 1.20 seconds on the Phoneme dataset), but still remains within acceptable thresholds for integration into automated data pipelines, particularly when there is a need to preserve nonlinear relationships without excessive computational cost.

The Autoencoder (AE) method consistently shows slower execution times across all datasets, ranging from approximately 4.4 seconds (Blood and Vehicle) to over 10 seconds (Phoneme), and nearly 9.1 seconds on the high-dimensional HighDim dataset. This level of low computational efficiency reflects the fact that AE involves multiple data passes during training, including parameter optimization and input reconstruction, making it more suitable for scenarios where latency is not the primary constraint, but where preserving complex nonlinear data structures is critical.

The most computationally inefficient method is t-SNE (t-distributed Stochastic Neighbor Embedding) with only three components, with reduction times exceeding several seconds in all datasets, and reaching up to 67 and 58 seconds in the Phoneme and HighDim datasets, respectively.

The reduction execution runtime per Datasets results emphasizes linear methods (PCA, LDA, FA) as computationally efficient dimensionality reduction methods and nonlinear methods (AE, KPCA, t-SNE) as more resource-intensive dimensionality reduction methods. Autoencoders are better in preserving the semantic structure of the data, with the cost in their execution time since it is considerably longer compared to PCA. This implies the need to balance accuracy and computational cost, preferably with objective function, when designing an optimal automated machine learning CRISP-DM-based process. Method PCA should be applied in time-sensitive applications, while AE and KPCA may be more suitable for applications involving complex nonlinearities and sufficient computational resources.

The results comparing linear and nonlinear methods by analysing the average values of performance metrics—accuracy, precision, recall, F1 score, and the area under the ROC curve (ROC/AUC)—it is observed that linear methods (PCA,

LDA, and FA) consistently outperform nonlinear approaches (Autoencoder, t-SNE, and KernelPCA) across all evaluated metrics.

Table 3. Group Averages: Linear and Nonlinear methods

Metrics	Group	
	Linear	Nonlinear
Accuracy	0.7572	0.7426
Precision	0.747	0.7296
Recall	0.7572	0.7426
F1 Score	0.7467	0.73
ROC/AUC	0.8909	0.8755

Linear methods achieved an average accuracy of 0.7572, whereas nonlinear methods yielded a slightly lower value of 0.7426. This pattern is consistent across other metrics as well: precision (0.747 vs. 0.7296), recall (0.7572 vs. 0.7426), F1 score (0.7467 vs. 0.73), and ROC/AUC (0.8909 vs. 0.8755). These results indicate a higher level of consistency and robustness of the linear approach, especially in structured classification tasks where relationships among features are predominantly linear or can be effectively approximated through linear transformations.

From a methodological perspective, linear methods benefit from low computational complexity, stable behavior across datasets of varying sizes, and transparent interpretability of the transformations. Although nonlinear methods are better in preserving complex relationships in the data, their higher variability and computational inefficiency efficiency may limit their applicability in automated machine learning systems. Therefore, the authors findings suggest that linear methods—particularly PCA—may be considered both methodologically and operationally optimal for most conventional classification problems within the CRISP-DM framework, whereas nonlinear methods are more suitable in scenarios where nonlinear structures are known to exist or where representation quality outweighs computational efficiency constraints.

The results of the statistical significance tests performed provide additional information on the differences between the classification models in connection with the different methods of dimensionality reduction. McNemar's test for binary classification tasks and Cochran's Q test for multiclass datasets are applied. Particularly when the PCA method was used with RandomForest, SVM, and XGBoost classifiers, a sizable number of statistically significant differences ($p < 0.05$) were found within the Phoneme dataset between models based on logistic regression (LogReg) and other classifiers. With a p-value of 0.0000 from the comparison between LogReg PCA and RandomForest PCA, the distribution of predictions was clearly highly significantly changed. Conversely, the matching p-values were 1.0000 and 0.9011 respectively when LogReg PCA was compared with LogReg LDA or RandomForest LDA. This suggests that models using different dimensionality reduction methods yet belonging to the same classifier family have not much variation. These findings imply that rather than the single impact of any one of these components, the combination of the classifier and the reduction technique determines the classification performance in major importance. Methodologically, these results validate the need of formal statistical validation inside the evaluation stage of the CRISP-DM framework. In this regard, McNemar's test helps to identify consistent prediction variations at the level of individual instances, so improving the dependability of choosing the most suitable combination of classifier and dimensionality reduction technique for application in automated analytical pipelines. Finally, the author’s findings imply that combinations like RandomForest PCA and XGBoost PCA differ greatly from simpler models like LogReg PCA, which methodologically supports their usage in more difficult classification tasks. All results, together with the Python code, are also available at the GitHub link.

5 Conclusion

This paper provides a comparative analysis of methodologically different dimensionality reduction techniques — PCA, LDA, Factor Analysis and AE, t-SNE, Kernel PCA — for multiclass and binary classification tasks on different datasets. The use of the adapted CRISP-DM methodology ensured a transparent implementation of all phases of the data processing and analysis process, from problem understanding to model implementation.

The results obtained confirm that the choice of dimensionality reduction method significantly influences the classification efficiency and computational feasibility of the model. The PCA method proves to be robust, stable and computationally efficient and is particularly suitable for classification tasks, while the AE method may have advantages in certain cases, e.g. for non-linear data structures. The combination of metrics and execution time provides a multidimensional evaluation of the model and thus contributes to a more comprehensive understanding of the application of dimensionality reduction methods in a production environment. The proposed automated approach, based on the CRISP-DM methodology, has been shown to be suitable and extensible for complex evaluations of machine learning models related to dimensionality reduction. Future research could aim to extend the analysis to other dimensionality reduction methods such as UMAP, as well as to develop and integrate new dimensionality reduction approaches and methods into the AutoML system to further extend the

applicability and scientific contribution of the proposed methodology.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., & Zhang, X. (2016). TensorFlow: A system for large-scale machine learning. https://doi.org/10.48550/arXiv.1605.08695

Baratchi, M., Wang, C., Limmer, S., van Rijn, J. N., Hoos, H., Bäck, T., & Olhofer, M. (2024). Automated machine learning: past, present and future [Article]. Artificial Intelligence Review, 57(5), Article 122. https://doi.org/10.1007/s10462-024-10726-1

Barrett, P., Hunter, J., Miller, J. T., Hsu, J. C., & Greenfield, P. (2005). matplotlib -- A Portable Python Plotting Package.

Bartal, Y., Fandina, N., & Neiman, O. (2019). Dimensionality reduction: theoretical perspective on practical measures. Neural Information Processing Systems,

Bratkovsky, E. (2024, 10.12.2024). A simple explanation of CRISP-ML for beginners. https://www.kaggle.com/discussions/general/487787

Chandra, S. B. (2024). Introduction to Autoencoders [Theory and Implementation]. Medium. Retrieved 2.7.2024. from https://medium.com/@c17hawke/introduction-to-autoencoders-theory-and-implementation-72fed7cf2f70

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS. https://books.google.hr/books?id=po7FtgAACAAJ

Cohen, E. (2017). Reducing Dimensionality from Dimensionality Reduction Techniques. Medium. Retrieved 2.7.2024. from https://medium.com/towards-data-science/reducing-dimensionality-from-dimensionality-reduction-techniques-f658aec24dfe

Costa, R. (2022). The CRISP-ML Methodology: A Step-by-Step Approach to Real-World Machine Learning Projects. Independently published.

developers, s.-l. (2024a). LogisticRegression. scikit learn. Retrieved 6.6.2024. from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

developers, s.-l. (2024b). RandomForestClassifier. scikit learn. Retrieved 6.6.2024. from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

developers, s.-l. (2024c). Support Vector Machines. scikit learn. Retrieved 6.6.2024. from https://scikit-learn.org/stable/modules/svm.html

developers, x. (2024). XGBoost. xgboost developers. Retrieved 6.6.2024. from https://xgboost.readthedocs.io/en/stable/

Fournier, Q., & Aloise, D. (2019, 3-5 June 2019). Empirical Comparison between Autoencoders and Traditional Dimensionality Reduction Methods. 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE),

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Ghobadi, F., Tayerani Charmchi, A. S., & Kang, D. (2023). Feature Extraction from Satellite-Derived Hydroclimate Data: Assessing Impacts on Various Neural Networks for Multi-Step Ahead Streamflow Prediction [Article]. Sustainability (Switzerland), 15(22), Article 15761. https://doi.org/10.3390/su152215761

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://books.google.hr/books?id=omivDQAAQBAJ

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Healy, J., & McInnes, L. (2024). Uniform manifold approximation and projection. Nature Reviews Methods Primers, 4(1), 82. https://doi.org/10.1038/s43586-024-00363-x

Jolliffe, I. T. (2002). Principal Component Analysis. Springer. https://books.google.hr/books?id=_olByCrhjwIC

Mayur Prakashrao Gore, A. C. A. A. S. (2024). Applying-Principal-Component-Analysis-and-Autoencoders-for-Dimensionality-Reduction-in-Data-Stream. International Journal of Innovative Research in Engineering & Management (IJIREM), 11(5), 121-126. https://doi.org/10.55524/ijirem.2024.11.5.17

Mendes Junior, J. J. A., Freitas, M. L. B., Siqueira, H. V., Lazzaretti, A. E., Pichorim, S. F., & Stevan, S. L. (2020). Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels armband approach [Article]. Biomedical Signal Processing and Control, 59, Article 101920. https://doi.org/10.1016/j.bspc.2020.101920

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825--2830.

Perktold, J., Seabold, S., Taylor, J., & statsmodels-developers. (2025a). statsmodels.stats.contingency_tables.cochrans_q. statsmodels. Retrieved 6.5.2025. from https://www.statsmodels.org/dev/generated/statsmodels.stats.contingency_tables.cochrans_q.html

Perktold, J., Seabold, S., Taylor, J., & statsmodels-developers. (2025b). statsmodels.stats.contingency_tables.mcnemar. statsmodels. Retrieved 6.5.2025. from https://www.statsmodels.org/dev/generated/statsmodels.stats.contingency_tables.mcnemar.html

Santosh, K. C., Das, N., & Ghosh, S. (2022). Chapter 2 - Deep learning: a review. In K. C. Santosh, N. Das, & S. Ghosh (Eds.), Deep Learning Models for Medical Imaging (pp. 29-63). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-823504-1.00012-X

Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud, Artificial Neural Networks — ICANN'97 Berlin, Heidelberg.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4).

van der Maaten, L., & Hinton, G. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative review. J Mach Learn Res, 10, 66--71.

Vantuch, T., Snasel, V., & Zelinka, I. (2016). Dimensionality Reduction Method's Comparison Based on Statistical Dependencies. Procedia Computer Science.