Exploring Dimensionality Reduction Techniques: Clustering and Neural Network Analysis

Mar Tejedor Ninou
CS7641 Spring 2024

I. Introduction

In this paper, we delve into the field of clustering and dimensionality reduction algorithms, exploring their applications and interactions with datasets. Our focus lies on Expectation Maximization (EM) and K-Means for clustering, along with Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction.

II. Datasets

For this project we utilized two distinct datasets for our experiments, the same datasets as for Assignment 1:

A. Wisconsin Breast Cancer Dataset

This dataset comprises features extracted from digitized images of fine needle aspirate (FNA) of breast masses. Each instance is characterized by various attributes such as texture, smoothness, and compactness. It is a small dataset with 569 samples and 30 different features which makes it interesting for dimensionality reduction. The primary goal of this dataset is to predict whether a breast mass is benign or malignant based on these features.

B. White Wine Quality Dataset

This dataset contains physicochemical properties and sensory attributes of white wines, along with quality ratings assigned by wine experts. Features include acidity, alcohol content, pH, among others. The objective is to predict the quality of white wines based on these attributes. This dataset is medium size, bigger than the previous one, with 4898 samples and 11 features. Following the structure of Assignment 1 the quality label column will be distributed between 'Good' and 'Bad' depending if the quality score is greater or less than 5. The label column will help us evaluate results comparing between results.

III. Hypothesis

Our hypothesis revolves around the application of dimensionality reduction techniques to these datasets and about how clustering can improve Neural Networks:

A. Hypothesis 1

"Dimensionality reduction techniques, when applied to the Wisconsin Breast Cancer Dataset and White Wine Quality Dataset, will reveal underlying patterns and relationships among the features, leading to improved classification performance"

Hypothesis 1 is based on the nature of the Wisconsin Breast Cancer Dataset and the objectives of our study. Given the complex nature of the dataset with numerous features extracted from FNA images, it is reasonable to assume that there exist underlying patterns and relationships among these features that may not be immediately evident in the high-dimensional space. By reducing the dimensionality of the dataset using techniques we aim to uncover these hidden structures and simplify the dataset while retaining its discriminatory power.

B. Hypothesis 2

"Introducing clustering algorithms as additional features in the dataset will improve the performance of classification tasks by providing the neural network with more discriminative information for accurate prediction."

For the second hypothesis, we came up with it based on the rationale that clustering algorithms can uncover hidden patterns and structures within the data that may not be apparent through traditional feature representations. By incorporating clustering results as additional features, we hypothesize that the neural network will be able to leverage this additional information to enhance its ability to classify data accurately. This is grounded in the idea that clustering algorithms can identify clusters or groups of data points with similar characteristics, which may correspond to distinct classes or categories in the classification task. By providing the neural network with more discriminative information derived from clustering, we anticipate that it will achieve better performance in accurately predicting the class labels of the data points.

Through our experiments, we aim to validate both hypothesis and gain insights into the effectiveness of dimensionality reduction techniques and clustering algorithms in improving the interpretability and predictive performance of models trained on the datasets.

IV. Methodology

In this study, we present the methodologies employed to investigate the impact of clustering and dimensionality reduction algorithms on Wisconsin Breast Cancer and White Wine Quality Datasets. Our analysis includes the utilization of Expectation Maximization (EM) and K-Means for clustering purposes, aiming to identify underlying clusters within the datasets. EM operates iteratively to refine estimates of cluster parameters, while K-Means partitions data into K clusters by minimizing within-cluster variance. Additionally, we apply Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction. PCA captures maximum data variance to reduce dimensionality, while ICA separates mixed signals into statistically independent sources. Randomized Projections approximate high-dimensional data efficiently, and t-SNE preserves local and global relationships for visualization. Through quantitative analysis and visual inspection, we evaluate the performance of these algorithms across datasets, aiming to validate our hypotheses and gain insights into their behavior in various scenarios.

V. Clustering Algorithms

In this section, we present the results of applying K-Means and Gaussian Mixture Model (GMM) clustering algorithms to both the Wisconsin Breast Cancer and White Wine Quality datasets. We utilized silhouette score as a metric to determine the optimal number of clusters for each dataset and algorithm, and interestingly, for both datasets, Figure 1 and Figure 2, the silhouette score was highest

when choosing 2 clusters for both algorithms. The specific number of clusters will be keep and used in future sections when applying it to optimize Neural Networks.

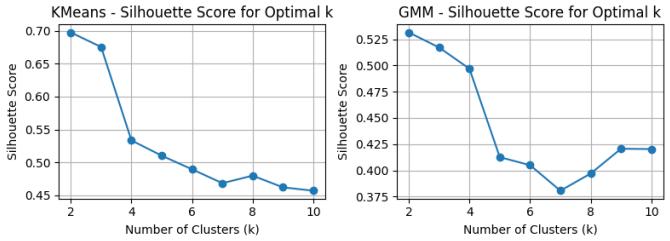

Fig. 1. Wisconsin Breast Cancer Dataset - Silhouette Score for clustering.

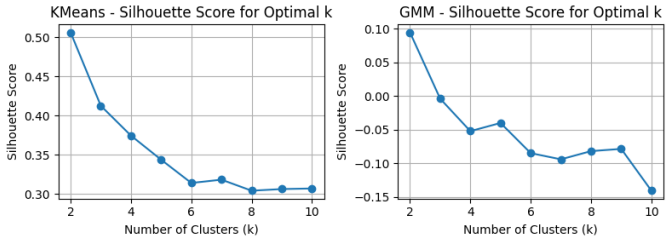

Fig. 2. White Wine Quality Dataset - Silhouette Score for clustering.

Upon analyzing the visual results and selecting the features that exhibited the best performance with the help of pair plots, Figure 3 and Figure 4, we observed distinct differences between the clustering outcomes for the two datasets. For the Wisconsin Breast Cancer dataset, the clusters were well-defined and closely aligned with the labels provided in the dataset for both of the algorithms. This suggests that the clustering algorithms successfully captured the inherent patterns in the data, leading to clear separation between benign and malignant breast masses.

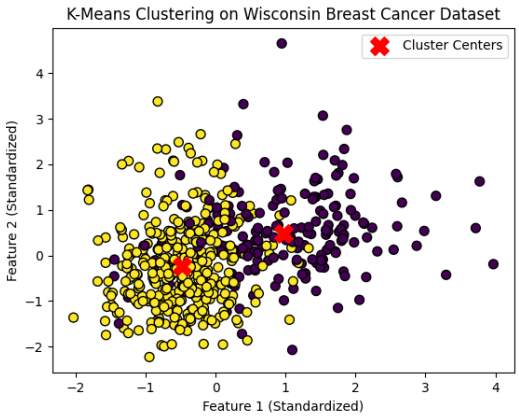

Fig. 3. Application of KMeans (2 clusters) on Wisconsin Breast Cancer Dataset.

Conversely, for the White Wine Quality dataset Figure 5 and Figure 6, while the clustering results were promising, we noticed that the distance from the cluster centers was slightly less pronounced compared to the Cancer Dataset. This observation could be attributed to the larger number of samples and fewer features in the wine

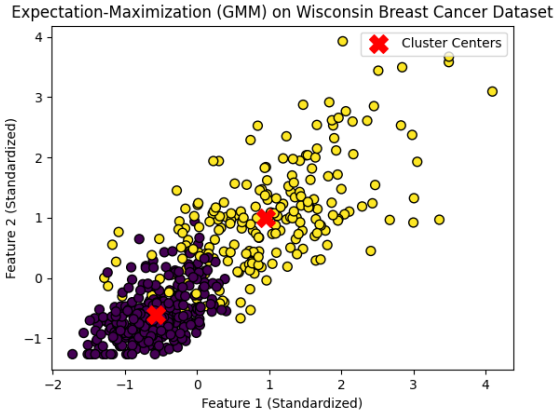

Fig. 4. Application of GMM (2 clusters) on Wisconsin Breast Cancer Dataset.

dataset, which may have introduced more complexity and noise into the clustering process. Despite this, the clusters still exhibited meaningful patterns, indicating that the algorithm was able to discern distinct groups within the dataset.

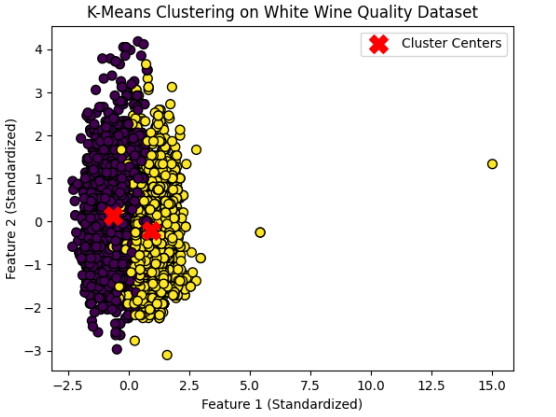

Fig. 5. Application of KMeans (2 clusters) on White Wine Quality Dataset.

The observed clusters align well with the inherent characteristics of the datasets, with the Cancer Dataset clusters exhibiting clear classification boundaries and the Wine Dataset clusters displaying slightly less pronounced separations. This disparity may stem from the differences in dataset size, complexity, and feature space, highlighting the importance of considering dataset-specific factors when applying clustering algorithms.

Comparing the performance of K-Means and GMM, both algorithms yielded similar clustering outcomes, with neither significantly outperforming the other. However, given the inherent differences in their underlying assumptions and methodologies, further analysis is warranted to determine which algorithm is better suited for each dataset.

To potentially improve clustering performance, modifications to the algorithms, such as incorporating feature scaling or exploring different distance metrics, could be explored. Additionally, conducting more comprehensive feature selection or engineering processes may help enhance clustering outcomes, particularly for datasets with high dimensionality and complexity.

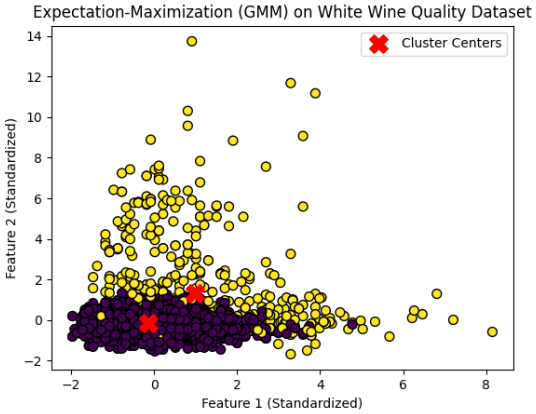

Fig. 6. Application of GMM (2 clusters) on White Wine Quality Dataset.

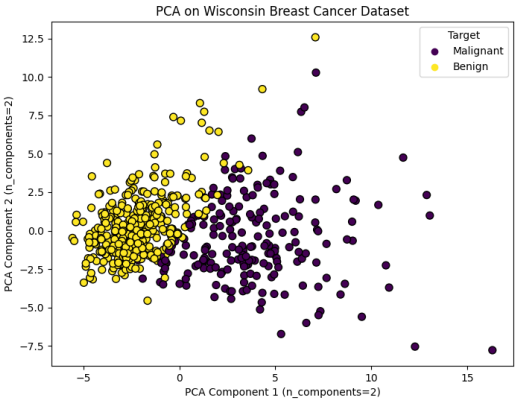

Fig. 8. Application of PCA to Wisconsin Breast Cancer Dataset.

VI. Dimensionality Reduction

A. Principal Component Analysis (PCA)

In this section, we employ Principal Component Analysis (PCA) to reduce the dimensionality of the Wisconsin Breast Cancer and White Wine Quality datasets. PCA facilitates the transformation of high-dimensional data into a lower-dimensional space while retaining as much variance as possible, enabling visualization and interpretation of the underlying structure.

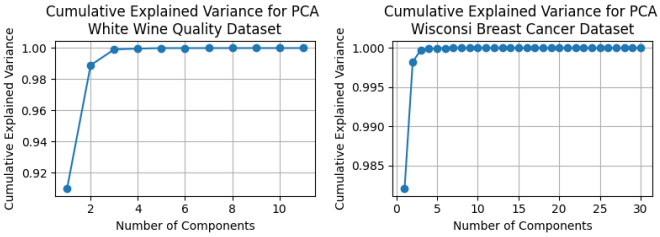

Fig. 7. Cumulative Explained Variance for PCA depending on number of components

To determine the optimal number of components, we utilize the cumulative explained variance plot, Figure 7. For the Cancer dataset, we observed that the cumulative explained variance continued to increase steadily until reaching a plateau after 3 components, suggesting that these components capture the majority of the data's variance. Similarly, for the Wine dataset, the cumulative explained variance reached a plateau after 4 components. Thus, we chose three components for the Cancer dataset and four components for the Wine dataset to balance dimensionality reduction with information retention. This optimal number of components values will be keep and used in future sections when optimizing Neural Networks with the help of dimensionality reduction.

Reducing the number of components to two for visualization purposes, we observe distinct patterns in the reduced space. In the Cancer dataset, Figure 8, the clusters exhibit clear separation, indicating meaningful distinctions between benign and malignant tumors. However, the separation is less pronounced in the Wine dataset, suggesting a more complex underlying structure, Figure 9.

Analyzing the distribution of eigenvalues in Figure 10, we observe a steep drop-off followed by a more gradual decline, indicating that the first few components capture the most variance in the data. This

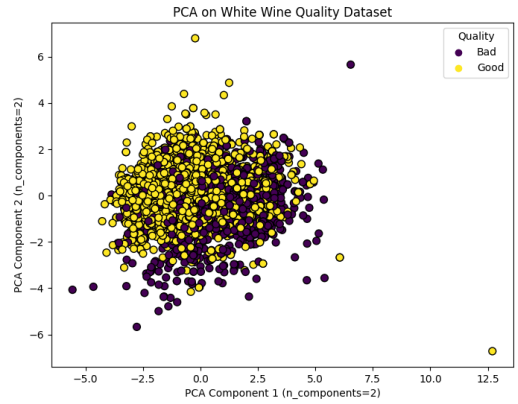

Fig. 9. Application of PCA to White Wine Quality Dataset.

distribution highlights the effectiveness of PCA in summarizing the data's variability with a reduced set of components.

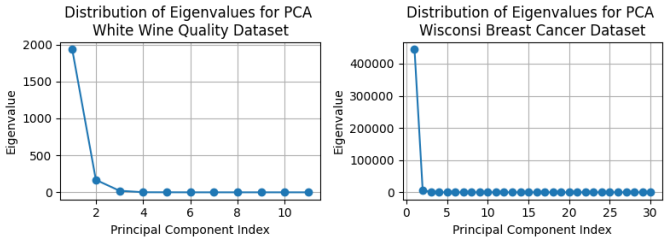

Fig. 10. Distribution of eigenvalues for PCA depending on number of components

Regarding the impact of collinearity, PCA inherently addresses multicollinearity by orthogonalizing the components, thereby reducing redundancy in the data. However, the effectiveness of PCA may be influenced by the specific properties of the dataset, such as the degree of collinearity and noise present. Noise can obscure underlying patterns and affect the performance of PCA, underscoring the importance of preprocessing and data quality assessment.

In summary, PCA provides valuable insights into the underlying structure of the datasets, facilitating visualization and interpretation.

By reducing dimensionality while preserving essential information, PCA enables the identification of meaningful patterns and relationships within the data, enhancing our understanding of complex datasets.

B. Independent Component Analysis (ICA)

In this section, we utilize Independent Component Analysis (ICA) to perform dimensionality reduction on the Wisconsin Breast Cancer and White Wine Quality datasets. ICA aims to separate mixed signals into statistically independent sources, allowing for the extraction of meaningful features and patterns from the data.

To determine the optimal number of components, we calculate the normalized mean kurtosis for each component and select the number of components with the highest kurtosis value, Figure 11. For the Cancer dataset, we identified 12 components with the highest kurtosis, while for the Wine dataset, 6 components exhibited the highest kurtosis values. These components are chosen to capture the most significant independent sources of variation within the datasets. This information about the number of components will be used in future sections when optimizing Neural Networks.

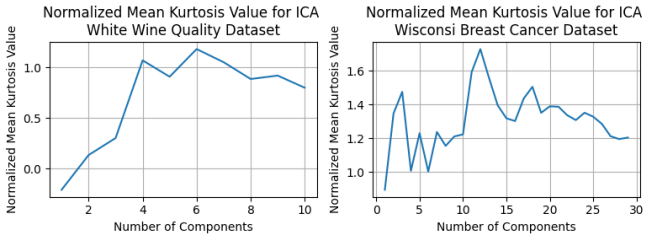

Fig. 11. Normalized Mean Kurtosis Value for ICA depending on number of components.

Reducing the number of components to two for visualization purposes, we observe the data in the new space, Figure 12 and Figure 13. However, unlike PCA, where clear separation between clusters was observed, the distinction is less pronounced in ICA-transformed space for both datasets. This suggests that while ICA effectively separates mixed signals, it may not capture as meaningful distinctions when compared to the original features or other dimensionality reduction techniques.

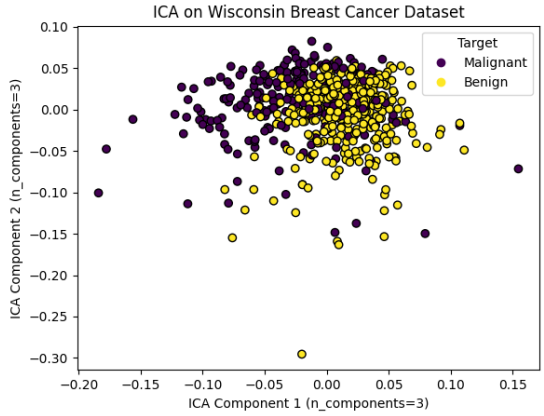

Fig. 12. Application of ICA to Wisconsin Breast Cancer Dataset.

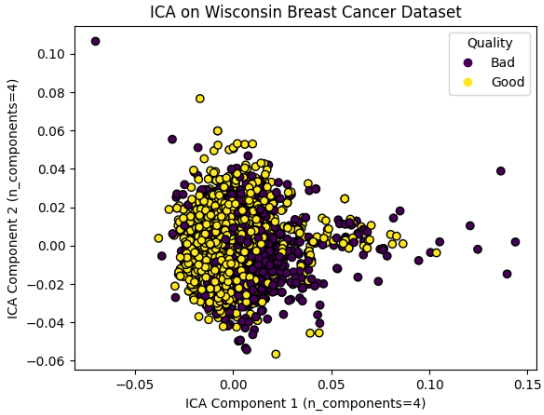

Fig. 13. Application of ICA to White Wine Quality Dataset.

Analyzing the distributions of kurtosis, we observe higher values indicative of non-Gaussianity in the selected components. However, the projection axes for ICA may not capture anything inherently meaningful, as the interpretation of the independent components may not directly correspond to the original features or class labels in the dataset.

Regarding the impact of noise and collinearity, ICA is robust to noise and can effectively separate sources even in the presence of noise. However, the performance of ICA may be influenced by the degree of non-Gaussianity and the specific properties of the data, such as the distribution of the features and the presence of outliers.

In summary, ICA offers a powerful approach to dimensionality reduction by extracting independent sources of variation from the data. While it may not capture as meaningful distinctions as other techniques in certain contexts as we can see on the plots, ICA remains a valuable tool for uncovering hidden patterns and features within complex datasets.

C. Randomized Projections (RP)

In the Randomized Projections (RP) section of the paper, we utilized the reconstruction error as a metric for selecting the number of components to reduce the dimensionality of the datasets as we can see in Figure 14. However, both graphs depicting the reconstruction error as a function of the number of components exhibited a consistent linear decrease without any distinct elbow points. This absence of clear inflection points makes it challenging to determine the optimal number of components for dimensionality reduction.

Given the lack of clear guidance from the reconstruction error graphs, choosing the best number of components becomes a more subjective decision. Researchers may opt to select a number of components based on domain knowledge, computational constraints, or specific requirements of downstream analyses. Alternatively, exploratory analysis techniques such as sensitivity analysis or experimentation with different numbers of components can provide additional insights into the trade-offs between dimensionality reduction and information preservation. In this paper, after trying all possible number of components and making use of the domain knowledge, the number of components that performed better was 14 for the Wisconsin Breast Cancer Dataset and 6 for White Wine Quality Dataset.

For visualization purposes, we reduced the dimensionality to two components to facilitate visualization in two-dimensional space, as il-

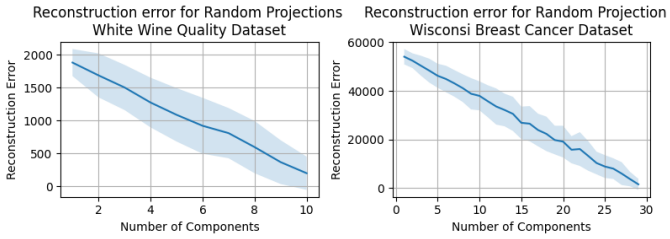

Fig. 14. Reconstruction Error for RP depending on number of components.

lustrated in Figure 15 and Figure 16. However, despite this reduction, the resulting scatter plots failed to display clear separation among data points for both datasets. This observation suggests a limited efficacy of RP in capturing meaningful patterns or structures within the data. Understanding the specific properties of the datasets and their impact on algorithm outputs is crucial for interpreting and contextualizing the outcomes of dimensionality reduction analyses.

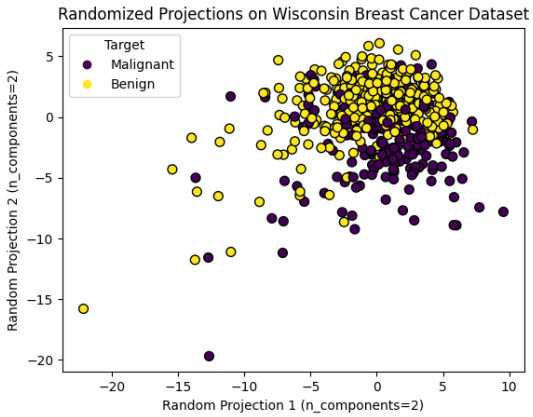

Fig. 15. Application of RP to Wisconsin Breast Cancer Dataset.

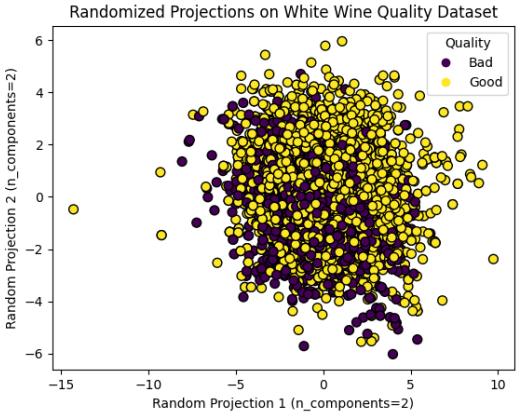

Fig. 16. Application of RP to White Wine Quality Dataset.

D. t-Distributed Stochastic Neighbor Embedding (t-SNE)

In the t-SNE section of the paper, we encountered challenges in determining the optimal number of components for dimensionality reduction. Unlike linear techniques such as PCA or ICA, t-SNE does not provide a straightforward method for selecting the number of components. Instead, researchers often rely on trial and error or domain knowledge to determine the appropriate number of components. Also, due to the restriction given by "sklearn" and the Barnes Hut algorithm the numbers of components cannot be more than 4. It is for this reason we picked 2 as the number of components for future experiments regarding t-SNE.

For visualization purposes and along with previous explanations, we reduced the dimensionality to two components to visualize the data in two-dimensional space. In the case of the Cancer Dataset, t-SNE exhibited excellent performance, Figure 17, effectively separating the data into distinct clusters. However, for the Wine Dataset, the visualization was less informative, as the clusters appeared to be mixed and lacked clear separation, Figure 18.

The difference in t-SNE's performance between the Cancer Dataset and the Wine Dataset can be attributed to various factors. Firstly, the inherent structure and distribution of the data points differ significantly between the two datasets. Additionally, t-SNE's sensitivity to parameters like perplexity, learning rate, and number of iterations can lead to suboptimal visualization outcomes, especially in datasets with complex structures or high dimensionality. Moreover, t-SNE's non-linear nature may struggle to capture intricate relationships within the data, particularly when the underlying structure is not well-suited for non-linear mapping. Overall, while t-SNE can be a powerful tool for visualizing high-dimensional data and uncovering latent structures, its effectiveness may vary based on the dataset's characteristics and parameter settings.

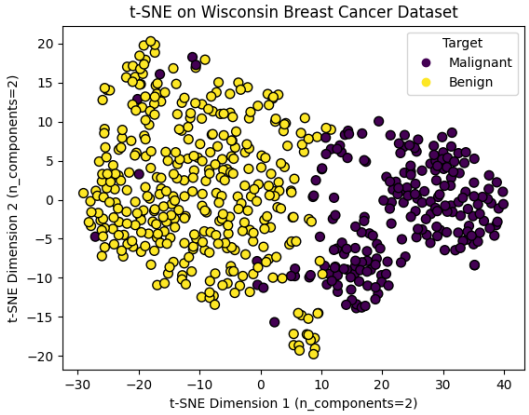

Fig. 17. Application of t-SNE to Wisconsin Breast Cancer Dataset.

Understanding the influence of these properties on algorithm outputs is essential for interpreting the results of dimensionality reduction analyses accurately. Further investigation into the impact of noise, collinearity, and other factors on algorithm performance can provide valuable insights into the behavior of dimensionality reduction techniques in diverse datasets.

VII. Clustering with Dimensionality Reduction

In the Clustering with Dimensionality Reduction section of the paper, we employed Principal Component Analysis (PCA) for linear dimensionality reduction and t-Distributed Stochastic Neighbor Embedding (t-SNE) for non-linear dimensionality reduction. We selected these techniques based on their demonstrated effectiveness in the

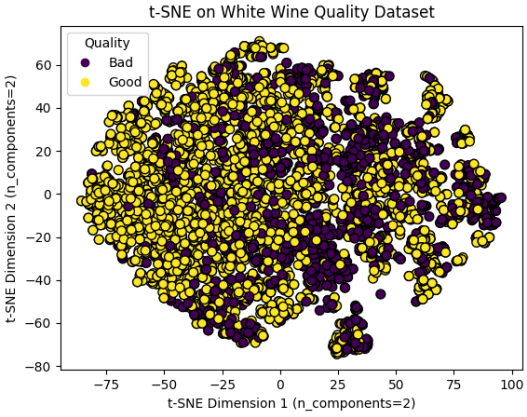

Fig. 18. Application of t-SNE to White Wine Quality Dataset.

previous sections. Additionally, we utilized KMeans clustering due to its superior performance with both datasets in earlier analyses.

Applying the optimal number of components determined in the previous section, Table I, along with the number of clusters set to 2 based on previous findings, we projected the datasets onto the reduced spaces created by PCA and t-SNE. The clustering results were then visualized in two-dimensional graphs.

TABLE I
NUMBER OF COMPONENTS PER ALGORITHM

DR Algorithm	Cancer Dataset	Wine Dataset
PCA	3	4
ICA	12	6
RP	14	6
t-SNE	2	2

The outcomes of the clustering experiments were highly promising, as evidenced by the remarkable accuracy achieved in both 2D visualizations and datasets (refer to Figure 19, Figure 20, Figure 21 and Figure 22). The clusters displayed clear separation and alignment with the ground truth labels from the original datasets. This success can be attributed to the effective reduction of dimensionality through techniques like PCA and t-SNE, which preserved meaningful patterns and structures in the data, facilitating accurate clustering.

The exceptional performance of the algorithms and clustering in this case can be attributed to several key factors. Firstly, the quality of dimensionality reduction played a crucial role, with PCA and t-SNE effectively preserving essential information while reducing the dimensionality of the original. PCA captured linear relationships, while t-SNE uncovered complex nonlinear structures, both contributing to improved data representation. Additionally, the robustness of the clustering algorithm, KMeans, played a significant role. Its simplicity and versatility enabled it to effectively separate the data into distinct clusters. Moreover, the selection of the optimal number of components and clusters based on metrics like variance and silhouette score ensured that the dimensionality reduction and clustering processes were finely tuned for the datasets, leading to well-defined and meaningful clusters. Lastly, the inherent structure and characteristics of the datasets, such as the Wisconsin Breast Cancer and White Wine Quality datasets, likely contributed to successful clustering, as they may inherently possess distinct clusters or patterns well-captured by the dimensionality reduction and clustering

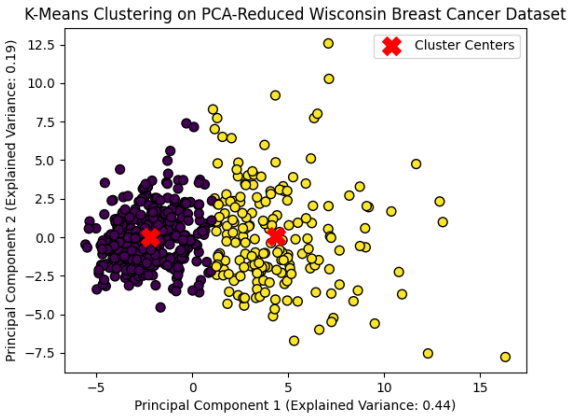

Fig. 19. Application of KMeans on a PCA-reduced Wisconsin Breast Cancer Dataset.

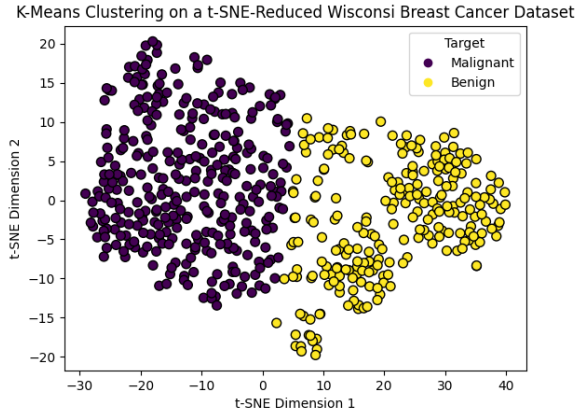

Fig. 20. Application of KMeans on a t-SNE-reduced Wisconsin Breast Cancer Dataset.

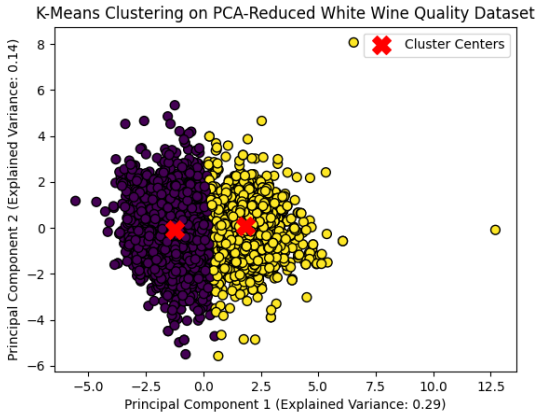

Fig. 21. Application of KMeans on a PCA-reduced White Wine Quality Dataset.

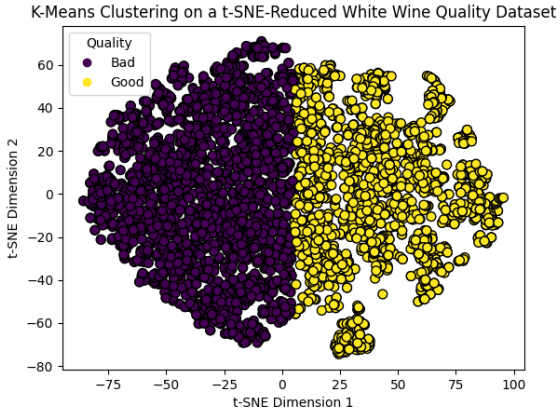

Fig. 22. Application of KMeans on a t-SNE-reduced White Wine Quality Dataset.

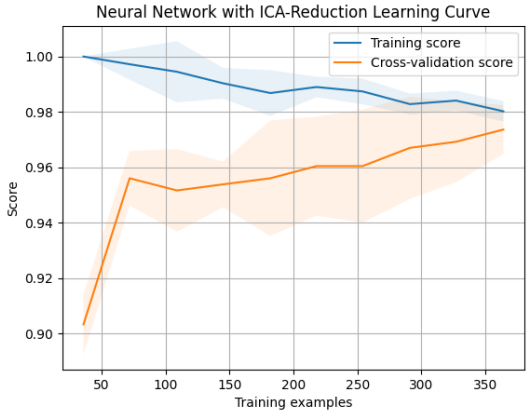

Fig. 23. Neural Network Learning Curve with ICA reduction on Wisconsin Breast Cancer Dataset.

algorithms. Collectively, these factors contributed to the outstanding performance observed in the algorithms and clustering outcomes.

VIII. Neural Networks with Feature Reduction

In the Neural Network with Dimensionality Reduction section, we utilized Independent Component Analysis (ICA) for linear dimensionality reduction and t-Distributed Stochastic Neighbor Embedding (t-SNE) for non-linear dimensionality reduction. Leveraging the optimal number of components determined in the previous section (Table I), we applied ICA and t-SNE to the Wisconsin Breast Cancer Dataset. The choice of this dataset over the White Wine Quality Dataset was based on its better performance as demonstrated in the previous section. We specifically selected ICA for linear dimensionality reduction due to its ability to capture the underlying structure of complex datasets like the Cancer Dataset. Additionally, ICA has shown really good performance in previous sections of this paper.

For the Cancer Dataset, we adopted an 80-20 split for training and testing, respectively. Grid search was employed to identify the best hyperparameters for the neural network. Remarkably, the hyperparameters obtained were consistent with those from Assignment 1, suggesting their efficacy in yielding favorable results for the neural network. Specifically, the selected hyperparameters included a layer size of (15, 5), an alpha value of 0.0001, and a ReLU activation function.

Upon analyzing the learning curves, ICA (Figure 23) demonstrated a favorable learning trajectory with a descending training percentage, indicating effective learning without signs of overfitting. This smooth decline in the training curve suggests that ICA effectively captures the underlying structure of the data, allowing the neural network to generalize well to unseen samples. In contrast, t-SNE exhibited a less smooth learning curve, with noticeable changes in direction in both the training and testing sets (Figure 24). The fluctuations in the t-SNE learning curve may stem from the non-linear nature of t-SNE, which can introduce complexities in the optimization process and hinder convergence to an optimal solution. Additionally, the non-linear transformations performed by t-SNE may lead to more intricate learning dynamics compared to linear techniques like ICA. While both reduction algorithms displayed improvements with larger sample sizes, ICA outperformed t-SNE in terms of accuracy, likely due to its ability to preserve meaningful linear relationships in the data, which are crucial for effective classification by neural networks.

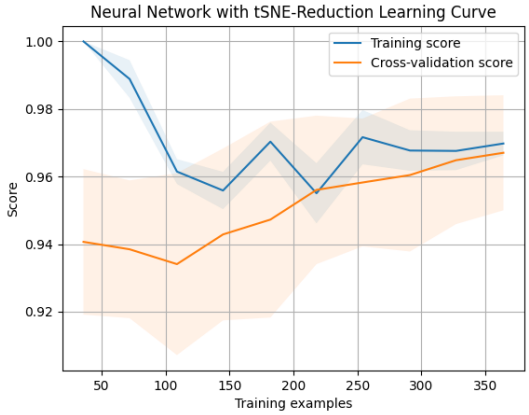

Fig. 24. Neural Network Learning Curve with t-SNE reduction on Wisconsin Breast Cancer Dataset.

When comparing the performance metrics between supervised learning (SL) and unsupervised learning (UL) utilizing ICA and t-SNE, as shown in Table II, distinct differences emerge. UL with ICA achieved a notably higher accuracy of 98.25% compared to SL, which attained 92.39%. However, this enhanced accuracy came at the cost of a slightly longer runtime. In contrast, UL with t-SNE exhibited lower accuracy at 96.34% and required an extended runtime compared to SL.

TABLE II
Time and Accuracy for Neural Networks

	Time	**Accuracy**
SL Neural Network	33.08s	92.39%
UL Neural Network with ICA	55.43s	98.25%
UL Neural Network with t-SNE	71.34s	96.34%

In evaluating differences in performance and speed, considerations include accuracy, runtime, and the stability of learning trajectories. The disparity in learning curves and runtime between ICA and t-SNE underscores the importance of selecting the most appropriate dimensionality reduction technique for a given dataset and task. Further exploration into the reasons behind the varying performance of different reduction algorithms can provide valuable insights into

their strengths and limitations.

IX. Neural Networks with Clustering

In this section, we integrated KMeans and Gaussian Mixture Model (GMM) clustering results as new features into the Wisconsin Breast Cancer dataset before applying neural networks algorithm.

Following an 80-20 split for training and testing the Cancer Dataset, we leveraged grid search to identify optimal hyperparameters for the neural network. Remarkably, the hyperparameters obtained mirrored those from Assignment 1, affirming their efficacy for neural network classification. Specifically, the selected hyperparameters included a layer size of (15, 5), an alpha value of 0.0001, and a ReLU activation function.

Upon evaluating the performance metrics and analyzing the learning curves, Figure 25 and Figure 26, it is evident that both KMeans and GMM clustering algorithms demonstrate favorable outcomes. The learning curves for both algorithms exhibit smooth training and testing phases, indicating no signs of overfitting. Particularly, as the number of samples increases, the testing accuracy tends to improve for both algorithms as it is expected. However, there are subtle differences observed in the behavior of the testing lines. KMeans' testing line increases smoothly as more samples are added, reflecting a consistent improvement in accuracy. In contrast, GMM's testing line shows more fluctuations, with occasional ups and downs, suggesting a slightly less stable performance compared to KMeans.

Despite the longer runtimes associated with both KMeans and GMM compared to supervised learning, their performance in terms of accuracy is notably higher. KMeans achieves an impressive accuracy of 98.24%, while GMM achieves 97.36%. This significant improvement in accuracy highlights the efficacy of incorporating clustering results as additional features in enhancing the neural network's performance. By enriching the dataset with additional discriminative information derived from clustering, the neural network gains better insights for classification tasks, leading to improved accuracy compared to traditional supervised learning approaches.

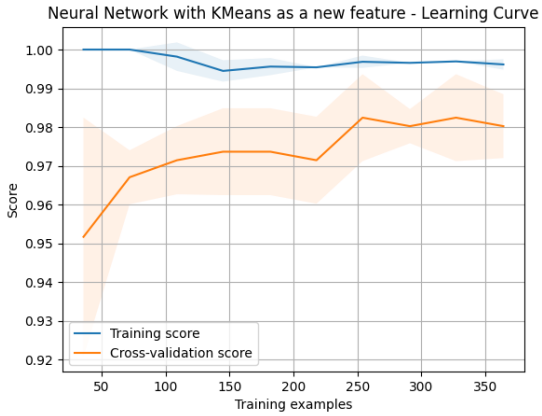

Fig. 25. Neural Network Learning Curve with Kmeans clusters as new features on Wisconsin Breast Cancer Dataset.

Comparing the performance metrics with the previous assignment, the neural network with clustering as a new feature showcased enhanced accuracy at the expense of increased runtime. Despite the longer runtime, the substantial improvement in accuracy underscores the effectiveness of incorporating clustering results as additional features for enhancing classification performance.

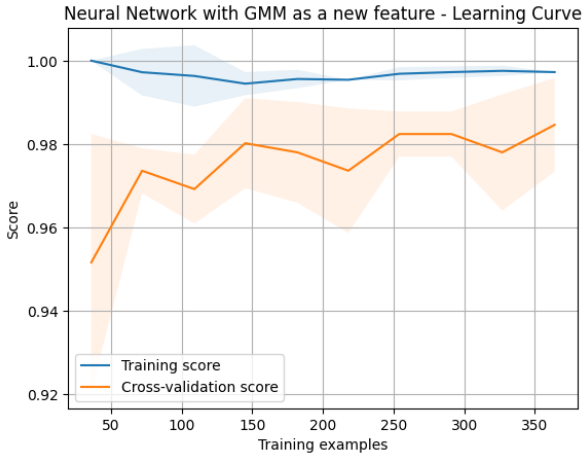

Fig. 26. Neural Network Learning Curve with GMM clusters as new features on Wisconsin Breast Cancer Dataset.

X. Conclusion

In this paper, we conducted a comprehensive exploration of clustering and dimensionality reduction algorithms, aiming to enhance classification performance and gain deeper insights into the underlying structures of the datasets. Our investigation yielded valuable findings regarding the effectiveness of these techniques and their impact on classification tasks.

We began by formulating hypotheses regarding the application of dimensionality reduction techniques to the datasets, anticipating improved classification performance through the discovery of underlying patterns and relationships among features. We also hypothesized that incorporating clustering results as additional features would enhance neural network classification accuracy by providing discriminative information.

Through meticulous experimentation, we validated our hypotheses and uncovered key insights. Dimensionality reduction techniques, including PCA, ICA, RP, and t-SNE, effectively captured essential data structures and reduced dimensionality while preserving critical information. These techniques provided valuable insights into the datasets' complexities, facilitating visualization and interpretation of intricate relationships.

Furthermore, our exploration of clustering algorithms such as KMeans and GMM revealed their ability to identify distinct clusters within the datasets. By incorporating clustering results as new features, we observed significant enhancements in neural network classification performance. This underscores the importance of leveraging unsupervised learning techniques to enrich supervised learning tasks and improve predictive accuracy.

Our findings also highlighted the influence of algorithm selection and dataset characteristics on performance outcomes. While some techniques, such as PCA and KMeans, exhibited robust performance across datasets, others, like t-SNE, demonstrated greater sensitivity to dataset complexities and characteristics.

In conclusion, our study emphasizes the importance of employing a combination of unsupervised and supervised learning techniques to uncover hidden patterns, enhance data representation, and improve classification accuracy. By continuing to explore and refine these techniques, researchers can advance our understanding of complex datasets and enhance predictive modeling capabilities in various domains.