****

**Applying Dimension Reduction techniques to measured cardiac blood flow data**

**Team members:**

Dantuluri Jayanth varma

Manish Sreerangam

**Supervisor:**

Dr.Monique Meuschke

Prof. Dr.-Ing. Bernhard Preim

March, 2020

**Contents**

Table of Contents

[List of Acronyms 4](#_Toc67256629)

[List of figures 5](#_Toc67256630)

[List of tables 6](#_Toc67256631)

[1. Introduction 7](#_Toc67256632)

[2. Background 9](#_Toc67256633)

[2.1 Cohorts: 9](#_Toc67256634)

[2.2 Outliers: 9](#_Toc67256635)

[2.3 Feature selection: 9](#_Toc67256636)

[2.4 Feature extraction: 10](#_Toc67256637)

[2.5 Principal component analysis: 10](#_Toc67256638)

[2.6 T-Distributed-Stochastic-Neighbor-Embedding: 10](#_Toc67256639)

[2.7 Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP): 11](#_Toc67256640)

[2.8 Multi-dimensional scaling: 11](#_Toc67256641)

[2.9 Silhouette coefficient: 11](#_Toc67256642)

[2.10 Correlation coefficient: 11](#_Toc67256643)

[2.11 Connectivity: 11](#_Toc67256644)

[3. Approach 13](#_Toc67256645)

[3.1 Principal component analysis (PCA): 13](#_Toc67256646)

[3.2 T-distributed stochastic neighbor embedding (t-SNE): 15](#_Toc67256647)

[3.3 Uniform Manifold Approximation and Projection (UMAP): 17](#_Toc67256648)

[3.4 Multidimensional Scaling (MDS): 19](#_Toc67256649)

[4. Evaluation 21](#_Toc67256650)

[4.1 Qualitative analysis: 21](#_Toc67256651)

[4.2 Quatitative analysis: 24](#_Toc67256652)

[5. Conclusion 25](#_Toc67256653)

[6. References 26](#_Toc67256654)

# List of Acronyms

**BAV** Bicuspid aortic valve

**4D PC-MRI** Four-dimensional phase-contrast magnetic resonance imaging

**PCA** Principal component analysis

**t-SNE** T-distributed Stochastic neighbor embedding

**MDS** Multi-Dimensional scaling

**isoMDS** Kruskal’s non-metric multi-dimensional scaling

# List of figures

# 

[1 PCA result on the dataset](#_yeypw9i40lqz) …………………………………………. 12

[2 Code snippet for t-SNE](#_ovu85dojbrcf) …………………………………………….. 13

[3 t-SNE projection result on the dataset .](#_bxm4lwvzi86q)........................................ 14

[4 Code snippet for UMAP](#_4nlcwaw4bbyz) …………………………………………….. 15

[5 UMAP projection result on the dataset .](#_eie6ghlrnf8t)....................................... 16

[6 Code snippet for MDS .](#_hilafp4jv01f)................................................................ 17

[7 MDS projection result on the dataset .](#_vr6sxxo3pb9x).........................................18

[8 Results of the user study. The first plot in each row represents the projection result. The second plot represents user rating on cohort separability. The third plot represents user rating on outlier findability from the projection results (for simplicity, we are using some notations, plots in the first row are considered as 1a, 1b, and 1c, and for the second row 2a, 2b, 2c and so on.)](#_myxgtvtjnp42)……………………………20

# List of tables

# 

[Table 1: Quantitative analysis of the projection results.](#_av7bkfzcm7l0) 22

# 

# 1. Introduction

The goal of this project is to differentiate between healthy patients and those suffering from Bicuspid-Aortic-Valve disease (BAV). Bicuspid aortic valve disease is a disease that is related to the aorta. The aorta is the largest and most important artery of the human body. Bicuspid aortic valve disease is related to abnormalities in the aortic valve. This disease makes the heart work harder than it should [12]. Researchers found out that analyzing the relations between hemodynamics and cardiac pathologies could identify the patients suffering from this disease [1].

To solve this task, we have patient-specific blood flow information measured using four-dimensional phase-contrast magnetic resonance imaging, also known as 4D-PC-MRI [1]. This imaging technique allows us to acquire temporally resolved, three-dimensional blood ﬂow information non-invasively. Analysis of such datasets can help us in assessing cardiovascular diseases and their severities. This dataset has around 197 attributes and 122 rows. Each row is a specific instance representing the characteristics of a patient. Attributes can be broadly categorized into five types [1].

1. attributes that are related to the speed of the blood flow are categorized as velocity-related attributes.
2. attributes that are related to the pressure of the blood flow are categorized as pressure-related attributes.
3. attributes that are related to the diameter of the aorta are categorized into diameter-related attributes. This category is for the attributes that relate to blood flow parts that show a vortical behavior.
4. In a healthy person, the blood flows laminar through the vessel, but in pathologically changed aortas, the flow starts to swirl inside the vessel. This is called vortical blood flow, and such attributes are categorized into vortex related.
5. And in the last category, the flow jet describes a part of the flow with the highest speed, which is usually centrally located in the aorta. Due to pathological vessels, changes in the flow jet can be displaced. Such attributes are called flow-jet related attributes.

Our dataset can be classified into three types.

1. Patients who are healthy
2. Patients suffering from bicuspid aortic valve disease (BAV).
3. Patients with a Tetralogy of Fallot.

Our main interest lies in cohorts that are related to healthy patients and patients suffering from BAV. The third category is not our major concern here. This is due to the fact that there are only ten instances of this category. Working with such few data could mislead the results. Hence, we choose not to consider it. We contain data about 90 healthy people and 22 BAV patients in the dataset.

The dataset is high-dimensional with 197 attributes. It is hard to visualize and interpret the structure of the data if the attributes present in the dataset are more than two. Though we can visualize in 3D space, it is hard to interpret the results. So, most of the time, it is preferred to visualize in 2D space. The dataset has 197 dimensions. It is impossible to visualize in 197D space. So, there are some dimensionality reduction techniques proposed to visualize the high dimensional data into just one-, two- or three-dimensional space without losing much information. We have applied several dimensionality reduction techniques starting with Principal Component Analysis (PCA) [6], Uniform Manifold Approximation and Projection (UMAP) [5], T-Stochastic Neighbour Embedding (t-SNE) [3] and Multidimensional Scaling (MDS) [4] to visualize the important information from the data like cohorts, outliers etc. Furthermore, to evaluate the projection results, we performed a qualitative and quantitative analysis of our projections. The qualitative analysis includes user study on 40 subjects, and for quantitative analysis, we measured silhouette and correlation coefficient for the projections.

# 2. Background

Before diving deep into the approach, we provide some background of some important concepts.

## 2.1 Cohorts:

Cohorts are groups of subjects or instances that share some common characteristics at a given time and, analyzing such underlying patterns are called cohort analysis.

## 2.2 Outliers:

Outliers are the data instances that differ significantly from other observations.

High-dimensional data is hard to visualize, and working with high-dimensional data is very complex. In real-life scenarios, most of the available data is highly sparse, and performing such data is desirable. Hence, there is a need to convert such high dimensional data into low dimensions (In simple terms, we are reducing the number of variables in our data) to simplify our search of finding interesting patterns in our dataset. Dimensionality reduction techniques can be broadly classified into two types.

## 2.3 Feature selection:

Feature selection methods [11] are expected to discover a subset of significantly relevant input variables from the given dataset. Feature selection follows three strategies: filter, wrapper, and embedded strategies. Feature selection is not in this project's scope as we are more inspired to perform feature extraction techniques.

## 2.4 Feature extraction:

Feature extraction techniques [11] are also known as feature projection techniques. These techniques aim to convert data from a high-dimensional space to a lower-dimensional space. Feature extraction techniques try to find smaller sets of input variables; these input variables combine input variables in our datasets. These techniques could be either linear or nonlinear in nature.

## 2.5 Principal component analysis:

Principal component analysis also known as PCA is a linear dimensionality reduction technique [6]. It maximizes the variance of the data in lower dimensions. PCA plots the correlation between data variables or instances into a 2D plot, and instances that are having similar characteristics are clustered together, and dissimilar are shown apart. PCA helps remove all correlated features (i.e., all principal components are independent of each other. Hence there is no correlation.), reducing overfitting and improving algorithm performance. The data must be standardized before implementing PCA. Selecting the number of principal components plays an important role, leading to a loss of information.

## 2.6 T-Distributed-Stochastic-Neighbor-Embedding:

It is also known as t-SNE. t-SNE is a non-linear dimensionality reduction technique [3]. This technique tries to model high-dimensional data points into lower-dimensional points so that nearby points model similar data instances, and distance points with some higher probabilities model the points that are not similar. Final outputs of t-SNE are highly dependent or strongly influenced by the parameters chosen. Hence it is very important to have a good understanding of the parameters beforehand. t-SNE can form well-separated clusters.

## 2.7 Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP):

This is also known as Umap. Umap is a non-linear dimensionality reduction technique [5]. It uses a Fuzzy topological structure to perform dimensionality reduction. The model tries to map lower-dimensional points by finding their closest or equivalent fuzzy topological data.

## 2.8 Multi-dimensional scaling:

Multidimensional scaling aims to find patterns by mapping data points/data instances to a 2D plot by analyzing the similarity or dissimilarity between data points/data instances [4]. Various similarity measures can be used like Euclidean distance, cosine similarity, or Manhattan distance.

## 2.9 Silhouette coefficient:

The silhouette coefficient is an evaluation metric that is used to measure the goodness of clustering [7]. The value ranges between –1 and 1. 1 indicates that the clusters are well separated, and –1 indicates there are wrong cluster assignments. It considers both, inter-cluster distance and intra-cluster distance.

## 2.10 Correlation coefficient:

The correlation coefficient indicates if there is a linear correlation between the points within a cluster [9]. The values range between –1 and 1. If the value is close to 0, then there is no linear correlation between the points. One indicated a positive correlation, and –1 indicates a negative correlation.

## 2.11 Connectivity:

Connectivity indicates to what extent the points are connected. It uses a k-nearest neighbor approach to calculate the connectivity value [8]. Value range between 0 to infinity. If the value is close to zero, then we can say that the points have good connectivity. So, values close to zero are better.

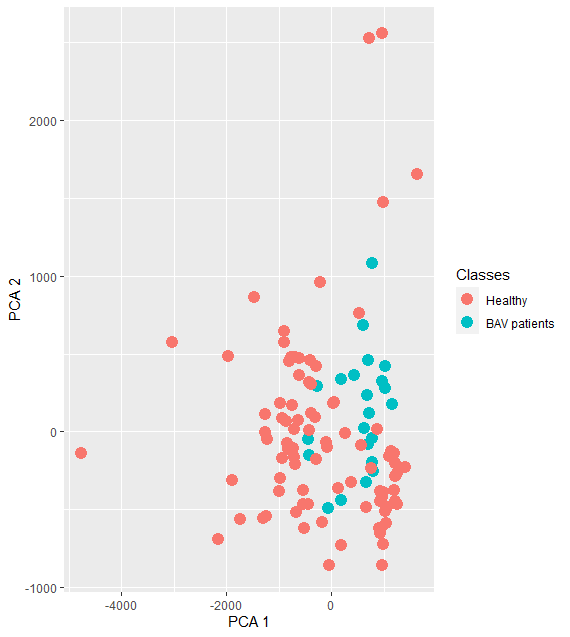
These cluster evaluation techniques(silhouette coefficient, correlation coefficient, connectivity) are chosen to quantitatively estimate the goodness of the projection result because they were reliable and easy to understand.

# 3. Approach

For the implementation, we have used R as our primary programming tool.

## 3.1 Principal component analysis (PCA):

The principal component analysis is the first method that we have used for visualizing the structure of the data. For implementing the PCA [6], we have used a method called "prcomp". It takes data in the form of a matrix and outputs the principal components. Out of many principal components starting, two components contain most of the variance. We consider the first two principal components for visualization, which covers 49% variance (PCA1- 38%, PCA2- 11%). We could also have considered PCA3

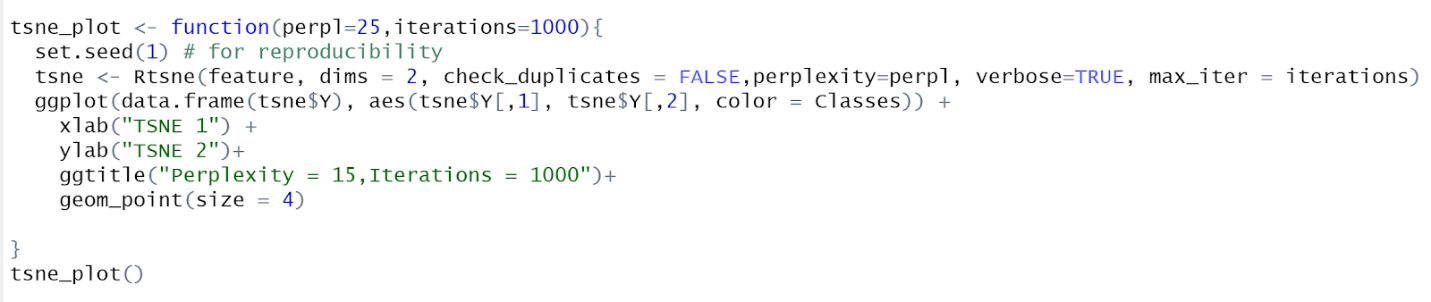


**Figure 1: PCA result on the dataset**

but, there are two major drawbacks. First, the visualization will be in 3D, which is hard to interpret, and second, PCA 3 shows only a 0.8% variance, which is not very useful. From Figure1, we can observe that there are two classes (cohorts) which are healthy and BAV patients. From the plot, we cannot find a clear separation between the two cohorts. So, we can understand that the structure of the data is not represented well by the PCA. The reason could be that the data might be nonlinear in nature. PCA works pretty well for linear data. When it comes to nonlinear data, it struggles. Based on this assumption, we decided to go with other dimensionality reduction techniques, which work quite well with nonlinear data.

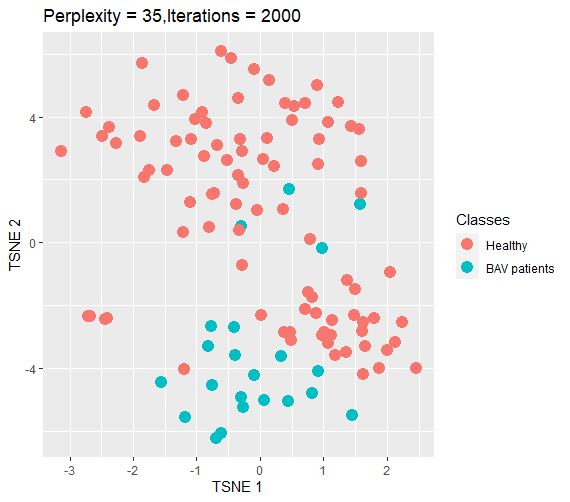
## 3.2 T-distributed stochastic neighbor embedding (t-SNE):

Many nonlinear dimensionality reduction techniques exist in the literature. Firstly, we have tried using a popular dimensionality reduction technique called T-distributed stochastic neighbor embedding, shortly known as t-SNE [3]. t-SNE tries to map the data from high dimensional space to low dimensional space without losing much information. A package called "Rtsne" can be used to compute t-SNE on our data. From Figure 2,



**Figure 2. Code snippet for t-SNE**

we can see that the Rtsne method takes many arguments. However, the important ones are input which is "feature" in Figure 2, perplexity, which represents the total number of neighborhood data points to consider. Based on our analysis, higher values are better to preserve global geometry and max\_iter, representing the maximum number of iterations the method should run. There are some limitations to the t-SNE algorithm. Every time we run the method with our dataset, we get different results. There is not much of a difference between different runs, but the result will not be the same. So, we set a seed at the beginning of the method "set.seed(1)". Here 1 is a random number. We can replace it with any number. This can be used to reproduce the results every time we run the method.

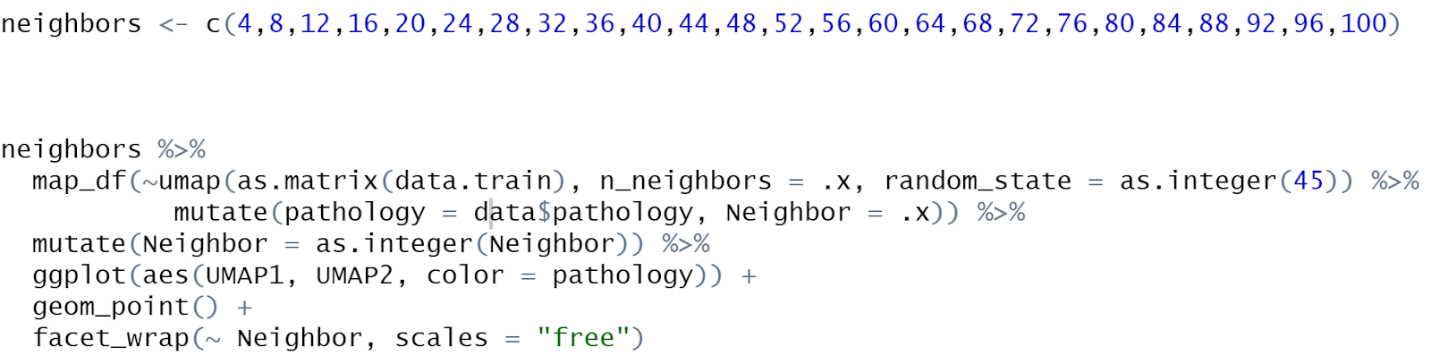


**Figure 3: t-SNE projection result on the dataset.**

The projection result of the t-SNE is shown in Figure 3 with parameters (perplexity = 35, iteration = 2000). The results were pretty good compared to PCA. The cohorts of healthy and BAV patients are not very well separated, but we can distinguish them. We can see that there are two groups formed inside the healthy people's cohort from the projection. Also, we can see four outliers from BAV patients, which falls under healthy people. We can see outliers in healthy people as well. There are three outliers that are very far from the two groups of healthy people. We can make some reasonable assumptions from the plot. However, this is just one parameter combination. The problem here is, we cannot say that this parameter combination (perplexity = 30, max\_iter = 2000) of t-SNE works best for every dataset. We tried and tested different parameter combinations and chose the best ones based on qualitative and quantitative analysis.

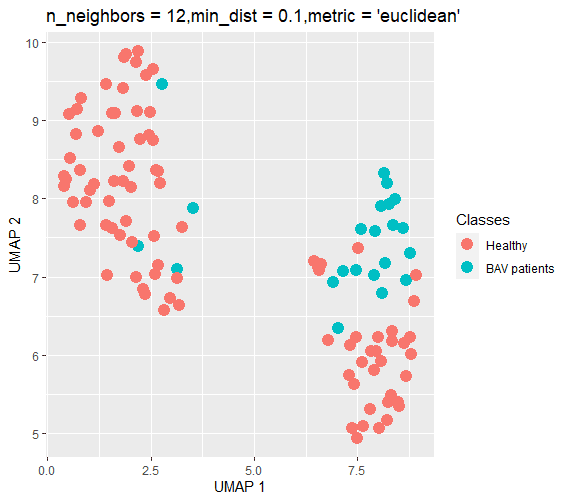
## 3.3 Uniform Manifold Approximation and Projection (UMAP):

One projection technique may not work for every dataset. So, there is a necessity to try and test with other nonlinear dimensionality reduction techniques. Next, we chose UMAP [5]. There are some advantages for UMAP when compared to t-SNE. UMAP is good at preserving the global structure, whereas t-SNE makes clusters of different sizes into the same size in the final projection. In other words, it makes small clusters look big, and significant clusters look small. And the processing time of the UMAP is faster compared to t-SNE.



**Figure 4: Code snippet for UMAP**

There is a library called "umapr", which can be used to implement UMAP in R. There is a method called "umap" in "umapr", which takes arguments like data, n\_neighbors. Here the data is passed in the form of a matrix. The n\_neighbors balance the local and global structure of the data. Lower n\_neighbors values will force UMAP to concentrate on the data's local structure, whereas higher n\_neighbors values will force UMAP to concentrate on the global structure of the data. From Figure 4, we can see that there is a list containing different values for n\_neighbors. We have plotted the results of many UMAP plots with different n\_neighbour values. There is also one more critical parameter called min\_dist. Just like n\_neighbours we have plotted the UMAP results for different min\_dist values by keeping n\_neighbor as constant. We have filtered the plots that are showing the best results, both qualitatively and quantitatively from all the plots.

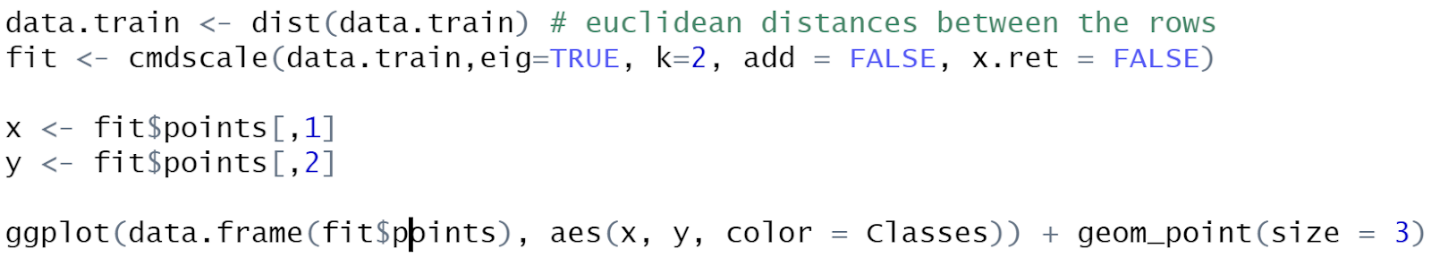


**Figure 5: UMAP projection result on the dataset.**

From the plot shown in Figure 5, we can see a clear separation in the cohorts. From the t-SNE plot, we have assumed that there are two groups within a healthy cohort. This plot clearly shows that, and also, we could see the four outliers in the BAV patients. Both the t-SNE plot (shown in Figure 3) and UMAP plot (shown in Figure 5) share complementary results. Nevertheless still, UMAP has some disadvantages, just like t-SNE. Unlike PCA, the axis does not convey any meaning, and the projection results will also change when we run the same UMAP method multiple times. However, a random state parameter in UMAP takes an arbitrary number as input, which can be used to reproduce the results.

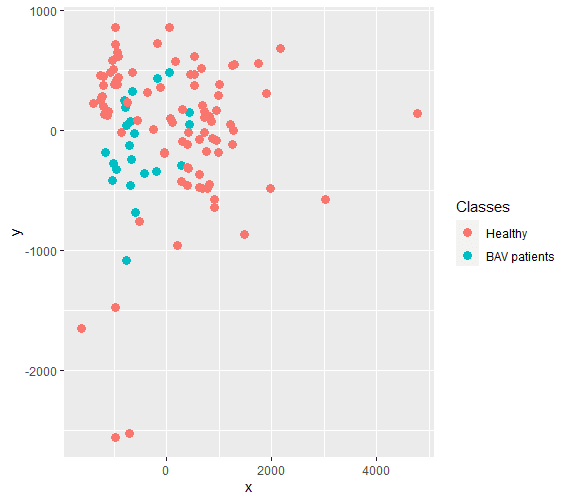
## 3.4 Multidimensional Scaling (MDS):

Multidimensional Scaling transforms the data from high dimensional space to low dimensional space such that the relative distance between the data points are preserved [4].



**Figure 6: Code snippet for MDS.**

For the implementation of Multidimensional scaling in R, we need a library called "Mass". This library contains many classic Multidimensional scaling functions, Kruskal's Non-Metric Multidimensional Scaling and Sammon's nonlinear mapping. The input to the classic Multidimensional scaling function (cmdscale) is the distance matrix of the input data and eig = True returns the eigenvalues. The k value represents the dimensions of the output. In our case, we output a 2D graph so, k = 2.



**Figure 7: MDS projection result on the dataset.**

Multidimensional scaling is not quite suited for this dataset. From Figure 7, we can observe that the cohort separability is not so good when compared to t-SNE or UMAP. The outliers of the BAV patients found in the previous methods were not found clearly in MDS. We have also tried other types of Multidimensional scaling (Sammon, isoMDS), but the results were not so good compared to the previous techniques. However, we can reproduce the results, which is not the case with UMAP and t-SNE and the computation time of the MDS is also comparatively low.

# 4. Evaluation

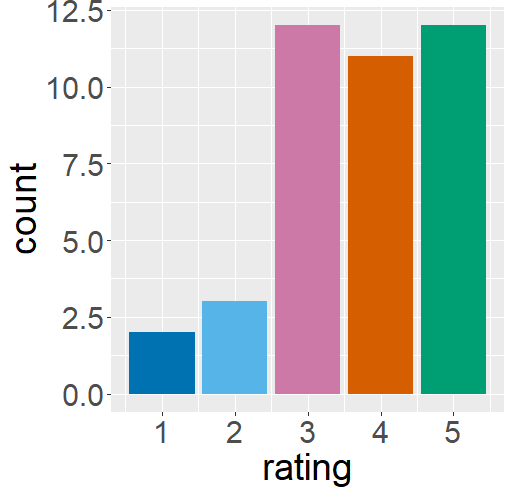
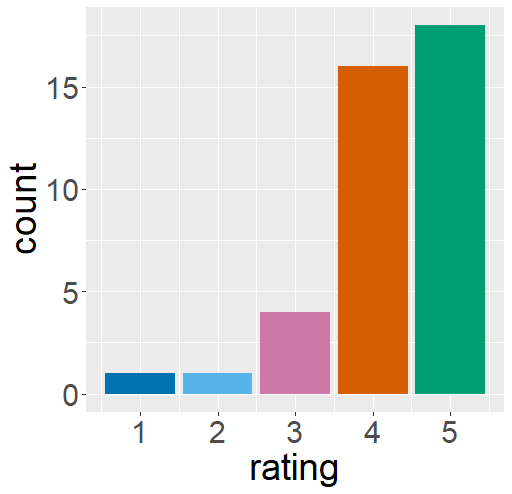
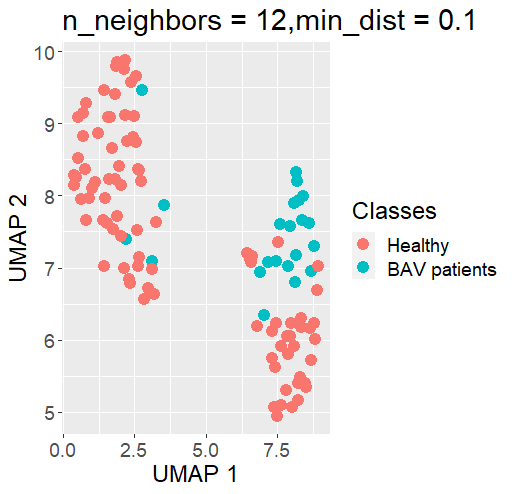
## 4.1 Qualitative analysis:

We evaluated the results, both qualitatively and quantitatively. This strategy is adopted from [10]. For qualitative analysis, a user study is performed on 40 subjects who are familiar with clustering and visualization. Out of 40 subjects, there were 12 females and 28 males. The age of the subjects is between 25-40 years. For each dimensionality reduction technique, we tested with different parameters and picked some plots as representatives. This selection of the plots is performed based on the diversity of the plots. The user study consists of two questions.

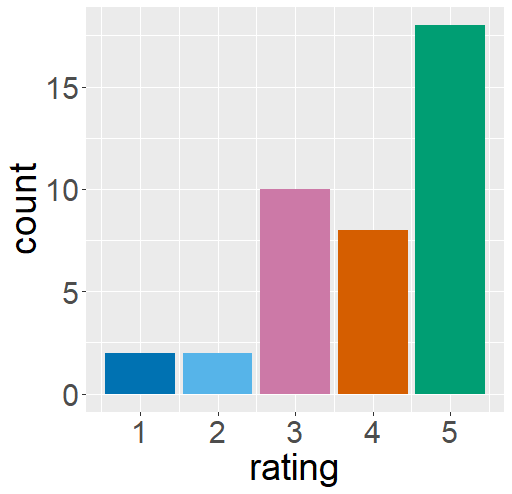
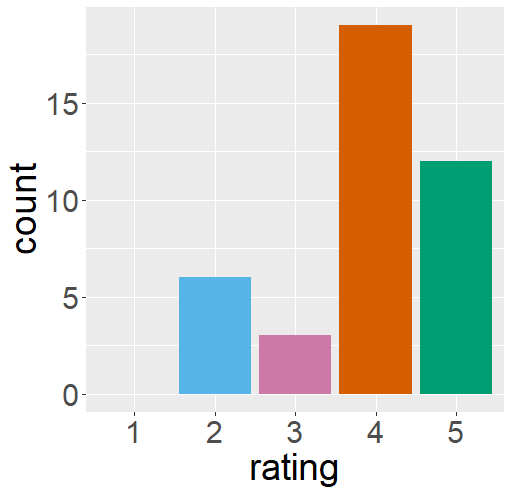
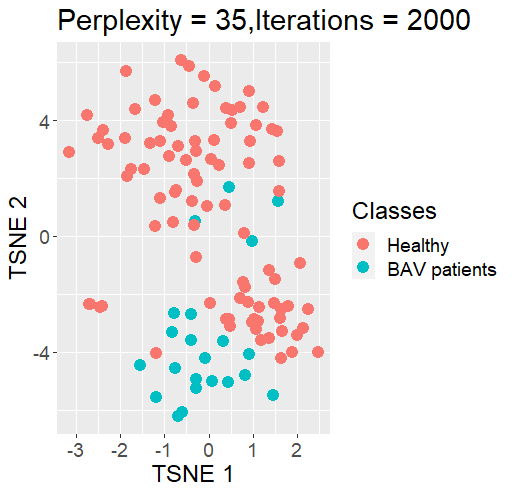
1. How well can you visually separate the cohorts?

2. How well can you identify outliers?

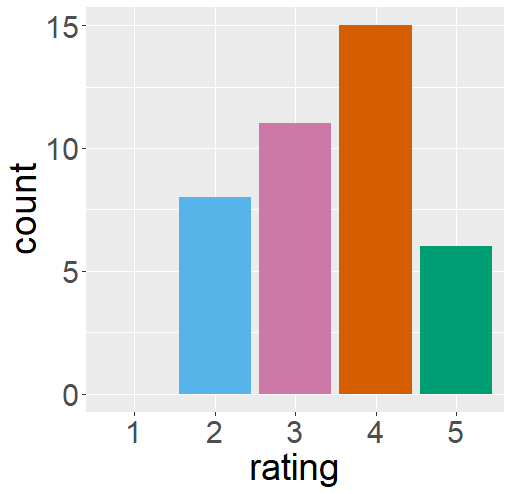
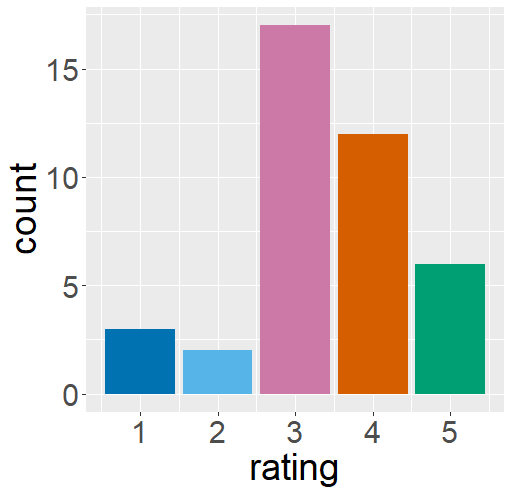
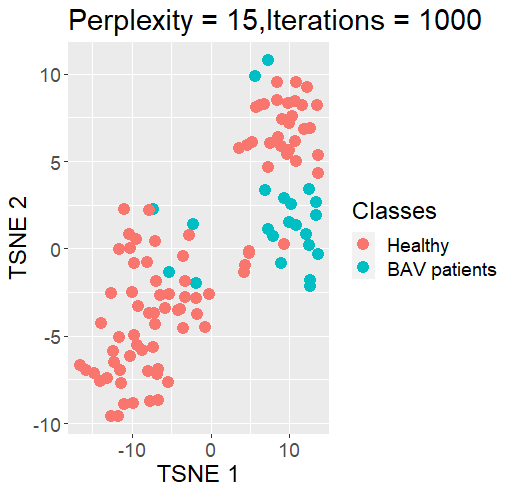
For each plot, these two questions need to be answered by the subjects using a Likert scale whose values are between 1 and 5. These responses are later used for analysis.



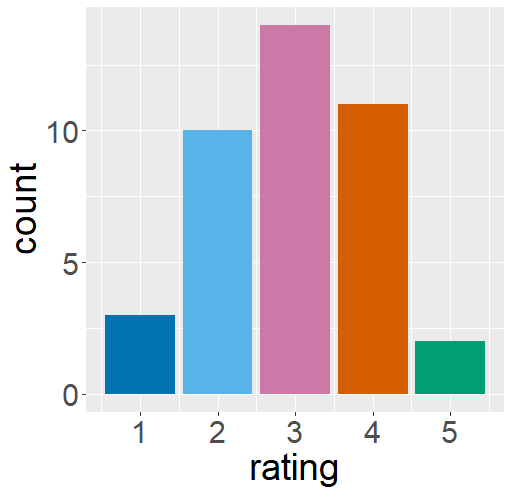
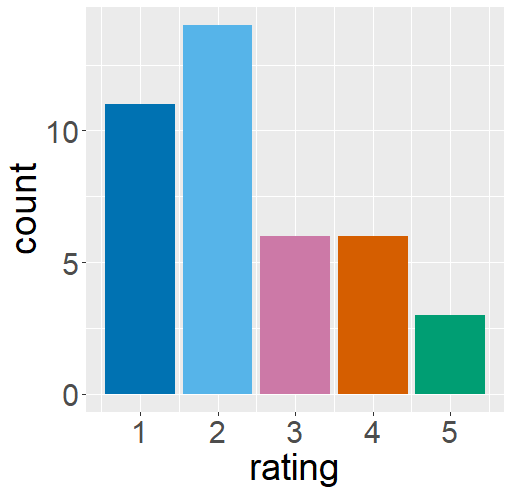
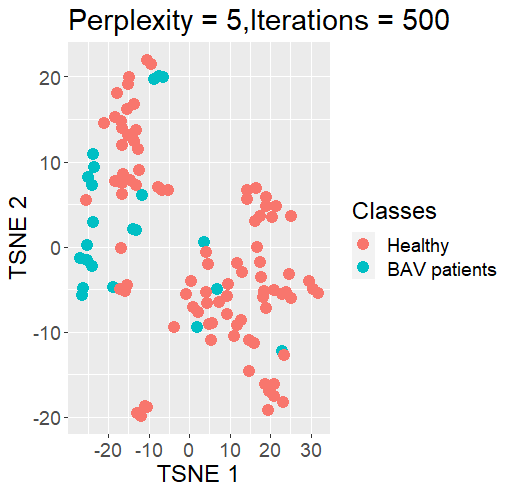
(1a) (1b) (1c)



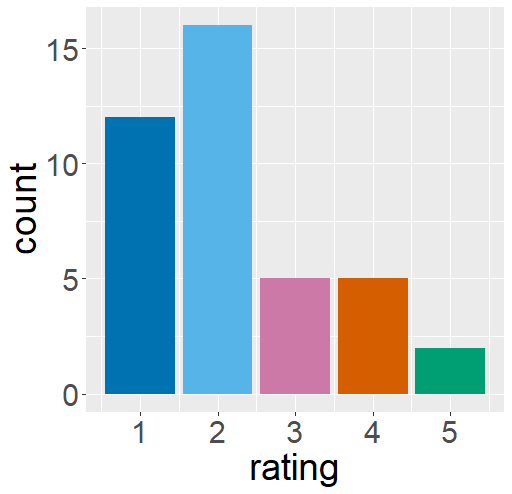
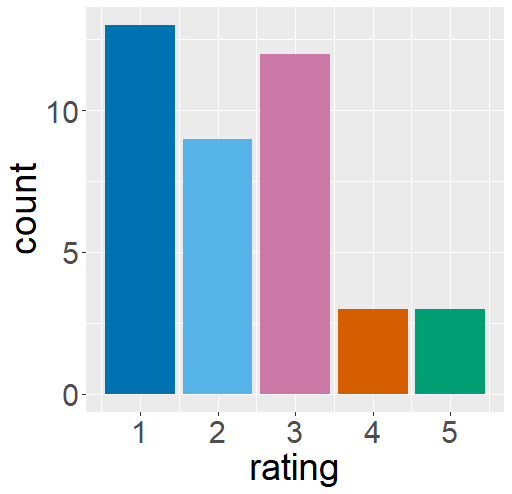
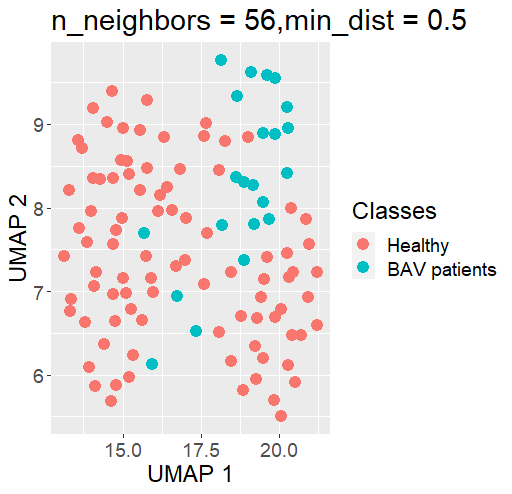
(2a) (2b) (2c)



(3a) (3b) (3c)



(4a) (4b) (4c)



(5a) (5b) (5c)

**Figure 8: Results of the user study. The first plot in each row represents the projection result. The second plot represents user rating on cohort separability. The third plot represents user rating on outlier findability from the projection results (for simplicity, we are using some notations, plots in the first row are considered as 1a, 1b, and 1c, and for the second row 2a, 2b, 2c and so on.)**

Figure 8 illustrates the responses of 40 subjects on five plots. The plots are arranged in a particular order (from best to worst). From Figure 8, we can observe that plot 1 and 2 show a clear separation between the cohorts, and outliers were also easily identified so, the subjects have given a high rating to these plots. Similarly, plots 4 and 5 receive poor ratings because of their poor cohort separability.

## 4.2 Quatitative analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| **Plots** | **Silhouette coefficient** | **Correlation coefficient** | **Connectivity** |
| Plot 1 | 0.128 | -0.70 | 30.41 |
| Plot 2 | 0.186 | -0.278 | 25.16 |
| Plot 3 | 0.113 | 0.771 | 29.75 |
| Plot 4 | 0.12 | -0.475 | 33.33 |
| Plot 5 | 0.094 | -0.05 | 28.5 |

**Table 1: Quantitative analysis of the projection results**

Later, we have considered the Silhouette coefficient, Correlation coefficient, and Connectivity as evaluation metrics for quantitative analysis. Based on the plots shown in Table 1, the Silhouette scores show that plot 5 has a poor Silhouette score because of its poor cohort separability. Next, we have used the Pearson Correlation coefficient, which measures linear dependencies between two variables. Plot 1 and 3 have better values which indicate that points within a cluster are correlated. The Connectivity validity measure used the k-nearest neighbor approach, and the value should be close to 0. Plot 2 shows a better Connectivity value. Based on quantitative analysis, plots 4 and 5 shows poor results on all three measures. The remaining plots are good at least for one measure.

# 5. Conclusion

In this project, we have worked with cardiac blood flow data, which is a high-dimensional dataset. Visualizing the structure and finding important information about the data is not an easy task. To achieve our goal, which is to distinguish between healthy people and BAV patients, we have used several dimensionality reduction techniques to visualize the structure of the data based on the features. Firstly, we tried with PCA, which doesn't give good results so, we assumed that the data might be nonlinear in nature, so we tried with different nonlinear dimensionality reduction techniques like t-SNE, UMAP and MDS. Then, the results were better. We evaluated our results, both qualitatively and quantitatively. For qualitative analysis, we performed a user study on 40 subjects, and we selected the best plots based on the responses of the subjects. Then, we performed quantitative analysis on the selected plots using the Silhouette coefficient, Correlation coefficient and Connectivity measure. Based on these measures, the best plots were selected.

# 6. References

[1] Köhler, Benjamin & Grothoff, Matthias & Gutberlet, Matthias & Preim, Bernhard. (2019). Bloodline: A system for the guided analysis of cardiac 4D PC-MRI data. Computers & Graphics. 82. 10.1016/j.cag.2019.05.004.

[2] G Hinton, S Roweis. “Stochastic neighbor embedding”, *Advances in neural information processing systems* 15, 833-840.

[3] L van der Maaten, G Hinton. “Visualizing data using t-SNE”, *The Journal of Machine Learning Research* 9 (2579-2605), 85, 2008.

[4] M Williams, T Munzner. „Steerable, progressive multidimensional scaling“, *Proc. of IEEE Symposium on Information Visualization*, pp. 57-64, 2004.

[5] Leland McInnes, John Healy, James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, arXiv:1802.03426

[6] Tharwat, Alaa. (2016). Principal component analysis - a tutorial. International Journal of Applied Pattern Recognition. 3. 197. 10.1504/IJAPR.2016.079733.

[7] S. Aranganayagi and K. Thangavel, "Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), Sivakasi, India, 2007, pp. 13-17, doi: 10.1109/ICCIMA.2007.328.

[8] Sriparna Saha, Sanghamitra Bandyopadhyay, "Some connectivity based cluster validity indices”, Applied Soft Computing, volume 12, Issue 5, 2012, pages 1555-1565,

[9] "Correlation coefficient”. Wikipedia, Wikimedia Foundation, 9 March 2021, https://en.wikipedia.org/wiki/Correlation\_coefficient

[10] Etemadpour, Ronak & Motta, Robson & Paiva, José Gustavo & Minghim, Rosane & Oliveira, Maria Cristina & Linsen, Lars. (2015). Perception-based Evaluation of Projection Methods for Multidimensional Data Visualization.

[11] "Dimensionality reduction”. Wikipedia, Wikimedia Foundation, 17 February 2021, <https://en.wikipedia.org/wiki/Dimensionality_reduction>.

[12] Richard N. Fogoros, MD. "Symptoms and Complications of Aortic Aneurysm” Verywellhealth, November 20, 2019, https://www.verywellhealth.com/aortic-aneurysm-symptoms-and-complications-4160769