CS 7641 Assignment 3: Unsupervised Learning

Ali Alrasheed
College of Computing
OMSCS

I. Introduction

This report aims to explore four different dimensionality reduction algorithms which are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Projection (RP), and Random Forest (RF). In addition, K-Means and Expectation Maximization (EM) clustering algorithms will be explored on two different datasets (Heart disease and spambase) and their reduced data using the above-mentioned dimensionality reduction techniques. Furthermore, the reduced data will be used to train neural networks to examine their performance. Finally, the outcome of the clustering labels will be used as new features that will be added to the original and reduced data to re-run the neural network experiments.

II. Datasets

A. *Heart Disease*

The Heart Disease dataset that was used in assignments 1 and 2 is also used in this assignment. The dataset was obtained from UCI Machine Learning Repository [1] which contains 300 instances with 14 features. However, since this dataset contains some categorical features that need to be hot-coded, the total features of this dataset will be 22. Since this dataset contains 22 features, it will be interesting to use it in clustering and dimensionality reduction algorithms. In addition, it will be also interesting to see the effect of having categorical data on the performance of clustering and dimensionality reduction techniques. For more details about the dataset, refer to the assignment 1 report.

B. *Spam*

The second chosen dataset was Spambase which is also provided by UCI Machine Learning Repository [2]. The dataset contains 4601 instances of spam and non-spam emails. Each email is represented by 57 features that indicate the frequency of a particular word occurring in the email. The spam emails were collected from individuals and postmasters that manually reported the emails as spam, while the non-spam emails were taken from personal and work emails. Since this dataset has high dimensionality features (57 features), it will be interesting to evaluate dimensionality reduction algorithms using this dataset.

III. part 1: Clustering (original dataset)

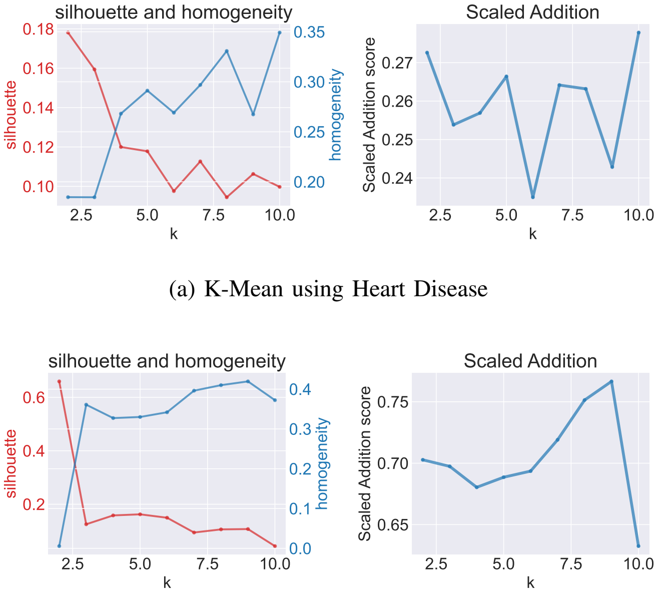
In this section, two clustering algorithms will be evaluated using the above-mentioned datasets. In particular, K-Means and Expectation Maximization (EM) will be evaluated using Silhouette, and Homogeneity. Since the Silhouette score does not peak into the labels, it will be used as the primary metric to decide the number of clusters, whereas the Homogeneity score will be used to validate the concluded hypothesis.

In addition, the elbow method is one of the most popular methods to decide the number of clusters (k) using the SSE graph (inertia). The number of clusters where the elbow occurs is considered the optimal k. The rationale behind the elbow method is that any k value after the elbow is considered a point of *diminishing return* where the cost is not worth the added improvement. However, in this report, the maximum Silhouette score will be used to decide the optimal number of clusters (unsupervised methods). The Silhouette score range from -1 to 1, where -1 means not well separated clusters, and 1 means very dense and separated clusters. The Silhouette score works by rewarding lower intra-clusters mean distance, and higher inter-cluster mean distance. Thus, the number of clusters (k) with the highest Silhouette score will be chosen as the optimal k in this report. Another criterion to evaluate the number of clusters is to look at their homogeneity. The homogeneity score needs the data and its label to calculate the score, and thus, it is only used for evaluation. It ranges from 0 to 1, where 1 means that each cluster consists of points of the same class (based on the label).

A. *K-Means*

The K-means algorithm aims to group data into k different clusters by minimizing the sum-squared error (SSE) within the clusters. The K-means algorithm has three main steps. First, it starts with random points as the centroids of the clusters. Then, it involves looping between calculating the new clusters based on the current centroids and then updating the centroids based on the new clusters from the previous step. The K-means is guaranteed to converge given enough iterations, although it does not guarantee optimally. In this section, the K-means algorithms will be used to cluster both heart disease and spambase datasets. The primary decision for the number of clusters will be taken based on the maximum silhouette score since it does not utilize the label to calculate the score. In addition, the Homogeneity score will be used to validate the concluded number of clusters based on the silhouette score.

Figure 1a shows both the Silhouette and homogeneity scores for Heart Disease data. As stated earlier, a higher Silhouette score is desired since it means the clusters are denser and better separated. It can be seen from figure 1a that the maximum Silhouette score for the heart disease data is at k=2. To validate the hypothesis from the Silhouette score, the Homogeneity score is used. By looking at the homogeneity score, it can be seen that the maximum is at k=10. This is expected since at homogeneity score is maximum when each cluster consists of points from the same class, which means that the homogeneity is likely to increase with a higher number of clusters since each cluster will need to classify a smaller number of points. In addition, a trade-off between the Silhouette and homogeneity score can be found by calculating the scaled addition of both scores which can be found in figure 1a. Since maximizing both scores is desired, we need to look for the maximum score of the scale addition score which can be found at around k = 10 then k =2.



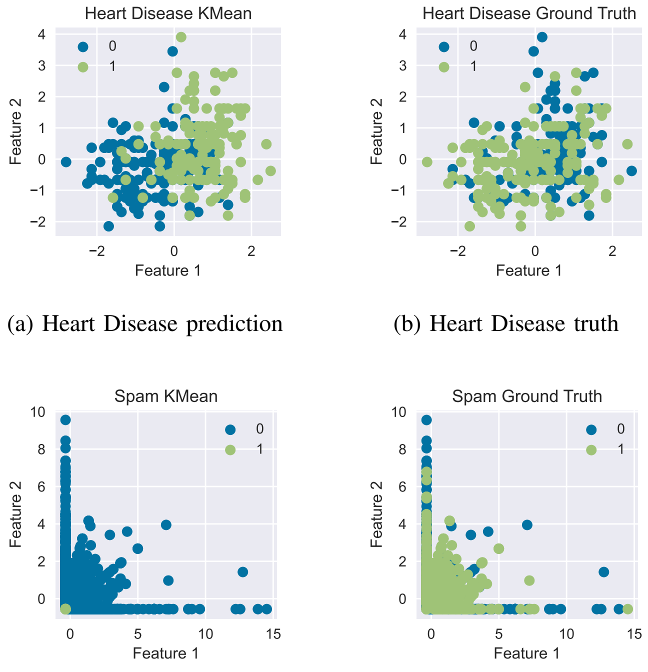
(a) K-Mean using Heart Disease

(b) K-Mean using Spam

Fig. 1: Silhouette and Homogeneity Scores for both datasets

Furthermore, figure 1b shows both the Silhouette and homogeneity score for the spambase dataset. Since the Silhouette is an unsupervised method to determine the number of clusters (k), it will be used as the primary metric to decide the optimal k. It can be seen from figure 1b that the maximum Silhouette score is at k=2 which is consistent with dataset labels (binary classification). However, the Homogeneity score is increasing with a higher number of clusters, and it is maximized at k=9, and it is likely to increase with more clusters. The reason is similar to what is stated in the previous paragraph. In addition, a trade-off between the Silhouette and homogeneity score can be found by calculating the scaled addition of both scores which can be found in figure 1b. The scaled addition graph in figure 1b shows that k=9 is the best trade-off between Silhouette and homogeneity scores.

It is important to visualize the result of the K-means clustering based on the Silhouette scores (k=2 for both datasets). Figure 2 shows the K-means clusters (k=2) for both heart disease and spambase datasets compared to their ground truth. As it is shown from the figure, the K-means did relatively well clustering the heart disease dataset compared to its ground truth. However, it did terrible job clustering the spambase dataset. This is expected since both datasets contain high dimensional features which tend to blow up the euclidean distance (curse of dimensionality) and thus deteriorate the K-means performance. It would be interesting to see the effect of dimensionality reduction algorithms on the K-means performance. Since the spambase has 57 features compared to 22 features (with one-hot encoding), the K-means performance on spambase is much worse which is expected. In addition, the heart disease contained categorical data which will affect the performance of the K-Means since it relies on Euclidean distance which is not meaningful for categorical data. K-modes could be used instead of K-Means, but since the heart disease dataset contains both continuous and categorical data, even K-modes might have bad performance.



(a) Heart Disease prediction

(b) Heart Disease truth

(c) Spam prediction

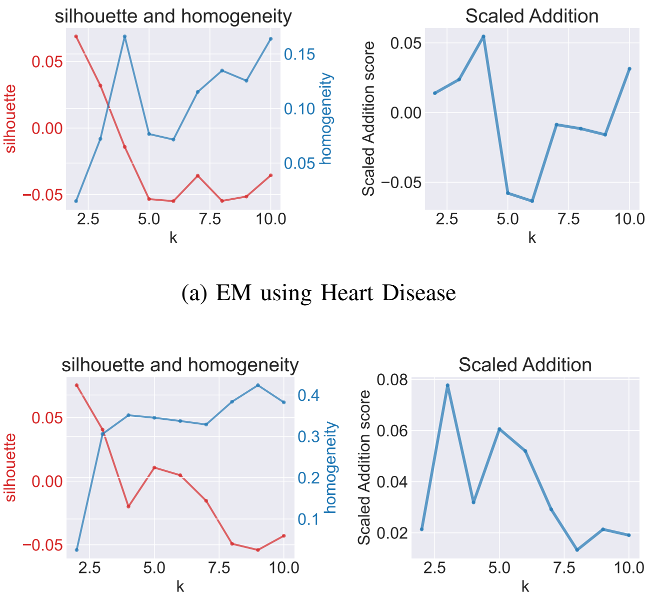
(d) Spam ground truth

Fig. 2: Predicted and ground truth clusters for both datasets visualized over two features

B. *Expectation Maximization (EM)*

Expectation Maximization (EM) is a common clustering algorithm that uses a mixture of Gaussian models to estimate the cluster boundaries in the given data. The most common way to evaluate the performance of EM is the Bayesian information criterion (BIC). However, for the sake of consistency, the Silhouette score will be used instead as the primary metric for deciding the optimal number of clusters (k). Similarly, the homogeneity score will be used as a validation method to assess the performance of Silhouette's hypothesis.

Figure 3a shows both the Silhouette and homogeneity scores of the Heart disease dataset. Similar to the analysis done in the K-means section, the desired is to maximize the Silhouette score since a higher value of Silhouette indicates more dense and separated clusters. The Silhouette score will be used to decide the optimal number of clusters (k) since it is an unsupervised method (without peaking into the labels). On the other hand, the homogeneity score will be used as an evaluation metric. Figure 3a illustrates that the maximum Silhouette score occurs at k=2. This is consistent with the dataset labels (binary classification). The result is also consistent with the result found using the K-means algorithm illustrated in figure 1a. In general, the homogeneity score tends to increase for a higher number of clusters. Figure 3a shows that the homogeneity score is at maximum at k=4 and k = 10, and it is likely will be higher with a higher number of clusters for the same reason explained earlier in the K-means section. The scaled addition of Silhouette and homogeneity scores provides a trade-off between them, which could be used to choose the best number of the cluster that maximizes both scores. In this case, it can be seen that k=4 and k =10 are the optimal value of k.



(a) EM using Heart Disease

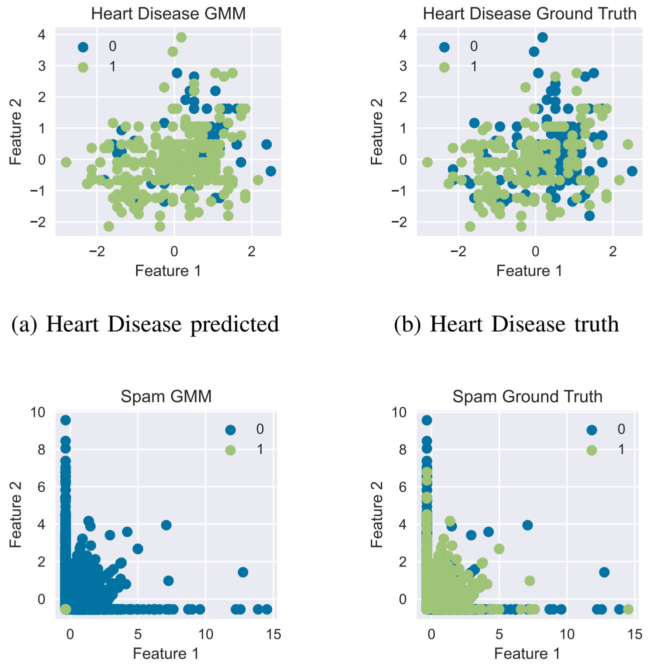
(b) EM using Spam

Fig. 3: Silhouette and Homogeneity Scores for both datasets

A similar analysis can be made for the spambase dataset. Figure 3b shows that the maximum Silhouette score is at k=2, and the maximum homogeneity score is at k=10. The scaled addition between the Silhouette and the Homogeneity score is shown in figure 3b. The scaled addition graph provides a trade-off between the Silhouette and homogeneity scores. If maximizing both scores is desired, the best number of clusters based on the EM on the spambase data is at k=3. It might be the case that there are some emails that are not easily distinguished between spam and not spam, and hence the data can be better clusters using 3 groups.

Based on the Silhouette scores, the optimal number of clusters for both datasets is two (k=2). The visualized clusters based on the expectation maximization (EM) algorithm can be found in figure 4. The figure shows that the EM did relatively well in clustering the heart disease data (figure 4a) compared to the ground truth (figure 4b). In addition, by comparing figure 2b and figure 4b, EM performed better than K-means in clustering the heart disease data. This is expected since EM does not relies on Euclidean distance to decide the number of clusters since the Euclidean distance may deteriorate the performance of any clustering algorithms when used with categorical data.

Similarly, the visualized clusters of spam data based on the EM algorithm are illustrated in figure 4c. The figures show that EM performs very poorly in clustering the spam data compared to the ground truth in figure 4d. This is expected since the spam dataset has 57 features which make it hard to cluster. Dimensionality reduction algorithms can help in improving the performance of the clustering algorithms, and this will be the focus of the next section.



(a) Heart Disease predicted

(b) Heart Disease truth

(c) Spam prediction

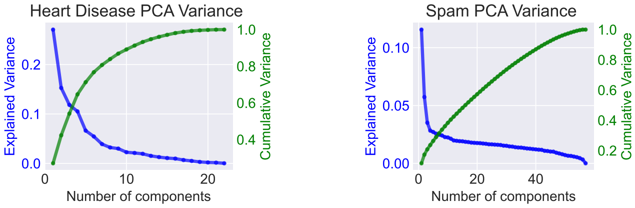
(d) Spam ground truth

Fig. 4: Predicted and ground truth clusters for both datasets visualized over two features

IV. Dimensionality Reduction

A. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction algorithm that transforms the data from its original coordinate into a new coordinate system where most of the variance of the data can be explained with only a few dimensions. The higher the variability of the data, the better it can be explained. The PCA technique looks for the vectors or dimensions that maximize the variability of the data and thus can eliminate dimensions with very low variability.



(a) Heart Disease Dataset

(b) Spambase Dataset

Fig. 5: PCA Analysis on both datasets

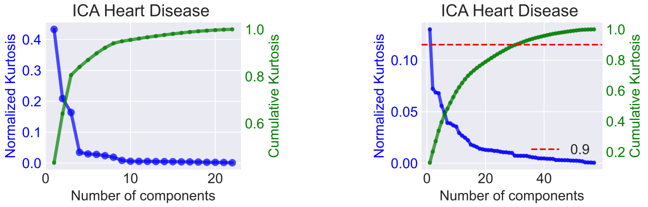
The cumulative explained variance will be used as the primary metric to decide the number of components of PCA. The explained variances are equivalent to eigenvalues of the covariance of the data which represent the variance of each principal component. The explained variance of the heart disease and spam datasets are shown in figure 5. They are sorted in descending order of the magnitude of the variance. To facilitate choosing the appropriate number of principal components, the cumulative explained variance is calculated and plotted in orange. The elbow method could be used to choose an appropriate number of principal components. However, it is easier to just pick the number of components that explained about 85-90% of the data. By looking at figure 5, it can be seen that any value between 10 and 15 seems to be an appropriate value for the heart disease dataset. For the rest of the report, the number of principal components that will be used for heart disease is 11. In addition, the PCA of the spam data shows that the majority of the components have similar variance which means that more components will be needed to explain about 80-90% of the data. In the case of the spam data, 50 PCA components are needed to achieve around 90% of the cumulative explained variance. Thus, it makes more sense here to aim for 80% of the total explained variance which is achieved after around 35 PAC components.

B. Independent component analysis (ICA)

ICA is another technique that can be used to reduce the dimensionality of the data. ICA transforms that data into a new coordinate system where the features are maximally independent of each other. In general, the ICA aims to reduce the mutual information between the new features and maximizes their Kurtosis (non-Gaussianity). Thus, the Kurtosis will be used as the primary metric to decide the appropriate number of components that will be used for each dataset.

Figure 6 shows the normalized kurtosis (blue) of each component, as well as the cumulative normalized kurtosis (green) for both datasets. The cumulative normalized kurtosis can assist us to visualize the trade-off between adding more features (more complexity) and increasing the kurtosis. Figure 6a shows the cumulative kurtosis has two clear elbows. The first captures about 80% of the total kurtosis, and the second captures around 90%. The second elbow is located at n=8, and since the second elbow captures around 90% of the total kurtosis, the first 8 sorted ICA's components will be used as the new features of the Heart disease dataset.

However, figure 6b shows no clear elbows in the cumulative kurtosis, and therefore, the optimal number of components will be at around 90% of the total kurtosis of the ICA's components. By looking at the figure, it can be concluded that the first 28 sorted ICA components will be used as the new reduced features for the spam dataset.



(a) Heart Disease Dataset

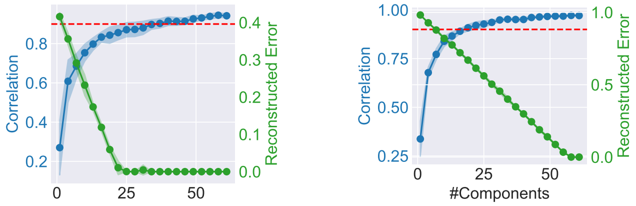
(b) Spambase Dataset

Fig. 6: ICA analysis for both datasets

C. Random Projections (RP)

Sparse Random Projections from Sklearn were used in this section which is another dimensionality reduction technique that projects the original data using a random sparse matrix. Since the process is random, it has the advantage of low computational cost compared to the other dimensionality reduction techniques. However, since the project is random,the accuracy could be also affected.

The pairwise distance correlation between the reduced data and the original data will be used as the primary metric to decide the appropriate number of components for each dataset. The pairwise correlation indicates the similarity between the original data and the randomly transformed data. High correlation means that the transformed data maintained the majority of the information from the original data. In addition, since the technique is inherently random, the means and variance of the correlation are recorded and shown in figure 7. The square of reconstructed error is also shown, but the pairwise distance correlation will be used as the primary metric since the main aim is to have new data that is highly correlated but not necessarily similar (very low reconstructed error).



(a) Heart Disease Dataset

(b) Spambase Dataset

Fig. 7: RP analysis for both datasets

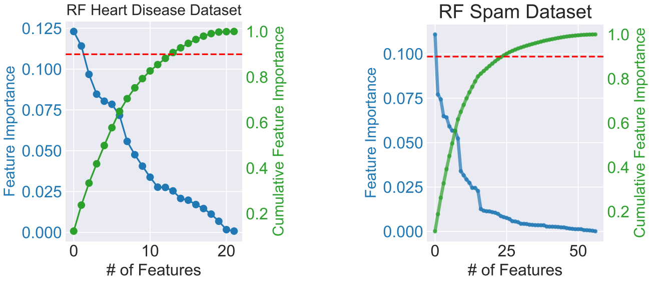
Figure 7a shows the pairwise correlations between the original heart disease data and the transformed data using the sparse randomized projection. The figure indicates that the random projection technique needed around 30 components to reach about 90% correlation with the original data. This is much higher than the original number of features. This might be because the original data contains categorical features which makes it hard for the random projection to reduce the number of dimensions while maintaining a high correlation with the original data. In this case, 15 components of the RP make more sense to use instead of 30 which is not a much worse option since its correlation is slightly above 80%. At k=15, the square of the reconstructed error is relatively higher, but since the generated data has a high correlation of the original data, the 15 components are acceptable.

On the other hand, things are more straightforward with the spam dataset. As it is shown in figure 7b that 20 components were enough to achieve around 90% correlation with the original spam data. Thus, 20 components using RP will be used as the reduced new feature for the spam data. However, although at k=20 the correlation is about 90%, the reconstructed error is very high and it slowly decreases with more components. However, as long as the generated data has a high correlation with the original data, it is not very important that the new data looks exactly the same as the original data if it was reconstructed. Thus, 20 components will be used as the reduced data for the spam algorithms using random projections.

D. Random Forest (RF)

Random forest (RF) is an ensemble technique that is capable of identifying the most important features in the data by looking that their respective Gini impurity. Thus, the random forest can be used to reduce the data to only the most important features in the data. Thus, cumulative feature importance will be used as the primary metric to decide the number of features to keep in the reduced data.

Figure 8 shows the sorted feature importance (blue), as well as the cumulative sum of feature importance (green). Our criterion to decide the number of features is similar to the previous methods which are based on achieving around 90% cumulative feature importance.



(a) Heart Disease Dataset

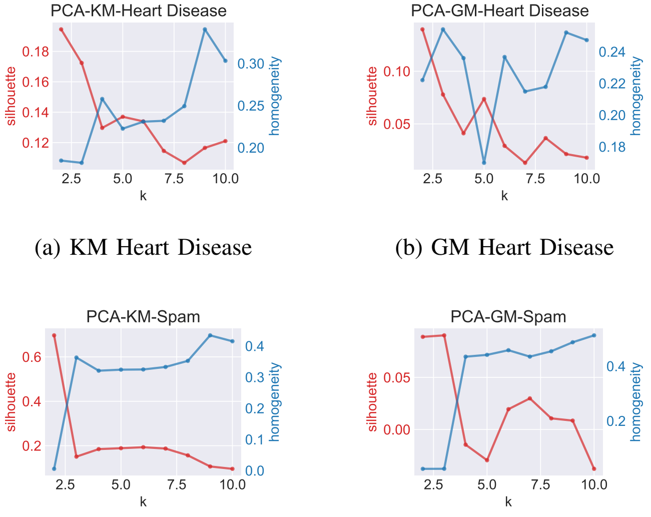
(b) Spambase Dataset

Fig. 8: RF analysis for both datasets

Figure 8a shows the cumulative feature importance for the heart disease dataset. The cumulative feature importance crosses the 90% line at around 12 or 13 features, thus 13 most important features will be used as the reduced new data of the heart disease dataset. Similarly, figure 8b shows the cumulative feature importance for the spambase dataset. The cumulative feature importance has an elbow around 15 or 16 features. However, to ensure good performance, it is better to stick with the number of features that can achieve 90% of the cumulative feature importance which is at around 23 features. Thus, the new reduced dataset for the spam will contain only the 23 most important features of the original data set.

V. Part 3: Clustering with Dimesionality Reduction

A. Principal Component Analysis (PCA)



(a) KM Heart Disease

(b) GM Heart Disease

(c) KM Spam

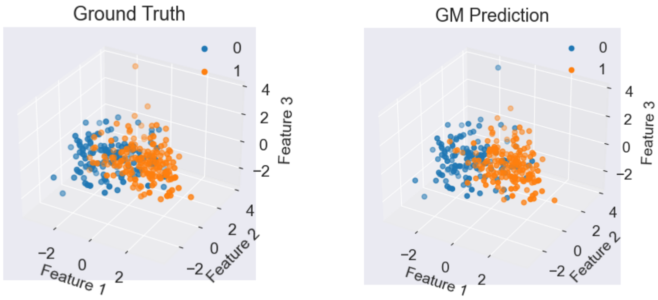
(d) GM Spam

Fig. 9: Silhouette and Homogeneity Scores for both datasets

In section IV-A, it was decided that 11 will be used as the number of components for PCA. Thus, the clustering analysis for the reduced data is shown in figure 9. Regarding the Heart disease dataset, figures 9a and 9b show the reduced data has not changed the location of the maximum Silhouette score which can be found at k=2. However, what is interesting is that the Silhouette score at k=2 has increased from 0.18 to around 0.21 for K-Means and from 0.075 to 0.14 for EM. This suggests that the dimensionality reduction technique using PCA has improved the clustering outcomes compared to using the original data. In addition, the Homogeneity score is also used to evaluate the similarity of the produced clusters compared to the labels. It can be seen by comparing figure 9a with figure 1a and 9b with figure 3a that the reduced data using PCA has caused the homogeneity score to increase more than two-folds. Thus, It is clear that reducing the dimensions for the heart disease data is beneficial for clustering.

PCA was also applied to the spam data using 35 components. However, by looking at figures 9c and 9d it can be seen that the effect of dimensionality reduction using PCA on improving the clustering on the spam dataset was much less obvious than on the heart disease. Both the Silhouette and Homogeneity scores remains almost the same with and with out applying PCA. Although PCA has reduced the data's dimensions from 57 to 35, the number of features is still very high which might the be the reason for the slight improvement when using the spam data. In addition, It was previously shown in figure 3b that using EM both k=2 and k=3 has same Silhouette score, and thus it was not clear which one to choose. However, it can be seen from figure 9d that k=2 is clearly the highest Silhouette score among the other number of clusters which might one of the advantages of using dimesionality reduction on high-dimensional data.

Looking back at figure 9, it is clear that K-Means finds clusters with higher Silhouette scores than EM. However, EM finds clusters that are more similar to the labels (higher Homogeneity scores). Thus, in this case, it seems there is a trade-off between getting a cluster that is more similar to the labels or clusters with more separated boundaries.



(a) Heart Disease Ground Truth

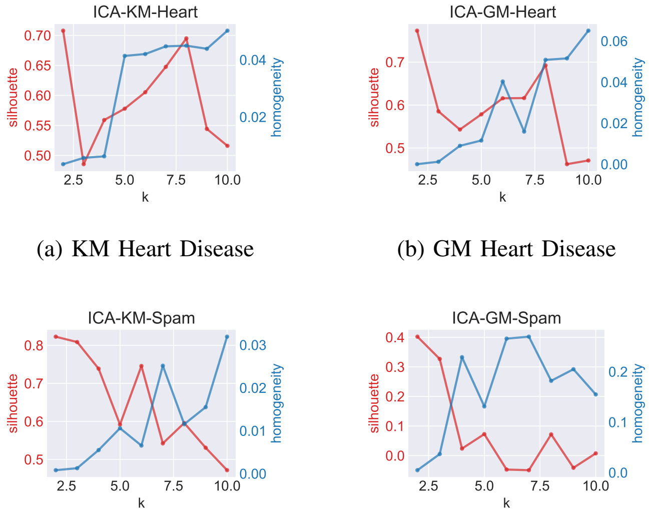
(b) Heart Disease Perdition

Fig. 10: Visualization of EM clustering using PCA on heart disease data

Figure 10 shows the ground truth of labels compared to the clustering labels from the Gaussian Mixture algorithm which is a type of Expectation Maximization (EM) algorithm. The figure shows the EM did a good in both separating the data and matching the ground truth labels which is reflected in its Silhouette and Homogeneity score is shown in figure 9b.

B. Independent Component Analysis (ICA)

From the previous discussion in section IV-B, it was decided that the highest 8 and 28 sorted features based on their kurtosis will be used as the reduced data for the heart disease and spam data respectively. Figure 11 shows the performance of clustering using ICA on both datasets. It is clear that the location of the highest Silhouette score has not changed compared to the original data which is at k=2 for K-Means and EM in both datasets. In addition, figure 11 illustrates two different findings. The first is the Silhouette scores have increased significantly when using ICA compared to the original data. This is not surprising since the ICA maximizes Independence between features which can help in producing more separated clusters and thus higher Silhouette scores. The second finding is that although Silhouette scores have significantly increased, the performance of clustering compared to the labels has become much worse. This can be concluded by looking at the Homogeneity scores have significantly deteriorated when applying ICA to the original data.



(a) KM Heart Disease

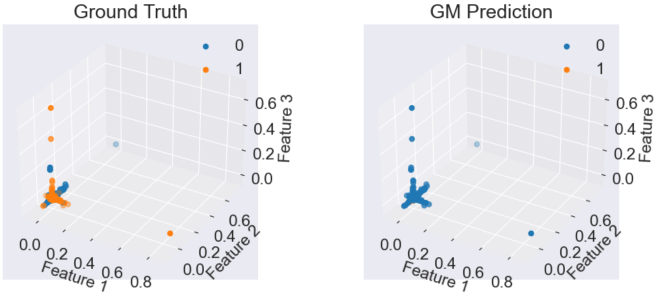
(b) GM Heart Disease

(c) KM Spam

(d) GM Spam

Fig. 11: Silhouette and Homogeneity Scores using ICA

Visualization of the EM clustering performance using the ICA of spam data is shown in figure 12. The purpose of this visualization is to emphasize the low homogeneity scores (almost zero) using the ICA techniques. It can be seen from figure 12 that the clustering method has completely failed to produce a cluster similar to the ground truth labels.



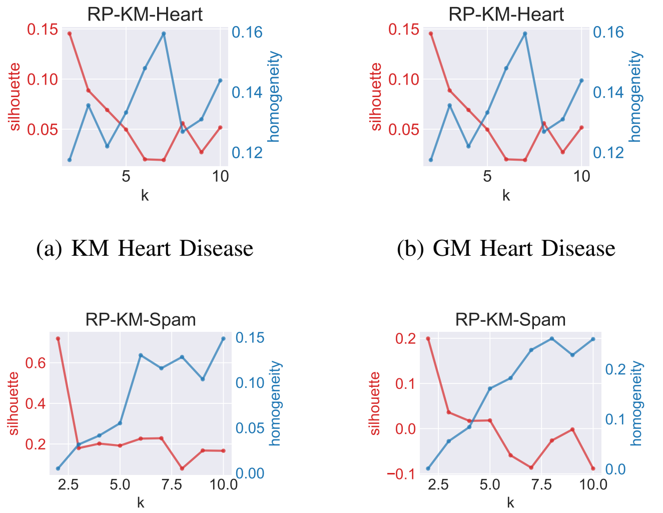
(a) Spam Ground Truth

(b) Spam Perdition

Fig. 12: Visualization of EM clustering using ICA on Spam data

C. Random Projections (RP)

Random projection techniques were also used to reduce the dimensionality of the datasets. Based on the correlation between the reduced and the original data, it was decided in section IV-C that 15, and 20 random projected components shall be used to reduce the original data. Figure 13 once again confirms that the location of the maximum Silhouette score is unchanged which is at k=2. It is interesting to see the performance of K-means has been negatively affected by the random projected data. This can be seen by looking at the Homogeneity scores of K-means on both datasets which have significantly reduced compared to using the original data, especially for a higher number of clusters.



(a) KM Heart Disease

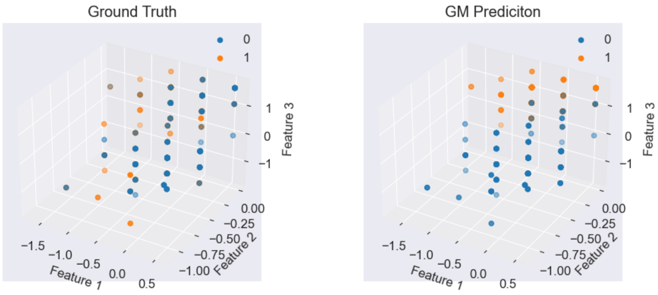
(b) GM Heart Disease

(c) KM Spam

(d) GM Spam

Fig. 13: Silhouette and Homogeneity Scores using ICA

On the other hand, it can be noticed the Silhouette scores produced from the EM have moderately increased which suggests that the result clusters are better separated. However, this did not translate when compares to the labels as the clusters produced from EM were completely different than the labels which resulted in Homogeneity scores of almost zero. In general, the lower clustering performance compared to the labels is expected due to the randomness of generating the features which may result in separable clusters but not necessarily in agreement with the labels.



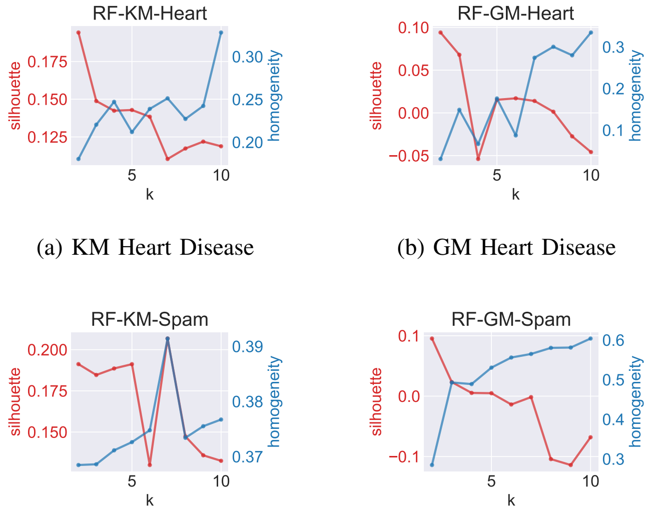
(a) Heart Disease Ground Truth

(b) Heart Disease Perdition

Fig. 14: Visualization of EM clustering using RP on Heart Disease data

Visualization of the EM clustering performance in figure 13a can be seen in figure 14. This visualization is interesting since it shows that EM did a good job in clustering the features based on their distance, but somewhat fail in matching the ground labels.

D. Random Forest (RF)



(a) KM Heart Disease

(b) GM Heart Disease

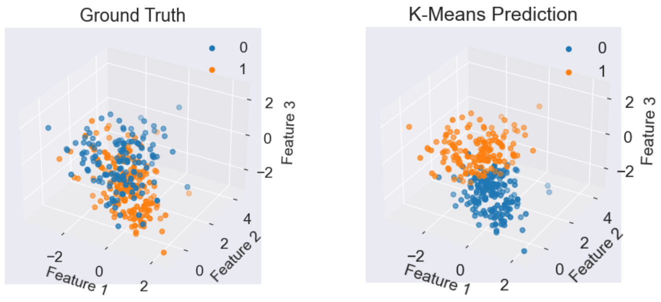
(c) KM Spam

(d) GM Spam

Fig. 15: Silhouette and Homogeneity Scores using ICA

Using Random Forest (RF) the most important features can be identified. Based on the importance of the features, we can decide the if the feature can be dropped, thus, the random forest can be used as a dimensional reduction technique. Based on the analysis done in section IV-D, it was decided that only 12 and 23 features are sufficient to represent the heart disease and spam data respectively. Figure 15 shows the performance of K-means and EM when using only the most important features of the data. It is interesting to see from 15c that the maximum Silhouette location has changed from k=2 to k=7 when using K-Mean for the reduced spam data. It can be also observed from figure 15c that the homogeneity score is also at maximum when k=7. This may suggest that the data can be better explained using seven different clusters. Other than that, the location of the maximum Silhouette score remains at k=2 for other combinations as can be seen in figure 15.

Another interesting finding is that the reduced data based on the feature's importance has significantly increased the homogeneity scores, especially in the spam data. The homogeneity scores of the original data at k=2 using both K-Means and EM were almost zero which can be seen in figure 1b and figure 3b. However, after reducing the features based on their importance, the homogeneity scores have increased significantly to 0.37 and 0.3 for the K-Means and EM respectively. This could be because of the high dimensionality of the original data (52 features) which probably does not add much information but adds significant overhead for the clustering algorithm. It must be mentioned that this increase in the homogeneity scores for the spam data came with a significant reduction in the silhouette scores.



(a) Heart Disease Ground Truth (b) Heart Disease Perdition

Fig. 16: Visualization of K-Means clustering using RF on Heart Disease data

Visualization of the EM clustering performance in figure 15a can be seen in figure 16. Similarly, this visualization is interesting since it shows that K-Means did a good job in clustering the features based on their distance (high Silhouette score), but somewhat fail in matching the ground labels (low Homogeneity score) which is not surprising since the clustering algorithms aims to increase the distance between clusters regards on the true labels.

VI. Part 4: Neural Network performance on the reduced data

In this section, the dimensional reduction techniques analyzed in this report will be used on the heart disease data to train Neural network models to compare their performance to training using the original data. Average cross-validation was used to assess their accuracy performance. Grid search was used to find the best hyper-parameters for each reduced data. The hyper-parameters used in this section can be seen in table I which were based on the grid search of the hidden layer and the analysis done in previous sections.

TABLE I: Hyper-parametes used for each dimensionality techniques

Technique	components	Hidden layer
Original	NA	(500,)
PCA	11	(500,)
PCA	8	(500,)
RP	15	(500,)
RF	12	(100,)

The performance of each dimensionality reduction technique based on table I can be seen in figure 17. It is clear that none of the reduced data was able to beat the model trained using the original data. This is expected since the reduced data contains only around 90% or less of the information of the original data. However, it is interesting to see that the average cross-validation accuracy of the reduced data was very close to the original data except for ICA. The ICA performance was worse compared with others with only eight independent components which were used based on the cumulative kurtosis of the transformed features. However, As it will be seen in figure 19 that ICA will need at least 16 of the highest kurtosis components to produce similar performance compared to the others which are still lower than the original data dimension (22).

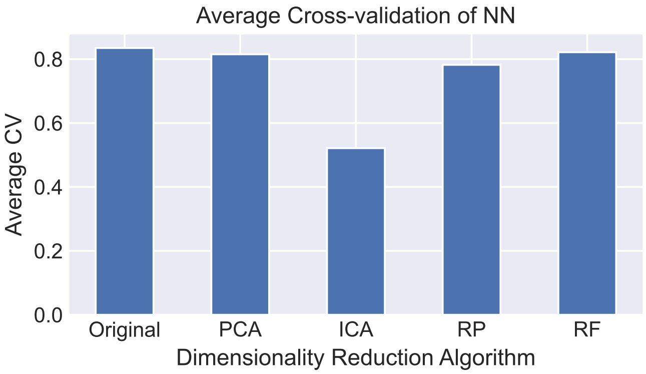
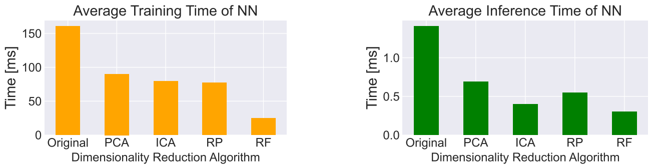


Fig. 17: Average Cross-validation Of Neural Networks

Overall, figure 17 shows it is useful to apply dimensionality reduction techniques on the original data since they can produce similar performance but with less features. However, the main advantage of using dimesnionality reduction is reducing time complexity which can be observed in figure 18. Both the training and testing (inference) time was significantly lower than using the original data. This is expected since number of features is reduced compared to the orignal data. It is also interesting to observe that model trained on the reduced data based on the random forest was to produced a model with similar accuracy to using the orignial data (figure 17) while having the lowest training and testing time as it can be seen in figure 18. It msut be noted that the reason for the low training time for RF compared to other is that the number of neurons of best performing model was found to be only 100 while the others had 500 neurons. This may suggest the heart disease data contained some unnecessary features which increased the neural network complexity for the original data.



(a) Training Time [ms] (b) Inference Time [ms]

Fig. 18: Time Complexity analysis

The analysis above was based on table I which can only give us a snapshot of the performance of the dimensionality reduction techniques. However, the chosen number of comments per technique may not produce the best-performing model. Thus, figure 19 shows the best average cross-validation per component for each of the dimensionality reduction techniques. In general, figure 19 shows that the performance of the PCA and

RF were less sensitive to the number of components compared to RP and ICA. In particular, ICA was very sensitive to the number of components which shows that 17 components are needed to produce a similar performance to the original data which can explain the low performance in figure 17 with only 8 components used. Another important note is that at a higher number of components the reduced data was able to beat the performance of the original data. This could be simply because the grid search did not cover the maximum performance of the model using the original data. However, it could be also that the reduced data removes the unnecessary features in the data which helped the neural network to generalize better.

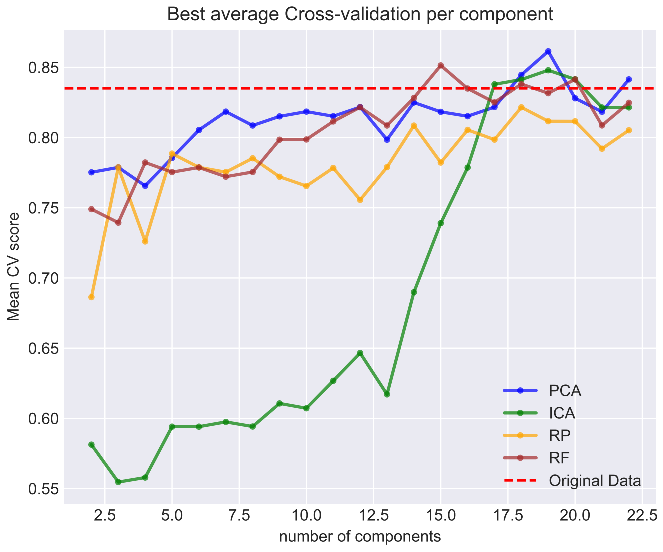


Fig. 19: Best Average Cross-validation score per number of components

VII. Part 5: Clustering as Features

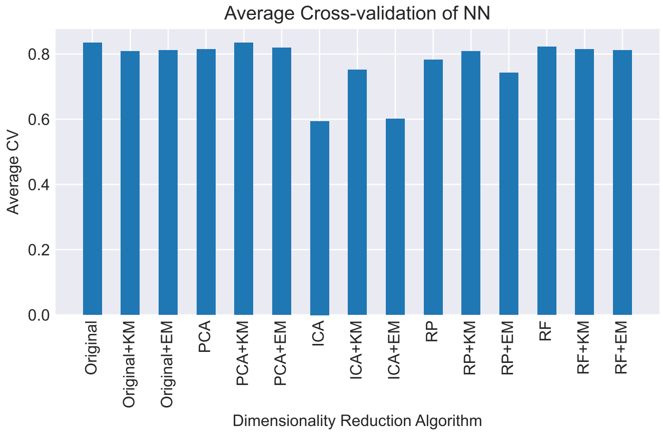


Fig. 20: Average Cross-validation Of Neural Networks with clustering

In this section, KMean and EM will be used to create labels that will be hot-encoded and then added as new features to the original data, and the reduced data (PCA, ICA, RP, and RF). The new data (15 combinations in total) will be used to train a neural network to evaluate its performance. Grid search was used for each one of these combinations to find their best-performing parameters, thus, their hyper-parameters might be different. Figure 20 shows the performance of the neural network with different data. The dimensionality reduction techniques use the number of components that were decided in the analysis in the previous section which can be seen in table I.

Figure 20 mean cross-validation of a diffident neural network trained using 15 different combinations of data. It can be seen that adding the clustering labels as new features helped increase slightly increase the model performance compared to only using the reduced data. Adding the K-Means in particular has shown to be beneficial, especially when adding it to ICA, PCA, and RP. However, adding the clustering labels to the reduced data from Random Forest (RF) does not appear to help to improve the best model's performance.

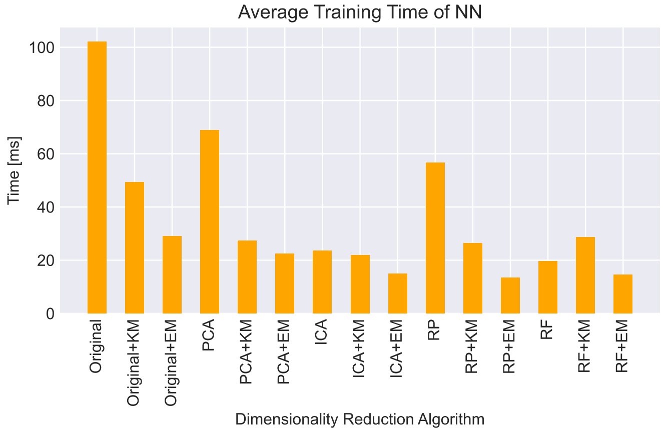


Fig. 21: Training Time [ms]

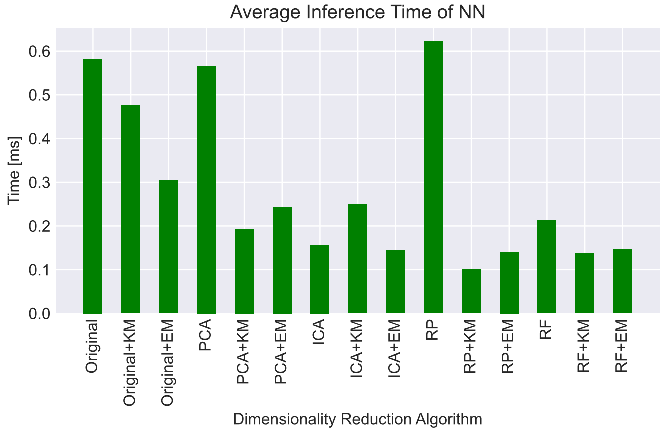


Fig. 22: Inference Time [ms]

Moreover, since reducing the time complexity is one of the main advantage of using dimensionality reduction, it is shown in figure 21 and 22 that in all cases, at least one of the clustering algorithm (EM or K-Means) helped to decreased the training and testing time compared to without adding the clustering labels using these algorithms. Although intuitively adding new features should increase the time complexity of the training and testing, the fact that these clustering labels could have high correlation with the output helps the neural network converge faster. In addition, the clustering data could add important information to the features that it could help reduce the model complexity (number of neurons) which can also decrease the inference time.

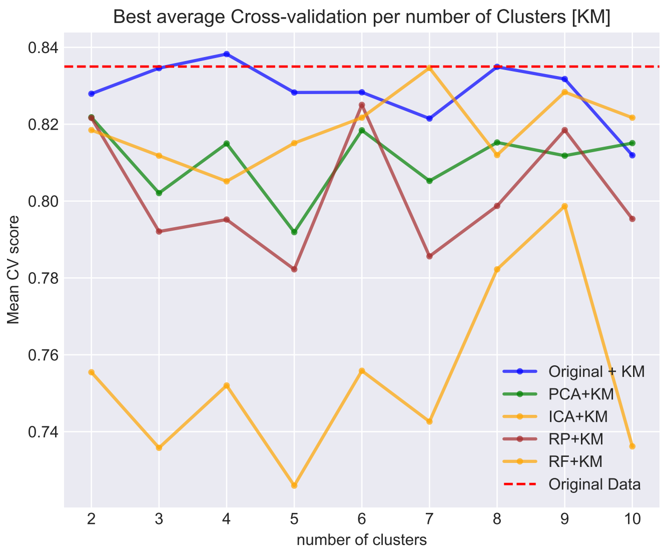


Fig. 23: K-Means Using Number of Clusters

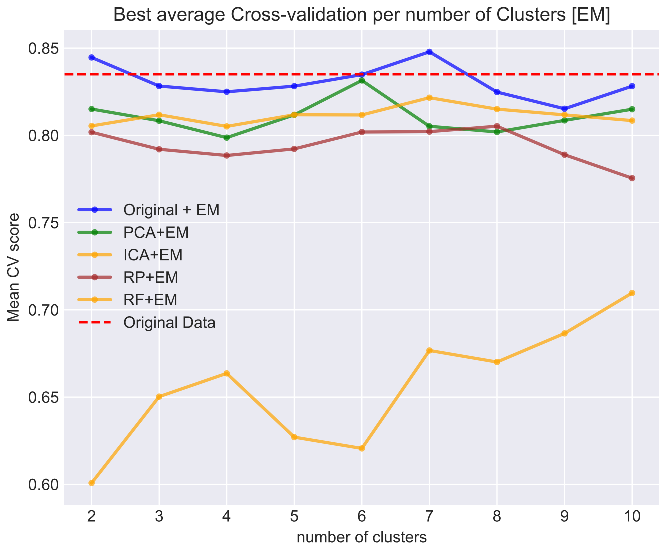


Fig. 24: EM Using Number of Clusters

Finally, figure 23 and 24 show the effect of changing the number of clusters that will be augmented to the reduced data. It is can be seen from the figure that the two clusters gave a good performance compared to the other clusters but were not the optimal number of clusters. This might mean that there is a number of clusters higher than two that could help explain the data better which led to better performance.

VIII. Conclusion

The first section of the report discussed K-Means and EM algorithms, and we found out the performance of these algorithms was better on the heart disease which only has 22 features than the spambase dataset which contains 57 features (higher dimensions). In addition, the Silhouette scores were used to identify the number of clusters for each data using an unsupervised approach, and we found out that both datasets are best divided by two clusters which are consistent with their given labels. In the section and third sections, dimensionality reduction techniques were discussed and applied to the clustering algorithms which shows that ICA maximized the silhouette scores compared to the other algorithms, but produced very low homogeneity clusters compared to the labels. High Silhouette scores using ICA are expected since ICA produces features that are maximally independent which results in clusters that are highly separated, and thus high Silhouette scores. On the other hand, the Random Forest (RF) performs the best compared to the true labels which can be observed by its high homogeneity scores. RF works by limiting the features to only the most important ones which seemed to help the clusters algorithms perform better as some unnecessary features were removed. The fourth and fifth sections focused on evaluating the clustering and dimensionality reduction performance using neural networks trained on the heart disease dataset. Section four showed that neural network models trained on the reduced data were able to produce a similar performance to the model trained using the original data. However, the main advantage that was found was that models trained on the reduced data have significantly low time complexity (training and inference) than the model trained using the original data. Finally, the last section shows that adding the clustering labels to the original and reduced data could help to improve performance and further reduce the time complexity. Although adding more features should intuitively increase the time complexity, the added features for the clustering algorithms could have high correlations with the labels which can help the neural network converge faster.

References

[1] Andras Janosi et al. *Heart Disease Data Set*. 1988. URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[2] George Mark Hopkins Erik Reeber and Jaap. *Spambase Data Set*. 1998. URL: https://archive.ics.uci.edu/ml/datasets/spambase.