Interactive Visual Clustering and Classification based on Dimensionality Reduction Mappings: A Case Study for Analyzing Patients with Dermatologic Conditions

M.A. Mohedano-Munoz[a], S. Alique-García[c,b,a], M. Rubio-Sánchez[a], L. Raya[d], A. Sanchez[a,e,*]

[a] *Universidad Rey Juan Carlos (Madrid, Spain)*
[b] *Hospital Virgen de la Luz (Cuenca, Spain)*
[c] *Hospital Universitario de Fuenlabrada (Madrid, Spain)*
[d] *U-tad (Madrid, Spain)*
[e] *Research Center for Computational Simulation (Madrid, Spain)*

Abstract

Multidimensional data sets are becoming more frequent in practically all research fields, and require complex data analysis techniques in order to extract knowledge from them. In this paper, we propose an interactive visualization tool for performing exploratory data analysis. The tool combines unsupervised and supervised dimensionality reduction methods, such as linear discriminant analysis, or t-SNE, with clustering and classification techniques. Analysts can use several machine learning methods for extracting data structure, and can group data into clusters interactively or through clustering algorithms. In addition they can visualize projections of the data to evaluate the quality obtained clusters, and to analyze the performance of classification methods. We have applied this tool to analyze a clinical data set related to patients with dermatologic conditions that are under photodynamic therapy. The analysis allowed medical doctors to identify several clinically interesting patient groups. In addition, clinicians discovered a greater

*Corresponding author
Email addresses: miguel.munoz@urjc.es (M.A. Mohedano-Munoz), seralique@gmail.com (S. Alique-García), manuel.rubio@urjc.es (M. Rubio-Sánchez), laura.raya@u-tad.com (L. Raya), alberto.sanchez@urjc.es (A. Sanchez)

efficacy in the treatment of patients with the photosensitizer 5-aminolaevulinic acid nanoemulsion gel compared to those treated with methyl-5-aminolaevulinate cream.

Keywords: Dimensionality Reduction, Clustering, Classification, Visual Analytics, Dermatology, Photodynamic Therapy

1. Introduction

The generation of multidimensional data sets, with a large number of observations and attributes, is increasingly common in a wide range of research fields such as healthcare, demography, or economics. The great complexity of these data sets has made it necessary in many cases to use sophisticated data analysis methods in order to extract knowledge from them.

A common issue when working with high-dimensional data is the well-known curse of dimensionality. It is related to the sparsity of the data as the dimensionality of the data increases (Bellman, 1957), and negatively affects the performance of data analysis and machine learning (ML) methods. One of the main strategies for tackling this problem is to pre-process the data in order to reduce its dimensionality (Friedman et al., 2001). Many dimensionality reduction (DR) methods have been proposed in the literature, and can focus on different goals such as preserving the structure of the data or maximizing class separation (van der Maaten et al., 2009). These methods define either linear or nonlinear mappings from the high-dimensional data space onto a low-dimensional space. In general, although nonlinear mappings are more powerful in the sense that they can represent data more faithfully, it is usually difficult to understand the role of the original features in the mapping. Alternatively, while linear mappings are simpler, it is possible to use visualization techniques to depict information about the features, which can be used to obtain insight regarding how they affect nonlinear mappings.

When the data is reduced to three or less features it is possible to visualize the transformed data in a Cartesian coordinate system. These data visualizations, which constitute a corner stone of exploratory data analysis, allow analysts to combine their domain knowledge with their ability to visually understand relationships and properties of the data. Moreover, it is acknowledged that user interaction is often essential for visual analytics, since it facilitates and speeds up tasks related to formulating hypotheses, making decisions, or drawing conclusions. Thus, interactive visualizations can facilitate and guide data analysis processes, and carry out diverse tasks such as exploration, feature selection, clustering, etc.

This paper describes an exploratory data analysis tool where users can visualize and interact with several two-dimensional plots of data that has been transformed through a DR method. Specifically, the tool implements (but is not limited to) linear methods like Principal Component Analysis (PCA) (Jolliffe & Cadima, 2016), Linear Discriminant Analysis (LDA) (McLachlan, 2004), and Locality Preserving Projections (LPP) (He, 2005), as well as nonlinear techniques such as Locally Linear Embedding (LLE) (Roweis & Saul, 2000), Multidimensional Scaling (MDS) (Cox & Cox, 2000), t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). Among other analysis tasks, the tool allows users to: i) obtain insight regarding the structure of the data through the DR methods; ii) understand the importance of features on diverse tasks; iii) partition the data, either with an automatic clustering algorithm or manually through the interactive components of the tool; and/or iv) build a classifier.

To the best of our knowledge, it is the only interactive tool that couples clustering, classification, and several DR techniques simultaneously in the same interface. Domain experts can perform different exploratory data analysis tasks combining supervised and unsupervised methods to work with both labeled and unlabeled data, which are common in medical records. In particular, we demonstrate the usefulness of our tool through a case study analyzing a data set related to dermatologic conditions. Clinicians have used the tool, together with our support, to obtain insight about how different photosensitizers behave, find interesting patient groups, learn the discriminatory power of the features to classify patients, and discover a greater efficacy of photosensitizer 5-aminolaevulinic acid nanoemulsion (ALA) compared to methyl-5-aminolaevulinate (MAL).

This paper is organized as follows. Section 2 analyzes related visualization tools. Section 3 describes the proposed interactive interface, the algorithms employed to perform DR, and how users can interact with the data and mappings to carry out data analysis tasks. Section 4 presents a specific data analysis pipeline as a case study analyzing dermatological patients under photodynamic therapy. Lastly, in Section 5 we draw the main conclusions and propose future work.

2. Related Work

Multidimensional visualization tools differ in how to transform the data into visual representations, and how analysts can interact with the visualizations. In this section we first describe several visual interfaces that employ DR mappings to visualize the data in a two or three-dimensional space. Subsequently, we present other visual tools that have been designed specifically for clustering or classification.

2.1. Visual Interfaces for Dimensionality Reduction

Multiple visual interfaces have been developed that incorporate different methods of DR to analyze multidimensional data. Most of them focus on a single method (Jeong et al., 2009; Molchanov & Linsen, 2014) or rarely allow users to select one technique among several possible options. For instance, iPCA (Jeong et al., 2009) is an interactive data analysis system that offers the possibility of adjusting data items, deleting data elements, or modifying the dimension contribution, but is limited only to PCA mappings. Analogously, Molchanov et al. (2015) propose to use PCA as a first step of an interactive supervised classification task of medical images. iVisClassifier (Choo et al., 2010) uses LDA linked with other visualizations, such as parallel coordinates (Inselberg & Dimsdale, 1987) and heat maps, to analyze class separation. This drives a visual analysis pipeline where users can ultimately compare the distance between two data instances on the heat map and manually alter their labels (classes). Our proposal also provides an analysis pipeline but does not suffer from the space limitations associated with parallel coordinates when working with a large number of features, or with the use of a single DR technique. Alternatively, our tool relies on a flexible and interactive coordination of views of different linear and nonlinear DR algorithms.

The tool developed by Turkay et al. (2011) also uses coordinated visualizations of various mappings of multidimensional data (e.g., PCA, MDS, or LDA). It allows users to focus on instances of interest to extract information and to cluster the data. However, although user interaction is a central part of the exploratory data analysis process (Liu et al., 2017), it only offers a limited set of interactive operations. In addition, it does not support classification tasks.

A recent review of the state of the art by Sacha et al. (2017) points out a few visual interactive interfaces (Mao et al., 2007; Rieck & Leitte, 2015; Nam & Mueller, 2013; Liu et al., 2014) that enable the selection of various DR methods. None of these tools are designed for classification tasks, and only the last two can be used for clustering. For instance, Persistent Homology (Rieck & Leitte, 2015) allows analysts to compare different DR methods, such as PCA, t-SNE or LLE, by ranking them according to several proposed quality measures. However, in contrast with our proposal, the tool does not provide functionality to carry out data analysis tasks.

Choo et al. (2013) present a tool for clustering and knowledge discovery where analysts can group unlabeled data or generate visualizations composed of several coordinated (i.e., linked) DR plots. Nonlinear DR methods (especially t-SNE) have been used as starting points for clustering (Gisbrecht et al., 2013). The process applies algorithms based on the Dunn index on representations in $\mathbb{R}^2$. Our proposal goes in the same direction but also allows analysts to interactively validate and modify the obtained clusters. In addition, it enables users to create pipelines that can employ the information obtained from nonlinear transformations, e.g. clusters, to try to provide meaning to the resulting groupings with the aid of linear DR mappings. In any case, in contrast to our proposal, these tools and many others (Mao et al., 2007; Bradel et al., 2014; Molchanov & Linsen, 2014) do not support classification.

2.2. Visual Interfaces for Data Analysis

In this section we describe several visual interfaces that focus mainly on clustering and classification, but do not rely on DR methods.

The Hierarchical Clustering Explorer (Seo & Shneiderman, 2002) is an early proposal to use visualization tools in clustering tasks. It uses heat map visualizations to group gene expression by building hierarchical clustering trees. Another work proposes a framework to visualize clusters in a variant of parallel coordinates that is designed to reduce clutter (Zhou et al., 2008). ClusterSculptor is an interactive application for finding cluster hierarchies that are represented in radial dendrograms. Specifically, it is based on k-means and visualizations of inherent characteristics of the high-dimensional data (Nam et al., 2007). Clustervision (Kwon et al., 2017) allows users to use different projections to understand the cluster structure. In addition, analysts can visualize the groupings obtained by automatic clustering algorithms, and update their hyperparameters interactively. Likewise, Lai et al. (2018) propose a method to discern and refine clusters trough projections based on the analysts' points of interest, which can be a single instance or a cluster of data. Several projections can be analyzed by observing different features of the points of interest, while maintaining the rest as context.

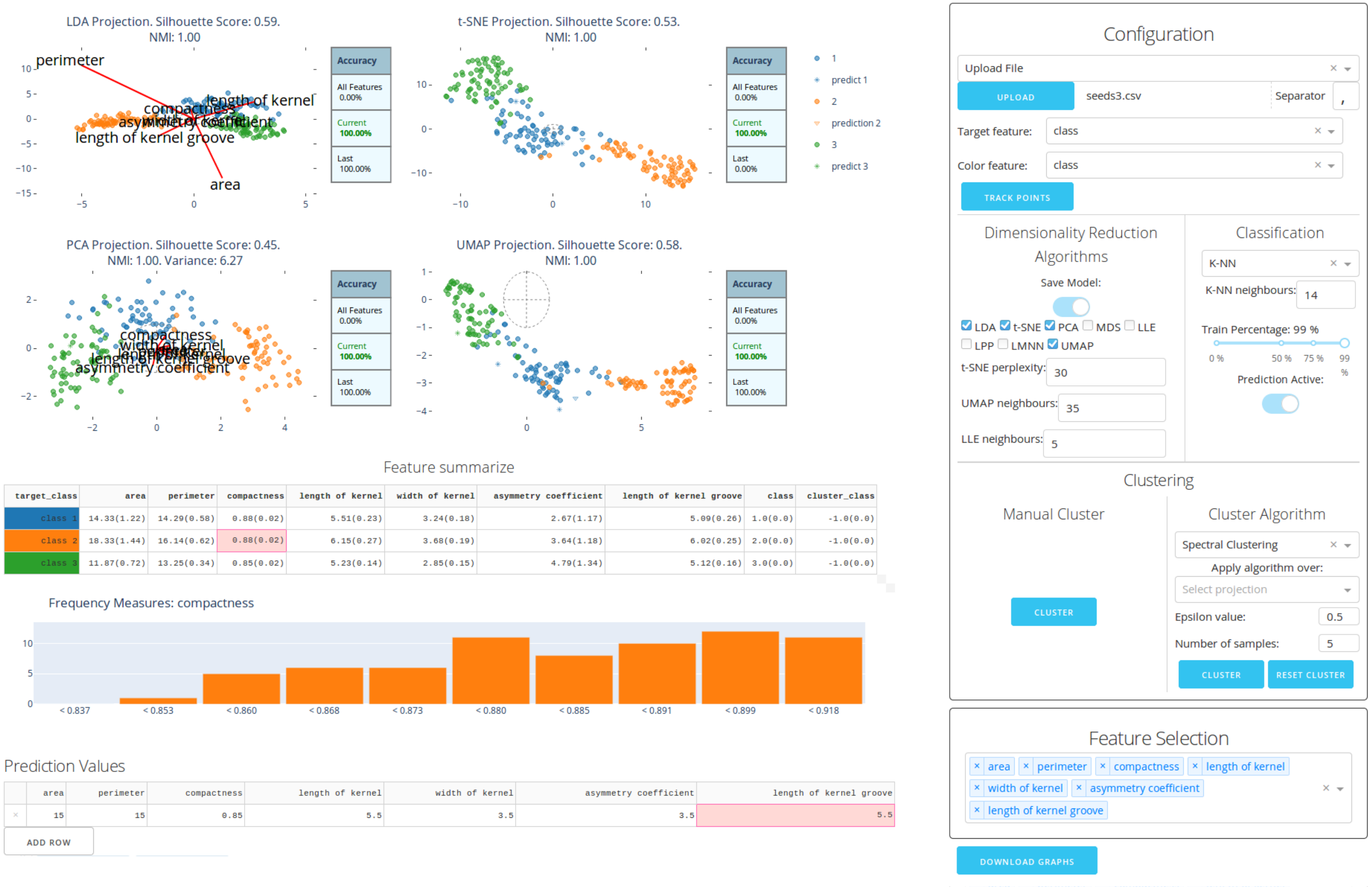
Clustrophile (Demiralp, 2016) was a first approach to coordinate the representation of scatter plots with discrete heat maps for evaluating clusters. This research work has evolved into the recently published Clustrophile 2 (Cavallo & Demiralp, 2019), a user-guided clustering tool that allows analysts to evaluate the quality of the resulting clusters. It is a powerful tool that incorporates numerous clustering methods, but our proposal allows users to additionally group the data interactively according to their own perception. Furthermore, our tool not only focuses on clustering, but also uses the obtained clusters to feed supervised ML methods.

Regarding classification, visualizations are seldom combined with ML algorithms. In general, automatic classifiers do not include information related to the users' domain knowledge. Instead, the few existing interactive data visualization tools for classification attempt to incorporate the users' expertise into the learning process (Choo et al., 2010). The particular interactive approaches depend on the ML technique employed. For instance, there are some visually guided classification methods (Ankerst et al., 1999; Teoh & Ma, 2003) where analysts can explore visualizations to split the data set and manually build a decision tree through a recursive process. The process described in Andrienko et al. (2009) first clusters and labels two-dimensional geographic points and trajectories, and subsequently builds a distance-based classifier according to the clustering partition. Our tool allows analysts to create a similar pipeline, but is not limited to two-dimensional data sets, since it incorporates DR methods.

3. Interactive Interface for Data Analysis

Our proposal aims to provide an interactive tool for knowledge discovery, where users can perform several data analysis tasks with the help of visualizations. Our tool, which we call DRCC (Dimensionality Reduction for interactive visual Clustering and Classification), combines DR mappings, unsupervised and supervised ML methods, and user interaction. In addition, it provides the usual tasks (overview, zoom and filter, details-on-demand) of the well-known Information Seeking Mantra (Shneiderman, 1996; Heer & Shneiderman, 2012). During an exploratory analysis, users can filter by category, get detailed information using the hover tool, hide/show classes by clicking on their legend entries, zoom in/out, select entries to highlight them or group them to create a cluster, export the obtained views, etc. To facilitate user interaction, the different views of the interface are linked and synchronized. Thus, any change in configuration parameters of DR or ML methods affects all active views.

We have developed our tool in Python, which is a well-known high-level programming language widely used in data analysis. The interface uses the package pandas to manage data structures, scikit-learn to perform dimensionality reduction, clustering and classification, Plotly to build interactive graphics, and Dash to manage interactions between the graphical elements and to create the web interface. Since the application will be freely available, visualization experts will be able to extend it to contain new features or adapt existing ones to their needs.



Feature summarize

target_class	area	perimeter	compactness	length of kernel	width of kernel	asymmetry coefficient	length of kernel groove	class	cluster_class
class 1	14.33(1.22)	14.29(0.58)	0.88(0.02)	5.51(0.23)	3.24(0.18)	2.67(1.17)	5.09(0.26)	1.0(0.0)	-1.0(0.0)
class 2	18.33(1.44)	16.14(0.62)	0.88(0.02)	6.15(0.27)	3.68(0.19)	3.64(1.18)	6.02(0.25)	2.0(0.0)	-1.0(0.0)
class 3	11.87(0.72)	13.25(0.34)	0.85(0.02)	5.23(0.14)	2.85(0.15)	4.79(1.34)	5.12(0.16)	3.0(0.0)	-1.0(0.0)

Prediction Values

	area	perimeter	compactness	length of kernel	width of kernel	asymmetry coefficient	length of kernel groove
×	15	15	0.85	5.5	3.5	3.5	5.5

ADD ROW

seeds3.csv

Separator ,

Target feature: class

Color feature: class

☑ LDA ☑ t-SNE ☑ PCA □ MDS □ LLE

□ LPP □ LMNN ☑ UMAP

t-SNE perplexity: 30

UMAP neighbours: 35

LLE neighbours: 5

K-NN

K-NN neighbours: 14

Spectral Clustering

Apply algorithm over:

Epsilon value: 0.5

Number of samples: 5

× area × perimeter × compactness × length of kernel × width of kernel × asymmetry coefficient × length of kernel groove

Figure 1: DRCC interface. On the left it shows different types of dimensionality reduction plots, together with a histogram and an interactive table with descriptive statistics. The panels on the right allow users to select hyperparameters (related to dimensionality reduction and clustering algorithms) and to choose a subset of variables to work with.

Figure 1 shows the interface of DRCC. The mappings in the top left (the group of four plots in the example) correspond to different DR methods. The interface is reactive to changes in the data set or the configuration parameters (right dialog box), which is crucial for exploring the data set efficiently. In addition, it is possible to recover information from the data samples quickly (in particular, attribute values, and predicted and true classes) by hovering the mouse over the projected points.

By using the coordinated views, users can obtain useful information about the cluster structure of the data. For example, users can verify if the points that appear to form a cluster in one mapping also form a cluster in the rest. This would support the claim that the points do indeed constitute a cluster in the high-dimensional space. In addition, analysts can highlight individual points, or select subsets of points by using a lasso tool. These groupings can be labeled and used in subsequent classification tasks.

The interface also allows users to filter the data according to the class labels (i.e., add or remove entire classes from the visualizations) by means of an interactive legend. This is useful when comparing a few classes, since users can filter out irrelevant data points that would otherwise introduce noise.

Data elements are represented by different symbols that indicate whether they belong to the training or test set for classification. Furthermore, users can visualize a measure of cluster quality that is related to the classification error of the projected points, where the labels are associated with the cluster groupings. The visualizations also use different symbols for the points to indicate whether they are classified correctly or incorrectly. Concretely, on the right side of each mapping DRCC shows a table describing the classification accuracy values obtained for the particular projection. Lastly, above each plot the interface shows interesting clustering and classification measures that can be used to compare the mappings.

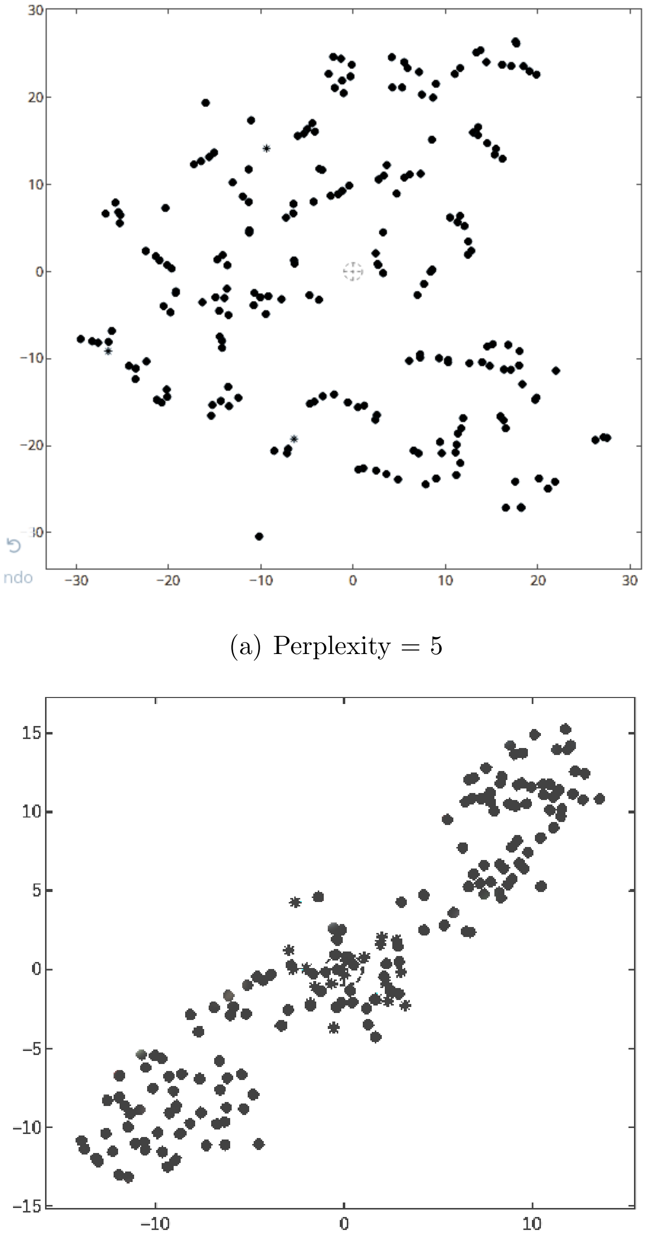
Underneath the projections we have included an interactive table of basic descriptive statistics that are useful for characterizing the groups of data instances and determining their homogeneity. Specifically, it shows the mean and standard deviation of each data attribute (columns) in each of the groups (rows). In addition, by clicking on a particular cell the interface shows the data distribution of its associated attribute and group through a histogram. Finally, at the bottom of the interface we can display another interactive table that we use for classification prediction (see Section 3.1.4).

The panels on the right allow users to choose a data set (which can be stored in several formats), select DR methods, define a clustering process, configure classification options, and select the features to use. Once the data file is loaded users can select the variable that contains the class labels that will be used as targets in the classification process. For unlabeled data sets users create a new variable C that can later be used to indicate cluster membership. The projected points can be colored according to these variables to show the classes or clusters, or to any other data variable (for example, to observe the distribution of the attribute values).

Analysts can then choose different DR methods, their hyperpameters (a tooltip explains briefly each of the DR methods and their goal), and the features to work with (through the panel at the bottom right). Since the projections are visualized simultaneously in different plots, users will be able to compare and relate the mappings in order to better understand the structure of the data. For example, analysts can obtain information from linear methods that can be useful for understanding more sophisticated nonlinear mappings (see Section 3.1.3).

3.1. Data Analysis Tasks

In this section we describe several data analysis tasks that can be performed with our tool.



(a) Perplexity = 5

(b) Perplexity = 40

Figure 2: t-SNE mappings of the Seeds data set for different perplexity values. We can appreciate the presence of clusters.

3.1.1. Extracting data structure through nonlinear mappings

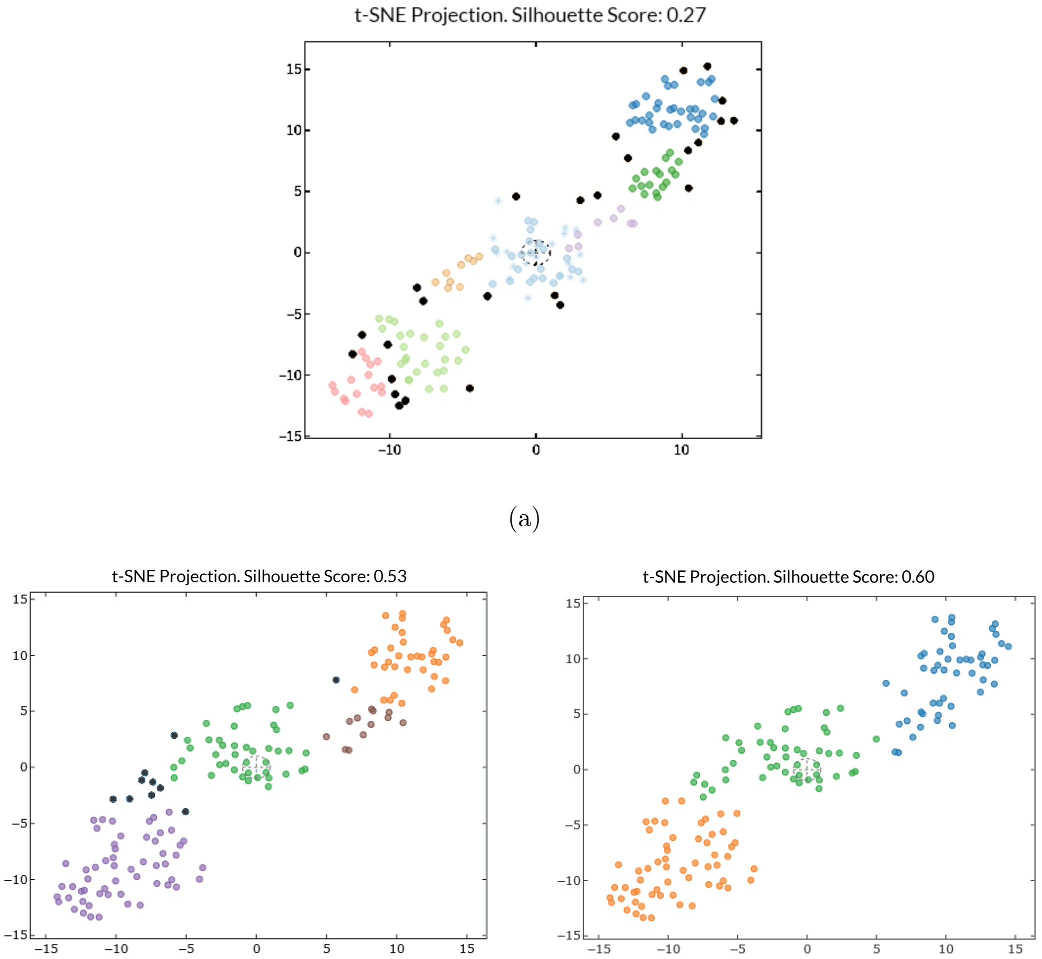
Our tool currently implements several nonlinear DR methods such as LLE, MDS, t-SNE and UMAP, but it can be easily extended to incorporate other techniques. Nonlinear DR methods can outperform linear ones regarding the ability to represent the structure of the data (e.g., nonlinear manifolds whose intrinsic dimensionality is two, or cluster structure) more faithfully. In addition, some unsupervised nonlinear methods (e.g., t-SNE or UMAP, which group data elements according to their similarity), despite not using information about class labels, often produce plots in which the classes appear better separated than in mappings generated through linear methods, such as LDA, which do make use of class membership information (Rubio-Sánchez et al., 2017) (see Section 3.1.3). Figure 2 shows two t-SNE mappings of the Seeds data set, available at the UCI Machine Learning Repository (Dheeru & Karra Taniskidou, 2017). The data is obtained from the analysis of three varieties of wheat (classes), 210 balanced instances of seven features (‘area’, ‘perimeter’, ‘compactness’, ‘length of kernel’, ‘width of kernel’, ‘asymmetry coefficient’, and ‘length of kernel groove’). The t-SNE method does not use the class variable (‘wheat’), but we can observe the presence of a few groups in the plots, which could be associated with the classes. When working with t-SNE, analysts must experiment with several perplexity parameters, since they can lead to different plots. For example, in Figure 2 it is easier to distinguish three groups in the plot associated with a perplexity value of 40. Our tool allows users to modify hyperparameters interactively in order to obtain different views of the data, which could reveal different properties of the data.

3.1.2. Clustering

If analysts suspect the presence of clusters in one of the two-dimensioal plots shown in the interface, they can either apply an unsupervised clustering algorithm (DBSCAN, spectral clustering, etc.) or trust their perception and domain knowledge to group the data. In case they want to use a clustering algorithm they simply have to select the specific DR representation and set the parameters of the clustering algorithm. This will form different groups of data elements whose membership will be encoded in a new clustering variable C. Alternatively, users can create their own clusters by relying on the DR plots. In order to create a cluster they first select a subset of points by using a lasso tool, and then click on the 'Cluster' button. This creates a new cluster label for the selected data elements, which is specified in the variable C. Lastly, they can color the projected points in all of the DR mappings according to the found clusters.

When users cluster the data visually they rely exclusively on two-dimensional information. In this regard, automatic clustering methods can use the original high-dimensional data, or the two-dimensional projected points. In our tool we run the automatic algorithms on the projected data mainly for efficiency, which is important in interactive applications. Currently the tool can cluster data automatically through two DBSCAN (Ester et al., 1996) or Spectral Clustering (Ng et al., 2001). It is important to note that both do not include points in clusters if they interpret that they correspond to noisy samples.

Figure 3 shows several ways to cluster the data according to the t-SNE mapping in Figure 2(b). Firstly, users can employ automatic clustering algorithms. Figures 3(a) and (b) show the result of applying Spectral Clustering and DBSCAN, respectively. Spectral Clustering determines the presence of seven clusters, while DBSCAN detects only 4 using as hyperparameters $\epsilon = 3.0$ and $samples = 20$. Alternatively, in (c) the particular clustering is specified by the user through the lasso tool, by considering the arrangement of the projected points, possible domain knowledge, and information gathered from the plots in (a) and (b). In this regard, the clusters obtained through an automatic clustering algorithm can be refined through user interaction. For example, the clustering in (c) stems from modifying the one in (b). In particular, we have chosen three clusters by assigning a cluster label to the points that were considered to be noisy samples, and by merging the orange and brown clusters. Moreover, in our tool all of the mappings are linked. Thus, when users select and highlight a set of points in one mapping, these also


t-SNE Projection. Silhouette Score: 0.27

t-SNE Projection. Silhouette Score: 0.53

t-SNE Projection. Silhouette Score: 0.60

(a)

(b)

(c)

Figure 3: Clustering process over t-SNE mapping on Seeds data set seen in Figure 2. Black points are considered as noise. (a) Spectral Clustering. The silhouette scores 0.27. (b) DBSCAN clustering with hyperparameters $\epsilon = 3.0$ and $samples = 20$. The silhouette scores 0.53. (c) Own clustering refining (b). The silhouette increases to 0.60.

appear highlighted in the rest of the projections. This allows analysts to use information from all of the mappings when assigning cluster labels. Lastly, in this example all of the points belong to some cluster. However, users can also define unlabeled points for which the cluster membership is not clear.

Finally, we have implemented the silhouette score (Rousseeuw, 1987) to interpret and evaluate cluster separation, with the aim of allowing analysts to compare different clustering processes. It measures the similarity of a sample to its own cluster, compared to the similarity to other clusters. The silhouette score can be between -1 and 1, where 1 indicates a perfect class separation.

3.1.3. Feature importance in linear mappings

Linear mappings from $\mathbb{R}^n$ to $\mathbb{R}^2$ are characterized by $2 \times n$ matrices. Most software libraries that compute linear mappings (such as PCA or LDA) will not only provide the coordinates of the projected points, but also the transformation matrix. Since each of its columns is a two-dimensional vector it is possible to represent these vectors together with the projected points. This is the main principle of Star Coordinates (SC) (Kandogan, 2000), which is a visualization technique based on radial axes that simply defines a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^2$, and shows both the projected points and the two-dimensional "axis" vectors (with the same origin point) of the transformation matrix in a single graphic.

Since each of the axis vectors is associated with a data variable, we can analyze their lengths and orientations to gain insight about their role and importance in the DR mapping. In particular, the vectors point towards directions in which the original attribute values of the projected points generally increase. In addition, longer vectors have a greater influence on the visualizations (Rubio-Sánchez et al., 2016; Wang et al., 2017), although their orientation can also influence their importance (Sanchez et al., 2018).

Figure 4 shows the SC plot associated with the LDA mapping of the Seeds data

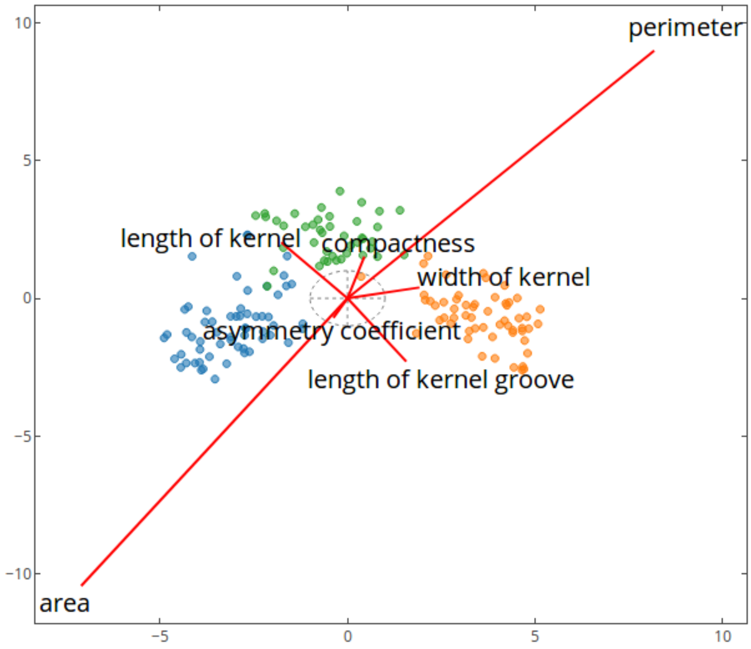


Figure 4: SC plot associated with the LDA mapping of the Seeds data set, where the labels were obtained from the visual clustering process related to Figure 3(c). The length and orientation of the axis vectors indicate the relevance of the features on the mapping (Sanchez et al., 2018). The axis vectors ‘perimeter’ and ‘area’ are clearly larger than the rest. Thus, those variables are the most relevant for separating the three groups.

set. Since LDA is a supervised method we have used the cluster labels identified in Figure 3(c) as the class variable. The plot not only shows the projected points, but also the axis vectors (as red line segments), together with their associated variable names. Since the goal of LDA consists of separating the three groups, the longer axis vectors indicate which variables are likely to be the most relevant for distinguishing the groups (i.e., the most discriminant features). In this case, the axis vectors ‘perimeter’ and ‘area’ are clearly larger than the rest, and therefore correspond to the most relevant variables for separating the three groups. However, other features such as ‘length of kernel’ or ‘length of kernel groove’ are also important for separating the orange and green groups. Lastly, the blue group has larger values of ‘area’ and smaller values of ‘perimeter’. This can also be verified through the interactive table that allows users to select and visualize histograms of particular variables for specific groups of data.

In the previous example we have determined the most relevant variables for a given linear mapping. In general, it is difficult to understand the role the variables in nonlinear mappings. However, note that the classes used in the LDA projection were obtained through the t-SNE nonlinear mapping (the original data does not contain class labels). If the classes are separable in the LDA plot we infer that the most important variables identified in that visualization are also relevant for the t-SNE plot. In this regard, a linear mapping has helped us to understand relevant features for a nonlinear projection.

3.1.4. Classification

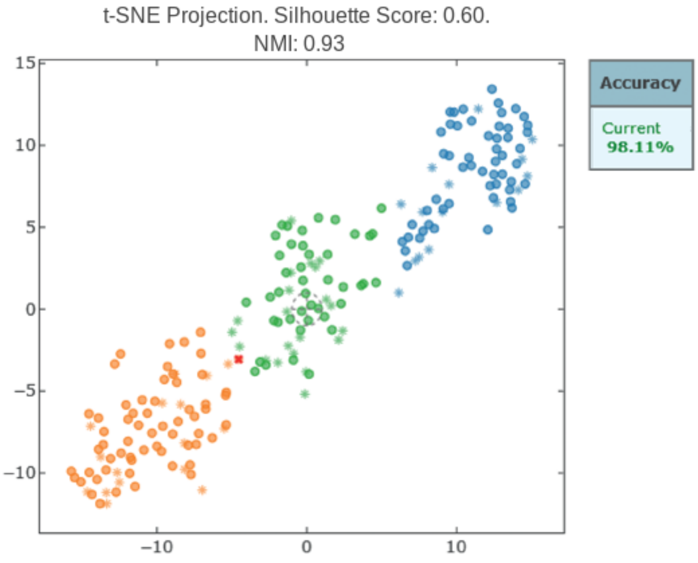
The tool also allows analysts to work with labeled data. The labels can be part of the data set, they can stem from a clustering algorithm, or they can be created by the user interactively through the process described in Section 3.1.2. In these cases users can obtain insight regarding the performance of classifiers for specific DR mappings. Firstly, the projected points in the plots can be colored according to the different classes or cluster labels in the data. This allows users to visually estimate the degree of overlap between the different groups, and therefore to assess the difficulty to classify correctly, given a particular set of features. This idea can be used to perform feature selection, since analysts can add or remove data variables to determine which ones help to separate the groups, and which simply add noise.

In addition, users can build classifiers. In this case users first divide the data set into a train and test subsets randomly, where the size percentages of these subsets are specified through a slider. Analysts can then choose the classification algorithm, set its parameters, and predict the classes of new data points. The interactive table at the bottom left of the interface allows users to specify the new point to classify, which are represented in each mapping through a triangle colored according to the predicted class. For some dimensionality reduction algorithms the two-dimensional coordinates of the new points can be obtained by simply applying the model (e.g., linear methods that provide the projection matrix) to the new point. Instead, other nonlinear methods may require implementing out-of-sample approaches (Pezzotti et al., 2016b). Analysts can then evaluate and compare different classifiers according to their accuracy on the test set (i.e., the proportion of true predictions among the total number of cases tested). Lastly, note that some clustering algorithms do not necessarily assign labels to all of the data samples (e.g., if they consider that samples are noisy). In these cases the classification algorithms simply ignore the unlabeled data.

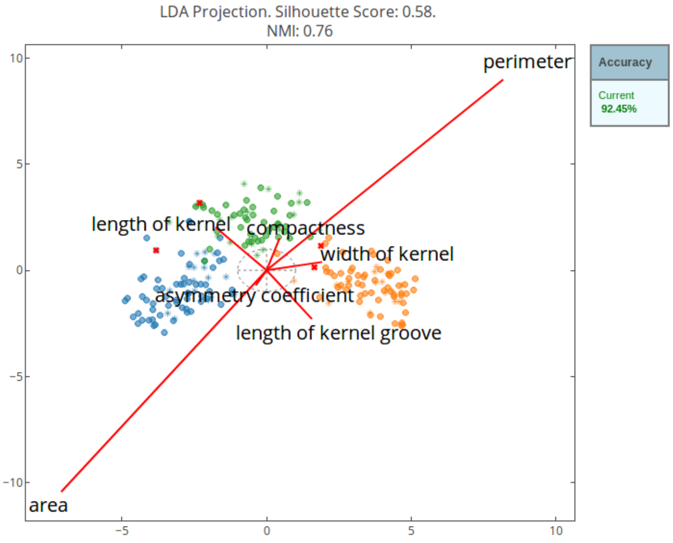
The interface shows both training and test sets by displaying the projected points through different shapes (circles for training data and star diamonds for test). In addition, the test samples that are correctly classified are presented by the color of their corresponding class, while the incorrectly classified samples are displayed in red. This allows users to observe the prediction errors pre-attentively.

We have currently implemented the K-NN classifier (Altman, 1992; Duda et al., 2001), which is based on distances. Specifically, the predicted class for a given sample is obtained through a majority vote of its k closest training samples. We have used $k = \sqrt{N}$ by default where N is the number of samples in the training set (Dasarathy, 1991). However, users can modify this value in order to analyze how it affects the classification accuracy. This is important since there is no rule for selecting a specific k that achieves optimal results for most data sets (Hassanat et al., 2014).

Figure 5 shows an example of the elements of the training (75%) and test (25%) sets for the t-SNE and LDA mappings in Figures 3(c) and 4, respectively. Note



(a)



(b)

Figure 5: Visual validation through a K-NN classifier of the classes in the t-SNE (a) and LDA (b) mappings of Figure 3(c) and Figure 4, respectively. The plots show the training data (75%) through circles, and the correctly and incorrectly classified test samples with star diamonds and red crosses, respectively.

that the cluster labels are chosen interactively by the user (they are not the labels included in the original data set). The panel to the right of the plots indicates the accuracy of the K-NN classifier for the default value of k, which in this case is 15 (since the data set contains $N = 210$ samples). Specifically, the classification accuracies are 98.11% and 92.45% for the t-SNE and LDA mappings, respectively.

We have also implemented the Normalized Mutual Information (NMI) (Strehl & Ghosh, 2003), which is another measure to compare sets of labels that constitutes an alternative to classification accuracy. Specifically, it returns a value in $[0, 1]$, where 0 indicates no mutual information between the sets of labels, and 1 signifies perfect correlation. Analysts can use these measures to compare different clusterings, but can also rely on the visualizations to observe cluster separation. In the example of Figure 5 t-SNE outperforms LDA.

4. Case Study

In this section, we describe a case study for analyzing a complex medical data set using DRCC. Specifically, we use data obtained from cancer patients under photodynamic therapy (PDT), which is a widely used within the field of dermatology. It consists of the topical administration of a photosensitizer, which accumulates selectively in certain cells or tissues, so that, when illuminated, in the presence of oxygen, with a light of adequate wavelength and in sufficient dose, produces the photooxidation of biological materials and the subsequent cancerogenous cell death (Fritsch & Ruzicka, 2006; Babilas et al., 2005). Currently, 5-aminolaevulinic acid nanoemulsion gel (BF-200 ALA, Ameluz®) and methyl-5-aminolaevulinate cream (MAL, Metvix®) are the two most employed photosensitizers in Europe and USA. PDT has been shown to be effective in the treatment of non-melanoma neoplasms, including actinic keratosis (Reinhold, 2017), Bowen's disease (Tarstedt et al., 2016; Alique-García et al., 2019a), and superficial basal

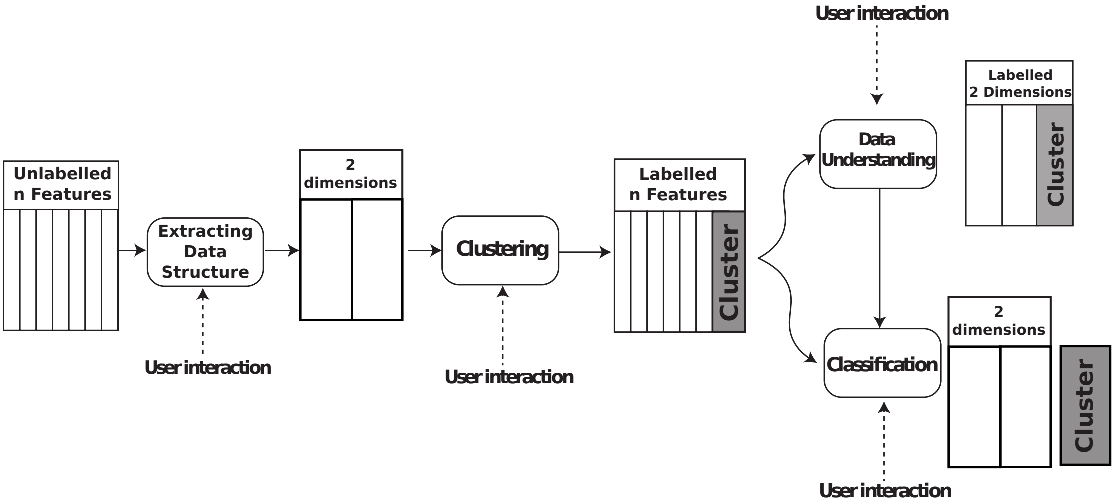


Figure 6: Proposed pipeline to carry out an exploratory data analysis during the case study. The interaction between users and the interface is essential in our approach and occurs in every step of the process.

cell carcinoma (Fernández-Guarino et al., 2014; Alique-García et al., 2019b). In addition, promising results have been obtained in recent years in the treatment of other tumors, inflammatory or infectious pathologies, as well as in cosmetic treatments (Gilaberte et al., 2006; Park et al., 2013).

The data was provided by the dermatologic PDT specialized section of Hospital Universitario de Fuenlabrada (HUF) in Madrid, Spain. It was collected by reviewing the Electronic Health Records of all of the patients (1225) treated with PDT during the period from June 2009 to June 2018. The data set contains information relative to patients' demographic features (age, gender, and skin phototype), diagnosis and lesions location, technical features of treatments (employed ALA or MAL photosensitizer, number of sessions, average and total dose, and incubation times), post-treatment events (tolerance, side effects, specific medical cares), clinical response to PDT (initial response, recurrence, and final healing rate) and "peri-treatment" measures (pre-treatment, support treatment, treatment of recurrences, and follow-up). Each patient is described by a total of 31 features, where each one is a PDT event. Note that in our study we treat patients with BF-200 ALA, which differs from the formulation in Tarstedt et al. (2016) that uses 5-aminolevulinic acid (Aminolevulinic acid hydrochloride 20% in standard ointment, Unguentum Merck).

This case study combines all the functionalities described in the previous section to assemble them in the data analysis pipeline shown in Figure 6. The goal consists of identifying interesting groups of patients and determining which features are relevant in the PDT treatment. The analysis was carried out and driven by clinicians belonging to the dermatology unit of HUF. We provided explanations to them about the use of DRCC, as well as assistance throughout the data analysis process.

The first step is to comprehend the structure of the data using a DR algorithm. For this purpose, clinicians preferred to use a nonlinear DR method to try to observe cluster structure in the data. Specifically, they used an UMAP mapping (McInnes et al., 2018) with 35 neighbours and default values for the rest of the parameters. They also discarded the healing feature (together with response and relapse, since these two attributes are correlated with healing) in order to avoid its information when forming groups. Thus, they did not use any feature as a class label.

The second step is to cluster the data set. In this case the clinicians applied DBSCAN over the UMAP mapping, which is shown in Figure 7. DBSCAN identifies the seven groups (silhouette score: 0.56) illustrated through different colors, where initially unlabeled patients marked as noise by DBSCAN were assigned to the group of their nearest labeled neighbor.

The clinicians then used linear dimensionality reduction methods to try to understand and characterize the obtained clusters and the importance of the features. Specifically, they applied LDA to separate and describe the clusters as shown in

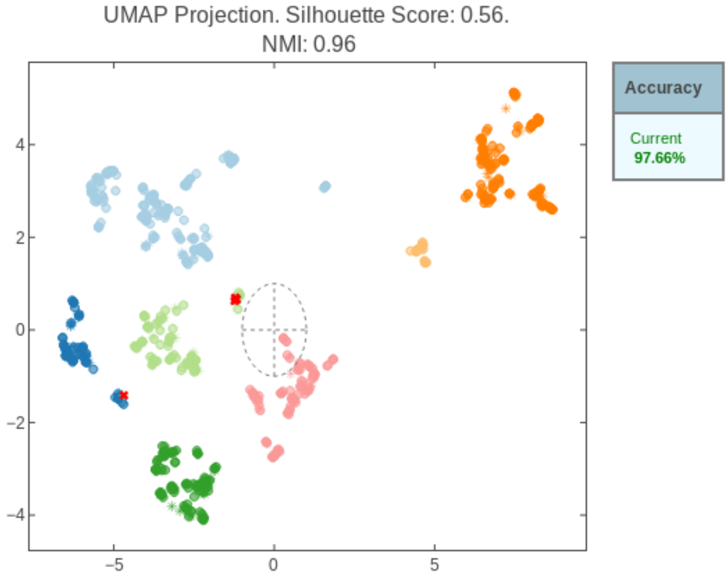
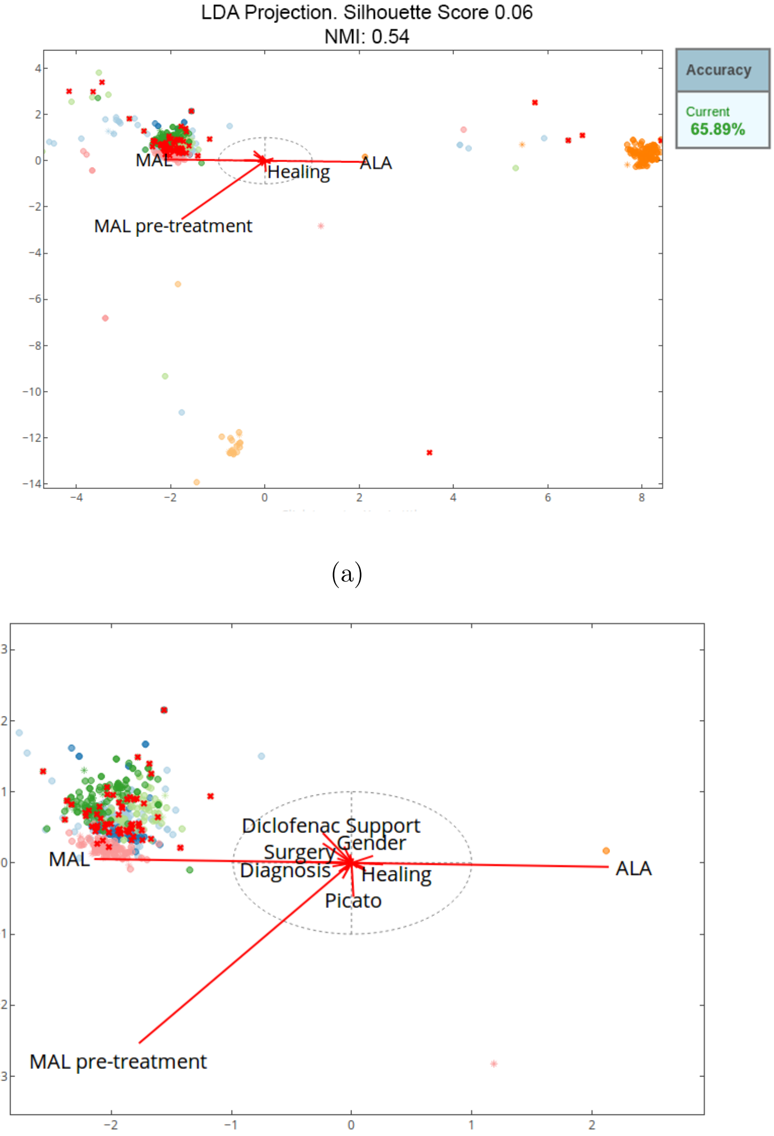


Figure 7: Clusters resulting from applying DBSCAN over the UMAP mapping. The algorithm detects 7 different clusters (group 0 in orange, 1 pink, 2 green, 3 light green, 4 light blue, 5 navy, and 6 yellow).

Figure 8, where the plot in (b) is simply a zoomed-in version of (a). In this case the class labels correspond to the seven cluster groups obtained through DBSCAN and UMAP. It is worth noting that in this case we included the healing feature when computing the mapping. The LDA plot shows the configuration of axis vectors, which provides information regarding the importance of the features for characterizing the cluster (patient) groups. The most discriminative features for this data set are the photosensitizers applied for the PDT (ALA and MAL), and MAL pre-treatment. Other features such as support treatments, diagnosis, gender, and healing play a minor role (the axis vectors of the rest of the features are imperceptible).

The clinicians then characterized all patient groups by examining the interactive table containing their means and standard deviations, and the corresponding



(a)

(b)

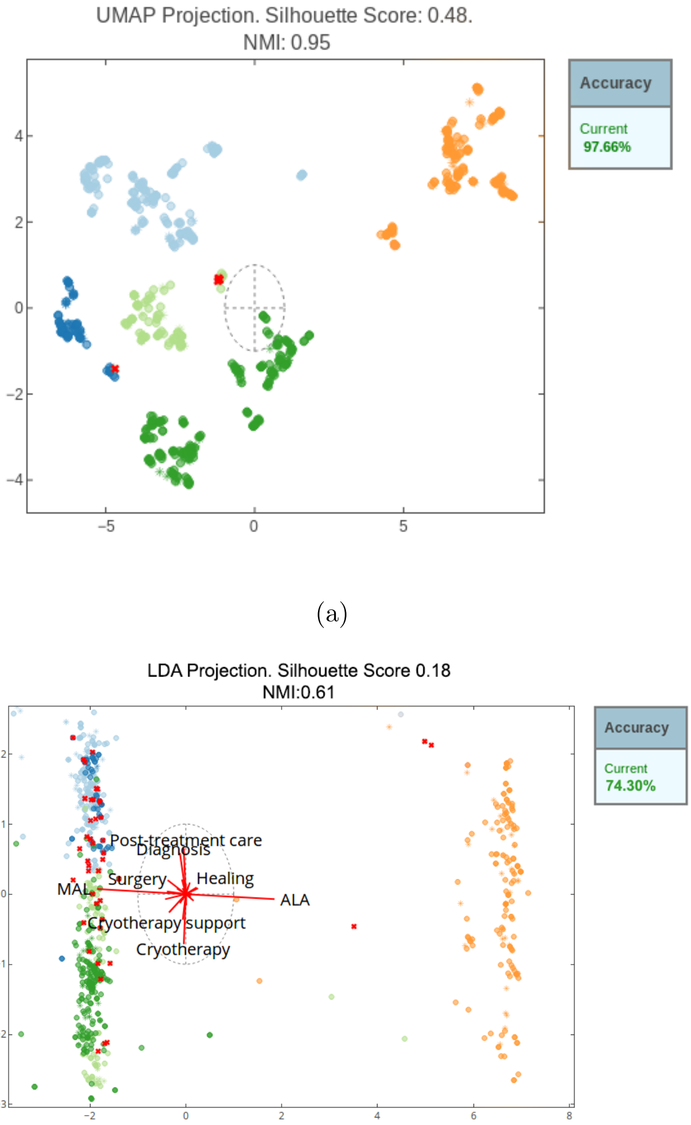
Figure 8: LDA mapping for characterizing patient groups through the length and orientation of the feature axis vectors. The class labels used by LDA correspond to the seven clusters obtained by applying DBSCAN over the UMAP mapping (see Figure 7). The plot also includes results of a K-NN classification ($k = 29$), where the red crosses are classification errors. The plot in (b) is simply a zoomed-in version of (a).

frequency histograms. The description of the clusters is as follows: group 0 (orange) is composed of patients treated mostly with ALA that obtained an average healing rate of 89%. It is a heterogeneous group in terms of diagnosis. Nevertheless, the proportions of patients with actinic keratosis, Bowen's disease, superficial basal cell carcinoma, and other diagnoses, are similar to the ones across the entire data set. There are also no relevant differences for this group in terms of gender. Group 6 (yellow) is very similar to the previous one. The majority of these patients were treated with ALA (the average healing response was 90%), and did not exhibit a predominance of a specific diagnosis. Regarding differences, the percentage of males was higher and, interestingly, about 90% of the patients were previously treated with MAL. In other words, it is a group of patients that were treated at first with MAL but did not heal, and were subsequently rescued with ALA with very good clinical response. Given the similarity between groups 0 and 6 the clinicians decided to join them in a single group that we will call group 7 in the remainder of the paper. The rest of the groups are composed mostly (95-100%) of patients treated with the MAL photosensitizer. In groups 1 (pink) and 2 (green) we find patients diagnosed with actinic keratosis (98% group 1, 90% group 2), which are homogeneous with respect to the rest of their characteristics, except for the previous application of cryotherapy (<1% group 1; >99% group 2), which results in very different healing rates (10% in group 1 and 41% in group 2). Since these percentages are both low from a clinical point of view, the medical doctors decided to merge both groups, creating group 8. Groups 3 (light green), 4 (light blue) and 5 (navy) are very similar to each other, except for gender. Thus, they are patients treated with MAL, without predominance of any particular diagnosis, and with final healing rates between 40 and 60%. The main difference is that in group 3 all of its members are male, in group 4 the majority are women (91%), and in group 5 the male-female ratio is essentially one. Regarding the healing rates, 60% of patients are free of lesions in group 3, 40% in group 4, and 42% in group 5. Taking into account these clinical differences, the doctors decided to maintain the initial cluster groups.

In summary, the clinicians refined the automatically obtained clusters (as the proposed analysis pipeline is a visually interactive process), reducing the seven initial groups to five according to their clinical criteria. Figure 9(a) shows the updated clusters (silhouette scores 0.48) in the UMAP mapping, while Figure 9(b) shows the LDA representation that results from using the new patient groups (i.e., new class labels). The ALA axis vector is the most important one that points towards the right and is therefore key for separating group 7 from the rest. The patient groups associated with MAL are on the left side, where their separation is due to the axes that are roughly parallel to the Y-axis, specifically diagnosis, cryotherapy treatment, support and post treatments, and gender. For instance, group 8 is pushed down in the mapping primarily due to cryotherapy treatment.

We can also color the points in the plots according to the values of a particular feature. Figure 10 shows the same LDA mapping as in Figure 9(a), but the points are colored according to the ‘healing’ feature. From a clinical perspective the plot suggests that the ALA compound obtains better results than MAL regarding patient healing. This is because the ALA group (7) is barely affected by the rest of the analyzed features. Note that the only two relevant axis vectors that point towards the right are ‘healing’ and ‘ALA’. However, for the groups of patients treated through MAL some feature axes (such as ‘diagnosis’ or ‘post-treatment’) take on a remarkable importance, influencing the final visualization.

Additionally, the clinicians evaluated the homogeneity of the clusters on the two-dimensional projections by measuring the performance of K-NN classifiers (with a default value of $k = \sqrt{N}$) that consider the cluster groups as class labels. For this purpose, they split the data set into 70% for training and 30% for testing,



(a)

(b)

Figure 9: Projections involving the user refined clusters. The UMAP plot in (a) shows the resulting groups of patients under PDT treatment classified on 5 categories after the clinical analysis (group 3 in light green, 4 light blue, 5 navy, 7 orange, 8 green). (b) LDA mapping to characterize the patient groups by the length and orientation of the feature axes. K-NN classification is made with $k = 29$.

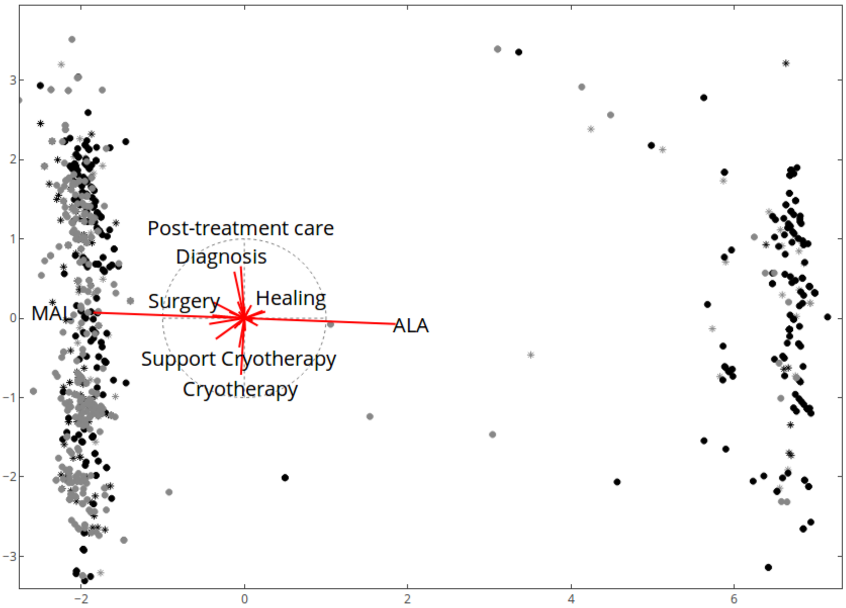


Figure 10: Analysis of healing through the same LDA mapping as in Figure 9(b). In this case the points are colored according to the 'healing' feature. Dark dots indicate cured patients, while lighter grey dots correspond to non-cured patients. Note that the 'healing' and 'ALA' axis vectors point towards similar directions (which usually indicates a positive correlation between the features). In this example, there is a higher proportion of cured patients on the right side of the plot, which are mainly the ones treated with ALA.

and analyzed the K-NN classification accuracy on the test set. The clinicians used this approach to compare the automatically generated clusters in Figure 8(a) and the refined clusters Figure 9(b). The graphics indicate the obtained classification accuracy in the panel to the right of the plots. In addition, the NMI (clustering quality) score is also shown above the mappings. As can be seen, the clustering created by the domain experts provides better results on LDA projections (74.30%, NMI = 0.61) than those obtained by DBSCAN (65.89%, NMI = 0.54). The classification accuracies on the UMAP mappings (see Figure 7 and 9(a)) are higher (97.68%, NMI = 0.95-0.96) as expected, since they usually capture the cluster structure of the data better.

5. Conclusions and Future Work

In this work, we have presented an interactive visual interface for data analysis of both labeled and non-labeled data based on DR projections. The tool allows users to import data sets and visualize them in a parallel and coordinated way with linear and nonlinear DR methods. In addition, they can carry out diverse data analysis tasks such as feature selection, clustering, and classification.

In comparison with other tools described in the state of the art, our method couples dimensionality reduction, clustering and classification through interaction and visualization. Similarly to other tools, we also allow users to cluster the data automatically in a low-dimensional space. However, with our interactive interface it is also possible to define clusters manually, by taking advantage of user perception and domain knowledge. In addition, our proposal complements unsupervised exploratory analysis with supervised (predictive) analysis, which can also be used to evaluate model performance.

We have presented a case study for analyzing a medical data set about patients with dermatologic conditions. The analysis of the data, which was carried out by clinicians of the HUF, revealed relevant information. In particular, the clinicians discovered that patients treated with ALA and MAL compounds obtained a global clinical healing rate close to 90%, and between 40-60%, respectively, according to the different analyzed groups. Therefore, the data analysis clearly suggests that the ALA drug obtains better results than the MAL drug in terms of clinical healing. Although new studies are required to validate these results, this would be an interesting finding, since there are no research works in the literature (with a large number of patients) comparing the ALA treatment (BF-200 ALA; Ameluz®) to the MAL (Metvix®) treatment.

We have published the tool and its source code (in Python, using Dash, Plotly, scikit-learn, pandas, and other packages) online (http://monkey.etsii.urjc.

es/drcc/visual_cluster_classification_tool) so that analysts can use it and data analysis experts can extend it and adopt it to their needs.

Finally, we have noticed that we are running several complex ML algorithms at once. If the data set is huge it can be time consuming to generate the visualizations or execute the algorithms. As future work, we are currently planning on implementing more scalable solutions, like Hierarchical-SNE (Pezzotti et al., 2016a) instead of t-SNE, or running the algorithms using big data technologies. Finally, we plan to introduce additional methods for training and validation, including bootstrap, cross validation, etc.

Acknowledgements

This work has been supported by the Spanish Ministry of Science, Innovation and Universities (grant RTI2018-098694-B-I00).

References

Alique-García, S., Alique, D., Company-Quiroga, J., Sanchez, A., Hernández, A., & Borbujo, J. (2019a). Treatment of Bowen's disease with photodynamic therapy. Observational study in 171 patients with 5-aminolaevulinic acid (BF-200 ALA) and methyl aminolaevulinate (MAL). *Photodiagnosis and Photodynamic Therapy*, *28*, 192 – 194.

Alique-García, S., Company-Quiroga, J., Sánchez, A., Hernández, A., & Borbujo, J. (2019b). Treatment of superficial basal cell carcinoma with photodynamic therapy. Observational study in 22 patients with 5-aminolaevulinic acid and methyl aminolaevulinate. *Photodiagnosis and Photodynamic Therapy*, *26*, 190–192.

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*, 175–185.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., & Giannotti, F. (2009). Interactive visual clustering of large collections of trajectories. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 3–10).

Ankerst, M., Elsen, C., Ester, M., & Kriegel, H.-P. (1999). Visual classification: An interactive approach to decision tree construction. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '99 (pp. 392–396). New York, NY, USA: ACM.

Babilas, P., Karrer, S., Sidoroff, A., Landthaler, M., & Szeimies, R. (2005). Photodynamic therapy in dermatology–an update. *Photodermatol. Photoimmunol. Photomed.*, *21*, 142–149.

Bellman, R. (1957). *Dynamic programming*. Courier Corporation.

Bradel, L., North, C., & House, L. (2014). Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 163–172). IEEE Computer Society.

Cavallo, M., & Demiralp, Ç. (2019). Clustrophile 2: Guided visual clustering analysis. *IEEE Trans. Vis. Comput. Graph.*, *25*, 267–276.

Choo, J., Lee, H., Kihm, J., & Park, H. (2010). iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (pp. 27–34).

Choo, J., Lee, H., Liu, Z., Stasko, J., & Park, H. (2013). An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Proc. SPIE 8654, Visualization and Data Analysis 2013* (p. 15). International Society for Optics and Photonics volume 865402.

Cox, T. F., & Cox, M. (2000). *Multidimensional Scaling, Second Edition.* (2nd ed.). Chapman and Hall/CRC.

Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.

Demiralp, Ç. (2016). Clustrophile: A tool for visual clustering analysis. In *KDD 2016 Workshop on Interactive Data Exploration and Analytics (IDEA16)* (p. 9). San Francisco, CA, USA: ACM.

Dheeru, D., & Karra Taniskidou, E. (2017). UCI machine learning repository.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. Wiley.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* KDD'96 (pp. 226–231). AAAI Press.

Fernández-Guarino, M., Harto, A., Pérez-García, B., Royuela, A., & Jaén, P. (2014). Six years of experience in photodynamic therapy for basal cell carcinoma: results and fluorescence diagnosis from 191 lesions. *J. Skin Cancer*, *2014*, 7.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* volume 1. Springer Series in Statistics.

Fritsch, C., & Ruzicka, T. (2006). Fluorescence diagnosis and photodynamic therapy in dermatology from experimental state to clinic standard methods. *J. Environ. Pathol. Toxicol. Oncol.*, *25*, 425–439.

Gilaberte, Y., Serra-Guillén, C., de las Heras, M., Ruiz-Rodríguez, R., Fernández-Lorente, M., Benvenuto-Andrade, C., González-Rodríguez, S., & Guillén-Barona, C. (2006). Photodynamic therapy in dermatology. *Actas Dermosifiliogr.*, *97(2)*, 83–102.

Gisbrecht, A., Hammer, B., Mokbel, B., & Sczyrba, A. (2013). Nonlinear dimensionality reduction for cluster identification in metagenomic samples. In *2013 17th International Conference on Information Visualisation* (pp. 174–179).

Hassanat, A., Abbadi, M., Altarawneh, G., & Alhasanat, A. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *International Journal of Computer Science and Information Security*, *12*, 33–39.

He, X. (2005). *Locality Preserving Projections*. Ph.D. thesis Faculty of the Division of the Physical Sciences, University of Chicago Chicago, IL, USA.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Commun. ACM*, *55*, 45–54.

Inselberg, A., & Dimsdale, B. (1987). Parallel coordinates for visualizing multidimensional geometry. In *Computer Graphics 1987* (pp. 25–44). Tokyo: Springer Japan.

Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., & Chang, R. (2009). iPCA: An interactive system for PCA-based visual analytics. *Computer Graphics Forum*, *28*, 767–774.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *374*.

Kandogan, E. (2000). Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics* (pp. 9–12).

Kwon, B. C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W. F., & Perer, A. (2017). Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 142–151.

Lai, C., Zhao, Y., & Yuan, X. (2018). Exploring high-dimensional data through locally enhanced projections. *Journal of Visual Languages & Computing*, *48*, 144–156.

Liu, S., Maljovec, D., Wang, B., Bremer, P., & Pascucci, V. (2017). Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, *23*, 1249–1268.

Liu, S., Wang, B., Bremer, P.-T., & Pascucci, V. (2014). Distortion-guided structure-driven interactive exploration of high-dimensional data. *Computer Graphics Forum*, *33*, 101–110.

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

van der Maaten, L., Postma, E., & van den Herik, H. (2009). *Dimensionality Reduction: A Comparative Review*. Technical Report TiCC-TR 2009-005 Tilburg University Technical Report.

Mao, Y., Dillon, J., & Lebanon, G. (2007). Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics*, *13*, 1208–1215.

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform

Manifold Approximation and Projection. *Journal of Open Source Software*, *3*, 861.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley-Interscience.

Molchanov, V., Chitiboi, T., & Linsen, L. (2015). Visual analysis of medical image segmentation feature space for interactive supervised classification. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine* VCBM '15 (pp. 11–19). Aire-la-Ville, Switzerland, Switzerland: Eurographics Association.

Molchanov, V., & Linsen, L. (2014). Interactive Design of Multidimensional Data Projection Layout. In N. Elmqvist, M. Hlawitschka, & J. Kennedy (Eds.), *EuroVis - Short Papers* (pp. 25–29). The Eurographics Association.

Nam, E. J., Han, Y., Mueller, K., Zelenyuk, A., & Imre, D. (2007). Clustersculptor: A visual analytics tool for high-dimensional data. In *2007 IEEE Symposium on Visual Analytics Science and Technology* (pp. 75–82).

Nam, J. E., & Mueller, K. (2013). Tripadvisor $^{\text{N-D}}$: A tourism-inspired high-dimensional space exploration framework with overview and detail. *IEEE Transactions on Visualization and Computer Graphics*, *19*, 291–305.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* NIPS'01 (pp. 849–856). Cambridge, MA, USA: MIT Press.

Park, J., Jang, Y., & Kim, Y. (2013). Ultrastructural changes in photorejuvenation induced by photodynamic therapy in a photoaged mouse model. *Eur. J. Dermatol.*, *23(4)*, 471–477.

Pezzotti, N., Hollt, T., Lelieveldt, B. P., Eisemann, E., & Vilanova, A. (2016a). Hierarchical stochastic neighbor embedding. *Computer Graphics Forum (Proc. of EuroVis)*, *35*, 21–30.

Pezzotti, N., Lelieveldt, B., van der Maaten, L., Hllt, T., Eisemann, E., & Vilanova, A. (2016b). Approximated and user steerable tSNE for progressive visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *23*, 1739–1752.

Reinhold, U. (2017). A review of BF-200 ALA for the photodynamic treatment of mild-to-moderate actinic keratosis. *Future Oncol.*, *13(27)*, 2413–2428.

Rieck, B., & Leitte, H. (2015). Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum*, *34*, 431–440.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326.

Rubio-Sánchez, M., Raya, L., Díaz, F., & Sanchez, A. (2016). A comparative study between radviz and star coordinates. *IEEE Transactions on Visualization and Computer Graphics*, *22*, 619–628.

Rubio-Sánchez, M., Sanchez, A., & Lehmann, D. J. (2017). Adaptable radial axes plots for improved multivariate data visualization. *Computer Graphics Forum*, *36*, 389–399.

Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., & Keim, D. A. (2017). Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, *23*, 241–250.

Sanchez, A., Soguero-Ruiz, C., Mora-Jiménez, I., Rivas-Flores, F., Lehmann, D., & Rubio-Sánchez, M. (2018). Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions. *Expert Systems with Applications*, *100*, 182–196.

Seo, J., & Shneiderman, B. (2002). Interactively exploring hierarchical clustering results [gene identification]. *Computer*, *35*, 80–86.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages* (pp. 336–343).

Strehl, A., & Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, *3*, 583–617.

Tarstedt, M., Gillstedt, M., Lark, A. W., & Paoli, J. (2016). Aminolevulinic acid and methyl aminolevulinate equally effective in topical photodynamic therapy for non-melanoma skin cancers. *J. Eur. Acad. Dermatol. Venereol.*, *30(3)*, 420–423.

Teoh, S. T., & Ma, K.-L. (2003). Starclass: Interactive visual classification using star coordinates. In *Proceedings of the 2003 SIAM International Conference on Data Mining* (pp. 178–185).

Turkay, C., Filzmoser, P., & Hauser, H. (2011). Brushing dimensions–a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, *17*, 2591 – 2599.

Wang, Y., Li, J., Nie, F., Theisel, H., Gong, M., & Lehmann, D. J. (2017). Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data. *Computer Graphics Forum*, *36*, 401–410.

Zhou, H., Yuan, X., Qu, H., Cui, W., & Chen, B. (2008). Visual clustering in parallel coordinates. *Comput. Graph. Forum*, *27*, 1047–1054.