Virtual Reality (2024) 28:41
https://doi.org/10.1007/s10055-024-00939-8

ORIGINAL ARTICLE

A virtual reality data visualization tool for dimensionality reduction methods

Juan C. Morales-Vega[1] · Laura Raya[1] · Manuel Rubio-Sánchez[2] · Alberto Sanchez[2]

Received: 13 December 2022 / Accepted: 9 January 2024 / Published online: 12 February 2024

Abstract
In this paper, we present a virtual reality interactive tool for generating and manipulating visualizations for high-dimensional data in a natural and intuitive stereoscopic way. Our tool offers support for a diverse range of dimensionality reduction (DR) algorithms, enabling the transformation of complex data into insightful 2D or 3D representations within an immersive VR environment. The tool also allows users to include annotations with a virtual pen using hand tracking, to assign class labels to the data observations, and to perform simultaneous visualization with other users within the 3D environment to facilitate collaboration.

Keywords Virtual reality · Information visualization · Natural interface

1 Introduction

The goal of data visualization is to graphically represent and communicate complex information, data, or results in order to provide valuable insights. This field is of paramount importance within data science, playing a pivotal role in extracting meaningful insights from data in various scenarios, encompassing scientific research as well as industrial applications. In medicine, clinicians may need to visualize many types of data, ranging from images or 3D models (Lawonn et al. 2018) to more challenging and complex data such as genomics information. In physics, it may be necessary to visualize complex data from simulations or several statistical analyses for exploring results (Ruder et al. 2008). In economics, analysts may need to inspect time-series data such as the evolution of sales or stocks (Schwabish 2014). In practice, analysts are encouraged to employ different techniques and to observe the data from multiple points of view, in order to discover properties, relations, or other insights.

The necessity to explore and visualize data is not new, but the study of visual perception has been becoming increasingly important over time. Many data visualization techniques try to benefit from the way the brain processes data and exploits pre-attentive uptake (Todorovic 2008) to provide effective and expressive visualizations (e.g., heatmaps or word clouds).

An important field of research considers the visualization of high n-dimensional data (where $n > 3$). When the data are numeric, a popular approach consists in applying dimensionality reduction (DR) techniques, which provide a transformed dataset of m features, where $m < n$. Some examples of these methods are PCA (Mishra et al. 2017), LDA (Tharwat et al. 2017), t-SNE (van der Maaten and Hinton 2008), UMAP (McInnes et al. 2018), star coordinates (Kandogan 2001), or parallel coordinates (Johansson and Forsell 2015). When m is sufficiently small, the data can be plotted and visualized in a 2D or 3D Cartesian coordinate system (if $m > 3$, it may be possible to represent the remaining features through graphical properties such as size, color, opacity, etc.).

Virtual reality (VR) has become increasingly popular in the past years, largely driven by the growing accessibility

✉ Alberto Sanchez
alberto.sanchez@urjc.es

Juan C. Morales-Vega
juan.vega@live.u-tad.com

Laura Raya
laura.raya@u-tad.com

Manuel Rubio-Sánchez
manuel.rubio@urjc.es

[1] U-Tad: University Center for Technology and Digital Art, C/ Playa de Liencres 2 bis, Las Rozas 28290, Madrid, Spain

[2] Department of Computer Science & Statistics, Universidad Rey Juan Carlos, C/ Tulipan s/n, Móstoles 28933, Madrid, Spain

of VR hardware and software. VR is a discipline that allows users to visualize a fully immersive 3D world generated by a computer using a head-mounted display (HMD). Most headsets also possess head and hand tracking functionality to transform user movements to movements in the virtual world by considering natural interactions. The recent technical advances in VR have extended its use to a large variety of applications. Among others, it is now possible to perform data analysis using stereoscopic visualization techniques in VR and in augmented reality (AR). In our work, we take advantage of this technology to represent numerical datasets in a real 3D space (not on a computer screen). This provides immersive visualizations that can be an alternative to monitor viewing. In addition, the multi-user capability of this technology also allows analysts to collaborate through the use of avatars and voice chat within the visualization itself.

This work proposes the design and development of a collaborative tool that leverages virtual reality (VR) technology for information visualization, specifically focusing on dimensionality reduction (DR) methods. The main objective is to provide analysts with a novel approach to observe and interact with their data in a three-dimensional (3D) virtual environment, enhanced by stereoscopic effects using a head-mounted display (HMD). By incorporating a natural interface based on handheld controllers and hand tracking, analysts can explore and engage with the data, gaining deeper insights. The tool presents data that have undergone transformation into a low-dimensional space using popular DR algorithms, such as principal component analysis (PCA), Linear discriminant analysis (LDA), star coordinates, and t-distributed stochastic neighbor embedding (t-SNE). Moreover, users have the capability to annotate the scene within a virtual room, enhancing collaboration and knowledge sharing among analysts.

The rest of the paper is organized as follows. Firstly, Sect. 2 reviews the state of the art and current applications in VR-based data visualization. Subsequently, Sect. 3 describes the developed application, while Sect. 4 presents its evaluation through a case study. Finally, Sect. 5 draws the main conclusions and discusses future work.

2 State of the art

It is acknowledged that interaction plays a fundamental role in information visualization. Not only can interaction make visualization processes less tedious (Betella et al. 2014), but many techniques require it in order to be effective for revealing insights. In general, graphical interfaces should be simple, while the mechanics should be at the perceptual level (van Dam et al. 2002). Thus, in visualization (not only VR-based), it is essential to design intuitive interfaces. One good example can be found in Kinetica (Rzeszotarski and Kittur 2014), which is a powerful data visualization tool that uses physical interactions (through a multi-touch screen) to process and manipulate the data. Thanks to these physical interactions, the tool is intuitive to use and easy to learn.

Instant Clue (Nolte et al. 2018) is another visualization tool that supports data visualization controlled through simple and intuitive gestures. It also offers statistical tools for interactive data manipulation and several machine learning techniques, such as support vector machines or decision trees. Again, its intuitive controls play a key role in the application.

Another highly interactive data visualization application is proposed in Mohedano-Munoz et al. (2021). It is a multipurpose tool that implements several dimensionality reduction techniques, such as PCA, LDA, or UMAP, to reduce the data to 2D. It also supports multiple plots at the same time, which greatly helps to obtain insights about the data.

In VR, there have also been various attempts to introduce realistic and intuitive data visualization tools. The work in Huang et al. (2001) integrated VR and geographical information systems for exploring spatial data. The VR visualization in de Haan et al. (2002) was developed to explore volumetric data and molecular dynamics. This early tool presented the drawback that the position for the virtual environment needed to be fixed with respect to a physical table, so it could not be used anywhere. In van Dam et al. (2002), the authors developed a set of early VR applications for archaeological data analysis, bioflow visualization in arteries, brain visualization, and Mars terrain exploration. However, given the early nature of the tools, they presented many issues, such as tracking errors, extremely low frame rates, and poor graphical and interaction designs.

Valdes et al. conducted several research works (Valdes and Barton 2006; Valdés and Barton 2007; Montes et al. 2008, 2010; Valdés et al. 2012) to couple visual data mining with VR spaces for data and symbolic knowledge representation. In short, they mapped the data, which could be heterogeneous, using different dimensionality reduction techniques into a homogeneous VR space that could be visualized in a VR cave. They provided promising results in different fields, such as microarray gene expression, cancer, computation performance, geophysical prospecting, earth sciences, or astronomy.

Nevertheless, it was not until 2014, with the emergence of commercial VR headsets (especially Oculus Rift), that VR applications started to become more powerful and beneficial in diverse areas. An economically more viable device with more advanced technology is capable of rendering more complete virtual worlds and natural interaction with positioning techniques. Donalek et al. developed an immersive and collaborative data visualization tool (Donalek et al. 2014) using the Unity Engine. In this tool, variables in a dataset could be mapped to several visualization channels,

such as position, color, or transparency. Moreover, multiple users could visualize the same data simultaneously. In the same year, Helbig et al. developed a tool specifically for visualizing atmospheric data in VR (Helbig et al. 2014).

VR also provides an intuitive way to perform visualizations in biology, thanks to the ability to represent 3D geometry accurately. In Ratamero et al. (2018), Ratamero et al. developed an application for visualizing proteins, acknowledging its usefulness for understanding their 3D structures. Nanome (Kingsley et al. 2019) is a collaborative tool that enables users to manipulate biomolecules in real time. Researchers found the tool very useful for drug discovery, since it allowed them to notice details, they would have missed in 2D.

Dimensionality reduction techniques can facilitate data processing, analysis, and visualization. There are many different methods that have been applied in numerous applications, such as word embeddings for language models (Devlin et al. 2019), molecular simulation (Ferguson et al. 2011), or cancer study using multi-omics data (Cantini et al. 2021), among many others.

In this work, we have developed a collaborative and immersive data visualization tool that can be used for exploring general datasets with numerical and categorical attributes. This tool can be fully operated within the virtual world using a HMD with stereoscopic vision, without the need to configure the visualizations externally. The user is also able to interact with the data using natural actions and movements, and take 3D annotations through hand tracking techniques. Our tool also includes several DR algorithms for plotting data in a 3D space. The multi-user capability of our tool allows us to perform collaborative analyses and meetings within the virtual 3D analysis space.

3 Proposal development

We have created a collaborative virtual environment that allows domain experts to analyze and visualize complex data in a 3D scenario that is modeled as an academic and work environment. With this tool, we are seeking to offer new ways to explore data that can complement classic information visualization techniques.

Our proposal, called VRDR (VR information visualization for Dimensionality Reduction) allows VR users to visualize a dataset and apply several dimensionality reduction methods over the data to get a clear visualization for analysis. The immersive environment we have developed allows multiple interactions with the data such as selecting or reclassifying records, taking 3D annotations or scaling the data by natural interaction through the HMD hand controllers and tracking.

For the development of VRDR, we have used the Unity Engine, version 2019.4.11, alongside the Oculus SDK, which enables easy controller binding and compilation for the Oculus Quest. The main advantage of Oculus Quest over other VR headsets is that it is a standalone platform that does not require a computer to work. This facilitates the analysis of data anywhere.

VRDR enables loading datasets and the selection of the parameters needed to perform the 3D scatter plot. Users can get an overview of the data by interacting with it and reducing occlusions thanks to stereoscopy and 3D navigation.

In accordance with visualization information seeking mantra (Shneiderman 2000), the user can scale the plot or filter out unnecessary classes to obtain an improved 2D or 3D visualization. Additionally, the user can select records for additional information via a floating hover tool, which allows further refinement of the representation by modifying other display options.

3.1 Dimensionality reduction methods

A fundamental part of VRDR is the dimensionality reduction algorithms, capable of reducing high-dimensional data automatically within the tool. Innitially, we have included the following four algorithms:

1. Principal Component Analysis (PCA) (Mishra et al. 2017). This algorithms finds the directions in the higher dimensional space where the variation of the data is maximum, orders them, and takes the first N, setting them as axes in the lower dimensional space and projecting all records into this new coordinate system. Although PCA works for any value of N that is less than the original dimensionality, since we are interested in visually displaying the data, the values available for N are 2 or 3 for 2D and 3D representations, respectively. Mathematically, this works as follows: The covariance matrix of the data can be easily computed as $C = X^T X$ and diagonalized, which gives us a new orthonormal vector basis ordered by eigenvalues. If we call that matrix V, projecting the data into that basis is as simple as computing the product $P = XV$. If instead of multiplying by the entire matrix V, we take only the top N columns to create a matrix V', we will get the projection over only the most important N axes $P' = XV'$. However, computing the covariance matrix is very expensive, so we can take a shortcut by computing the SVD decomposition of the data $X = UWV^T$, and plug it into the projection, which gives $P = UW$. In the same way as before, we can keep only the N upper rows of U to get $P' = U'W$.
2. Linear Discriminant Analysis (LDA) (Tharwat et al. 2017). This algorithm tries to maximally separate the data according to the unique values of a categorical

attribute. A way to achieve that is to first compute a within-class scatter matrix and a between-class scatter matrix of said attribute. We can then compute an overall matrix, extract the eigenvectors ordered by eigenvalue, keep the N with the highest eigenvalue, and multiply the resulting matrix with the data to reduce it to N dimensions. The mathematical process is similar to the one explained for the PCA.

3. Star Coordinates (Kandogan 2001). This algorithm interprets each one of the attributes to plot as a vector in the lower dimensional space. The value of the attribute represents the length of that vector for the record. In short, it interprets the D attributes of the data as the coordinates of a vector basis in N dimensional space. The direction of the vectors given to each attribute can be created arbitrarily. Since there is no particular criterion for their directions or lengths, the algorithm presents a natural way to interact with the data by manually manipulating the axes, which changes the position of all records.
4. t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008). t-SNE is a nonlinear algorithm that tries to maintain similar distances between the records in the lower dimensional space as they have in the higher dimensional space. To do that, the distances in the higher dimensional space are modeled using gaussian distributions over the records. In the lower dimensional space, the records are initially placed in random positions, and iteratively, they move until reaching a similar distribution. One classic method to reduce the error in each iteration is by using a gradient descent algorithm. For better separation of the records, instead of a normal distribution, in the lower dimensional space, a t-distribution is used.

3.2 Virtual environment

We have designed a realistic environment. This favors the use of the tool as a place for different users to work and meet, and for collaborative data analysis and an academic meeting point. Specifically, we modeled a large office room, with enough space for several users to interact and a large visualization where the user has multiple tools to analyze the loaded data (see Fig. 1a). The design of this scenario is based on the use of flat colors without much contrast or brightness, to avoid interfering with the visualization of the data.

We have designed a user interface that allows the analyst to work and interact with datasets. Located at the front end of the virtual environment, we allow the user to load the different datasets and select the different options for visualization. In the event of an error in loading the data, the tool gives visual feedback to the user. All of these options are explained in Sect. 3.2.1.

Located at the back of the virtual environment, the tool displays a table with application instructions. In addition, the coordinate system is displayed for visualization. The walls surrounding the coordinate system have been colored white (see Fig. 1b) to minimize the impact that the background color may have on the set of marks and channels in the visualization (Bertin 1983; Reinhard 2008). In the virtual environment, interactive objects are made available to the user, implemented as pencils to make annotations within the virtual world and an eraser to allow the user to erase them. These tools are explained in more detail in Sect. 3.2.3

Data are loaded using a hovering file explorer in the UI. The user can navigate through the computer files and select a CSV dataset. The data in the CSV need to be separated by semicolons.

The dataset columns are also automatically separated in two different groups: numerical and categorical variables. This automatic detection is effective for the analyst, since the

(a) Front part of the room (b) Plot area

Fig. 1 Modeling of the collaborative virtual environment. **a** The floating menus that allow the loading of datasets. **b** The 3D and 2D visualization area through a white or black board

user does not have to do it manually, and the system detects that the two types of data should be treated very differently when visualized.

Once the data are loaded, the configuration menu appears and the user can select the different options for the plot. The interaction area of the menus is adapted to the user's distance from the menu. Thus, the user can interact from different distances without requiring a high precision of the controller movement, facilitating the user experience.

3.2.1 Configuration options

VRDR has two main setup menus. To enhance the user experience, all menus can be moved and positioned wherever the user wants via the controllers. They are not in fixed positions as in a 2D interface. In addition, the menus are oriented according to the camera's point of view, allowing information to always be visible to the user. The first one is a menu located at the front right side of the virtual environment (shown in black and blue in Fig. 2). This menu contains information about the columns of the loaded dataset and their type of variables. The second is a floating menu that the user can open and close via the headset controllers. This menu contains the visualization tools that can be used to manipulate the plot in real time. This menu is always displayed at a distance of 3 ms from the user to be fully visible.

On the left side of this floating menu, the user can select the type of plot as well as the columns (variables of the dataset) needed to create it. The available columns in the menu depend on the type of plot. There are four attributes to choose that are common to most of them:

1. Color: select the column or variable to give color to the data marks. If the variable is categorical, then each mark is mapped to a different color tone depending on the unique values in the dataset, up to a maximum of twelve. At first, we followed Erik Reinhard's advice to use at most seven different colors for categorical variables (Reinhard 2008), but as in many cases, we found examples with more than seven classes, we finally extended the number of colors to twelve. The colors are constructed from the RGB primaries, and the rest are created in the same way as the secondary and tertiary colors are calculated. An interactive color legend is displayed and can be placed wherever the user prefers. If the variable is numeric, the color of the marks will behave like a heat map (values are mapped to color brightness, following Bertin's recommendation Bertin 1983). Finally, if the default option "None" is chosen instead of selecting a variable, all marks will be gray.
2. Size: select the column or variable that gives sizes to the marks. This attribute is intended to work with numerical variables. It can be left as "None" to give all marks the same size.
3. Form: select the column or variable that that gives form to the mark. This column needs to be categorical according to Bertin (1983). Each record will be mapped to a different mesh depending on the unique values of the selected column. The mesh mapping is also shown in the legend. It can be left as "None" to display all marks as spheres.
4. Type: The user can choose between 3D visualization or a more classic 2D layout, both stereoscopic. In the first case, the user can walk inside the data, obtaining a 360-degree view of it. In the second case, marks are drawn on a 2D blackboard (see Fig. 3) located on the wall next to the 3D view. Both visualizations can be displayed simultaneously, and the interaction with the data is the same for both types. Note that the blackboard used for 2D visualization can be hidden if the user prefers.

In the blue dropdown list of the floating menu, users can select the type of plot. We have implemented five different types of graphs, four of them corresponding to the dimensionality reduction algorithms shown in Sect. 3.1. The types of plot and their respective attributes are described below:

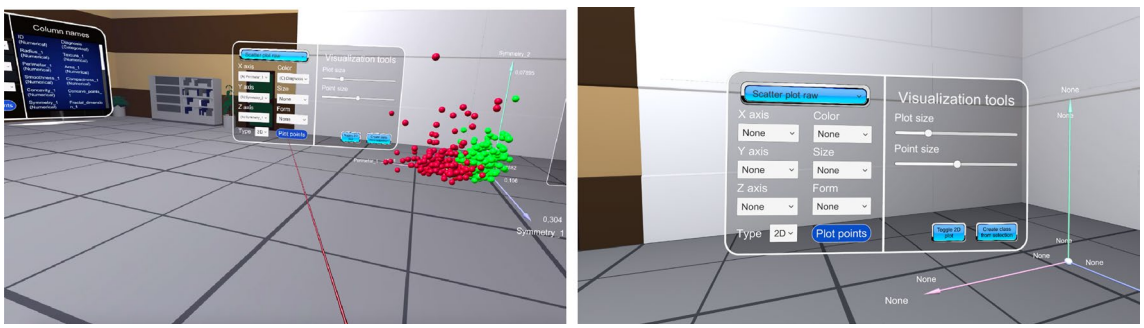


Fig. 2 Menus for configuring the displays and the application of dimension reduction methods

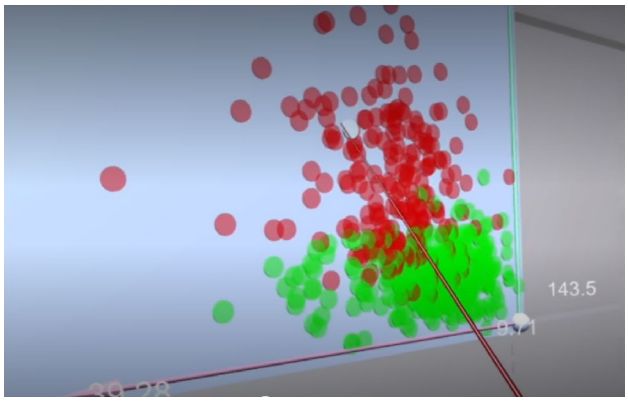
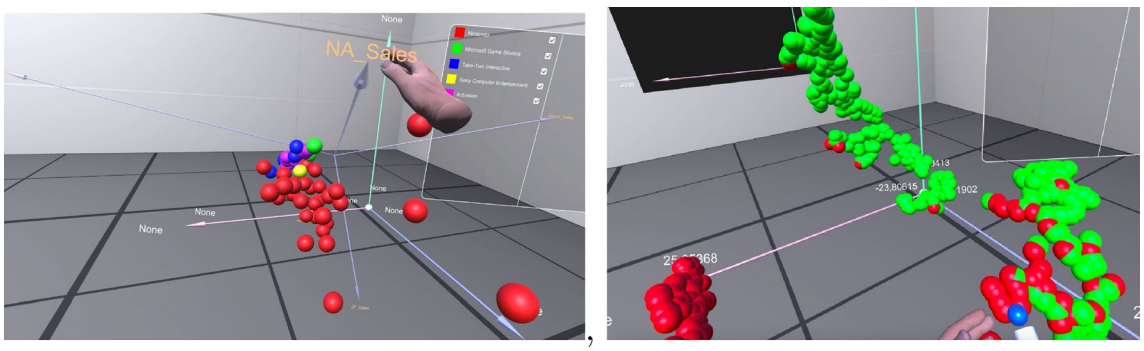


Fig. 3 The tool allows the user to visualize the data in 2D if a more traditional visualization is preferred. Interaction with the data is also done through the HMD controller and a laser pointer

1. Three-dimensional scatter plot: This option has three attributes named "X-axis," "Y-axis," and "Z-axis." The data are spatially distributed according to the raw values for the chosen columns for each attribute. Both numerical and categorical columns can be selected. If one or more attributes are left as "None," the data will be displayed as a scatter plot or 1D plot.
2. PCA: This options applies the principal components analysis algorithm over a set of columns to reduce the dimensionality to three. It possesses the attributes "First column" and "Last column" that define the column interval taken for the algorithm (Mishra et al. 2017).
3. LDA: This options applies the linear discriminant analysis algorithm over a set of columns to reduce the dimensionality to three. In the same way as the PCA, it possesses the attributes "First column" and "Last column" that define the column interval. It also has the "Class" attribute that define the column used for the algorithm to separate the data. The selected column for the class needs to be categorical (Tharwat et al. 2017).
4. Star Coordinates: This option applies a star coordinates visualization (Rubio-Sánchez et al. 2016). This also possesses the attributes "First column" and "Last column." It has also a initialization attribute that defines the initial position of the axes. High-dimensional data are represented in 3D space by constructing one 3D axis per each selected column and making the weighted sum over all of them. These axes are also displayed along with a tag to identify which one correspond to which column. They can be moved and scaled using the virtual hands of the user and the data marks react in real time adopting their corresponding position according to how the axes are being moved. An example is shown in Fig. 4a.
5. t-SNE: This option computes the nonlinear t-SNE dimensionality reduction algorithm (see Fig. 4b). It also presents the attributes "First column" and "Last column." It also has a "Perplexity" attribute that controls the number of neighbors that will be used for the algorithm. Since this algorithm is computationally expensive, it runs in parallel to not block the user movement (van der Maaten and Hinton 2008).

3.2.2 Data exploration

Note that not all datasets will be easily separable by classes (e.g., see Fig. 5). In these cases, the tool offers different options for processing and interacting with the data, enabling interaction and manipulation of the visualized data in real time, to facilitate the user's analysis. For example, it is possible to visualize the data in 3D in a 360-degree view, using VR and stereoscopy enabling different perspectives of the data. In this sense, it is possible to start from an overview of the dataset (see Fig. 6a) and move to the foreground by simply walking to the center of the data, as shown in Fig. 6c. The user can brush, select the desired data without losing



(a) Star Coordinates (b) t-SNE

Fig. 4 Examples of different types of plots. The user's hand can be viewed by selecting the data of interest

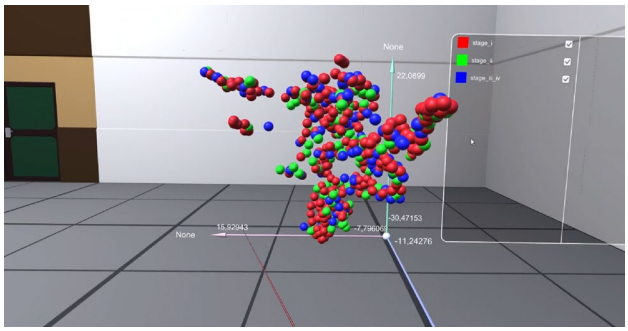


Fig. 5 Example of a dataset that does not provide a clear separation

the general context, or select a record and view its information in detail.

Using the floating menu, the user can change the size of the coordinate system, scaling everything in it, including the data marks. This size change acts as a multiplicative value on the original change, so if a column has been selected for the "Size" attribute, the relative size between the marks will still be preserved.

Also, as mentioned above, our tool has an interactive legend automatically generated following the color and shape attributes that were chosen in the initial configuration of the plot. The legend is always oriented toward the user, even when the user scrolls through the plots. An example of the legend is shown in Fig. 7. Classes can be selected or deselected interactively in the legend. Records belonging to a selected class will have an opaque color. Records belonging to a deselected class will have a transparent color, allowing the rest of the data to be seen while providing context (see Fig. 6b). Stereoscopic vision reduces the occlusion of the projected 2D data.

A record is considered selected if its entire class (selected through color and shape) is selected, although it is also possible to highlight individual records manually directly through the HMD controller (see Fig. 8). Using the natural interaction of VR, the user can point the HMD controller's laser at a record by keeping the index trigger

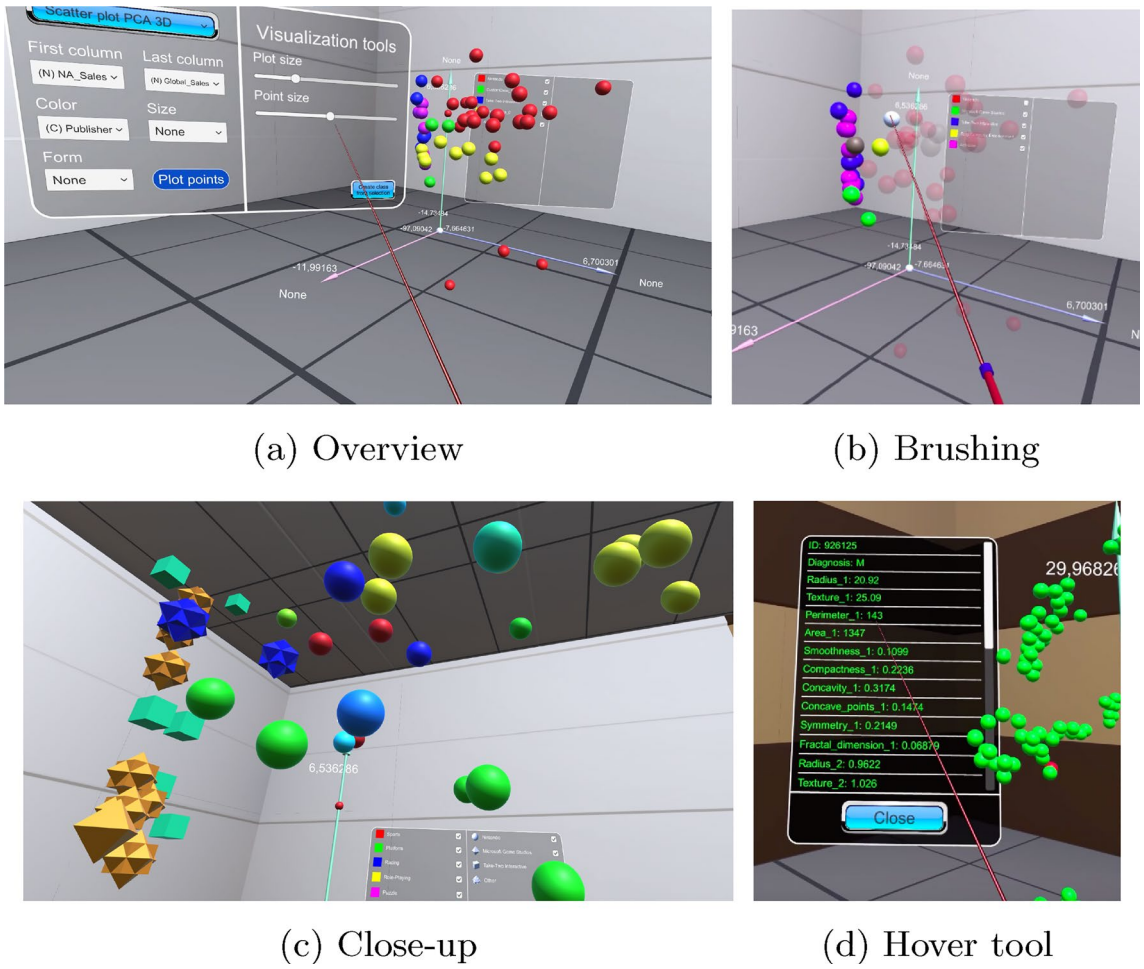


Fig. 6 Different options and steps to explore the data

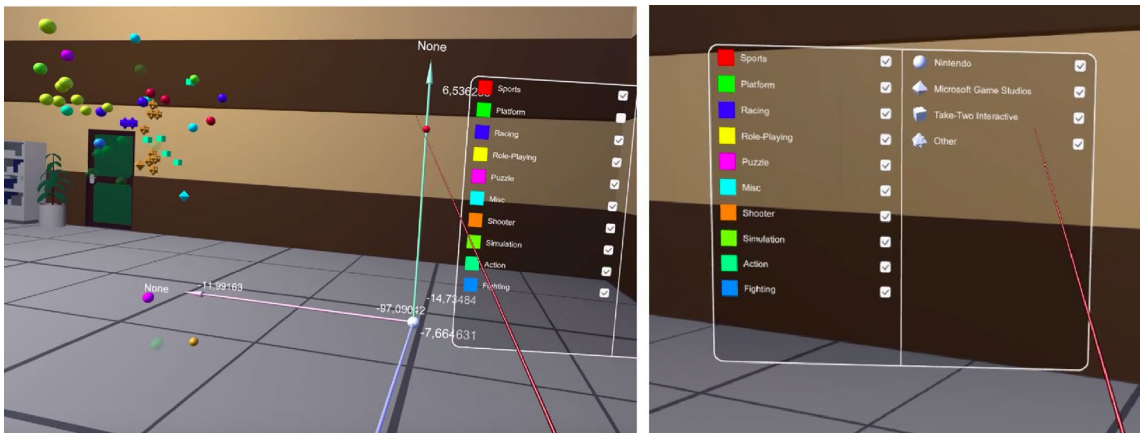


Fig. 7 The user can filter data by color and shape, using the floating legend menu

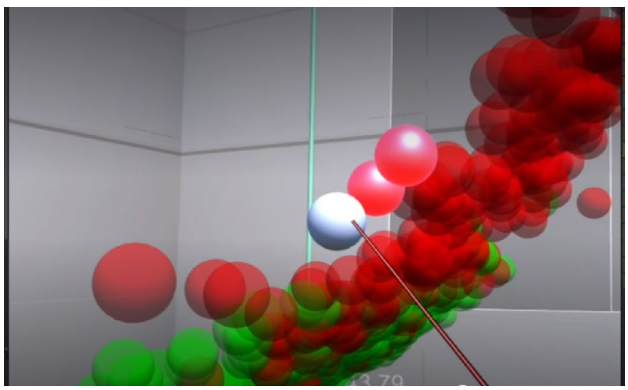


Fig. 8 The user can select the desired data or group of data manually via the HMD controller pointer. A specular material is displayed to identify the selected data

pressed for half a second to highlight it. This allows the user to not switch their focus to a menu or a button, as in a traditional 2D interface, without having to take their eyes off the data. The material of the marks will change to a much brighter one. Non-selected records can also be highlighted, resulting in a bright but transparent mark. A record can be de-emphasized by doing the same process.

Highlighting records is also used for another function: the creation of custom classes. Once users have made a selection, it is possible to transfer the whole selection to a new complete class. To do this, just click the "Create class from selection" button in the floating menu. This new class is created as a color class. These classes are also fully functional, appear in the legend, and are taken into account for algorithms that use class information, such as LDA.

Finally, we can also scan all variables contained in a single record. To do this, users can quickly click the index trigger while the laser pointer is pointing to a mark. This opens a Hover tool that displays the values of each of the columns in the dataset for that particular record (see Fig. 6d).

3.2.3 Virtual annotations

Taking manual notes within of the data can be useful to highlight certain values for later visualization. Using VR and hand tracking methods, we allow the user to create annotations on the data itself within VRDR (see Fig. 9). For this purpose, we have implemented virtual pens in three different colors (red, green, and blue). The user will be able to paint on the 3D or 2D graphic itself simply with his hand. To paint, the user must first pick up the desired virtual pen on the stage with his/her virtual hands and then press and hold the index finger trigger to write. You can also erase strokes with the virtual eraser by picking up the virtual eraser object on the stage and placing it over the stroke to be erased by pressing the index trigger.

3.2.4 Collaborative visualization

VRDR enables work meetings within the virtual environment and immersive, collaborative visualization of data. This facilitates collaborative analysis between different users, for example, one being the data analyst and another the domain expert. Users can be in different physical locations. When logging into the tool, they will have three different options: start the tool in offline mode, create a virtual room, or join a room.

The offline mode option allows the user to start the app immediately with all the functionality explained before. The other two options correspond to the collaborative mode, which has been implemented using the Photon Engine, integrated in Unity. A green or red dot in the collaboration section will inform the user whether the connection to the server was successful or failed. Users will be able to upload

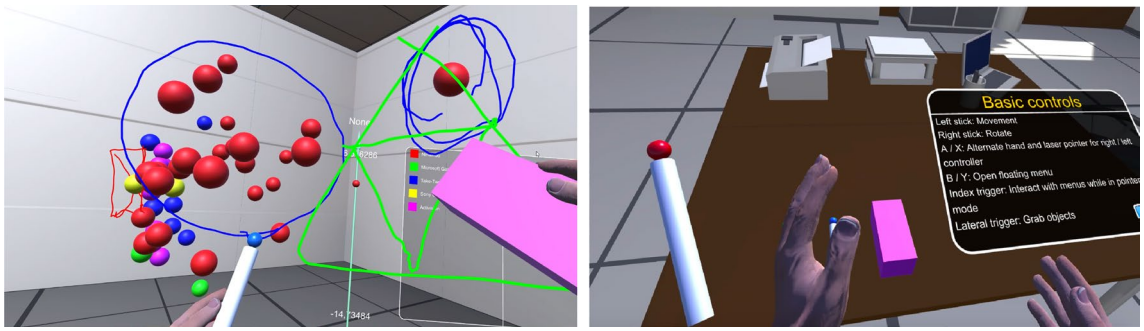


Fig. 9 Annotating in VRDR

datasets, configure the plot, and interact with menus and data.

Position and scale are synchronized natively with Photon Views, while attributes (such as color) are sent via simple RPC messages. Virtual annotations are similarly synchronized, transmitting to all viewers the positions of all points that are part of the drawn line. This facilitates analysis and learning, as data are analyzed collaboratively among users. Users can talk during their collaborative visualization, sharing progress and interaction strategies.

For the avatars, we decided to go with a very simple approach and just keep the virtual hands for every user connected (see Fig. 10).

3.2.5 Controls

The application is controlled using a Oculus Quest or Oculus Rift headset and the Oculus Touch. The user can move around the virtual environment through the room-scale method of locomotion. This method enhances the immersive user experience as it causes little motion sickness and allows a natural interaction with the environment. Additionally, in order to extend the virtual workspace in a reduced physical workspace, our interface allows movement via joystick. The left joystick is used for basic movement. The selected type of movement was smooth locomotion where, similar to a videogame, the user moves in the direction, the joystick is pushed. We deemed unnecessary the use of a teleportation system, since in this case, the movements are simple enough to not cause motion sickness. The right joystick is used for rotations. In the same way as the movement, it was implemented as a smooth turn.

The B and Y buttons open and close the floating menu. This menu can be interacted in the same way as the main menu in the front part of the room and provides the same functionality (choose plot options). It also has several visualization tools that allow the user to, e.g., modify the size of the data marks or create new classes, as seen above.

The A and X buttons swap the hand model between a normal human hand and a laser pointer. Each of those has a different purpose. The human hand is used to interact with physical objects, such as pencils to take notes, and the laser

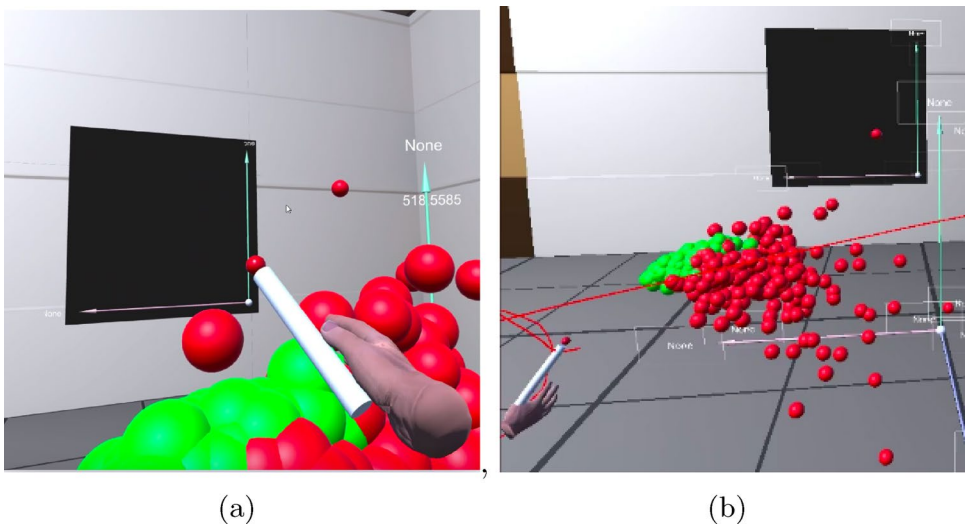


Fig. 10 Two users visualizing the data from different points of view. **a** A user sees his/her virtual hands and paints on the visualization. **b** Another user sees how the previous one paints with the virtual red pen

pointer is used to interact and select options from the different menus. Note that the bounding boxes for the interaction of the different buttons and check boxes of the menus are wider than their display. Thus, facilitating the interaction of the raycast of the controllers by extending their activation zone.

The lateral trigger, also known as "held" trigger, is used to grab objects, as it is usually the case in many different VR applications. Pressing it with the hand model allows the user to grab objects. If pressed with the laser pointer allows the user to grab and move the floating menu from a distance.

The index trigger is used to select options in the menus when the laser pointer is active. When the hand model is active and an object is grabbed, the index trigger is used to activate the object (for example, paint in the air when the pencil is grabbed).

The controls are summed up in Fig. 11.

4 Case study

The evaluation of our virtual reality data visualization tool was conducted in strict adherence to ethical principles and guidelines. We are pleased to report that our study received the formal approval of the Ethics Committee of U-tad. This approval underscores our commitment to ensuring that all aspects of our research, including participant involvement and data handling, comply with the highest ethical standards. All participants were provided with clear and comprehensive informed consent forms detailing the nature of their involvement, the purpose of the study, the procedures, and the handling of their data.

We have carried out two evaluations with our tool. To prove how easy it is to visualize data with VRDR, firstly, five data analysts used our tool as a case study to analyze the well-known WDBC dataset (Breast Cancer Wisconsin Data Set) (Mangasarian et al. 2022). This dataset contains several (+500) cancer cases, labeled as benign or malign, as well as many other numerical parameters characterizing each case, such as radius, texture, or perimeter among others. The VRDR users were between 20 and 30 years old, four men and one woman. Three had previously worked with virtual reality HMD and two had not.

Performed individually at different times, users had to tell what they were doing, and the tester could see the actions inside the virtual world as a second avatar. All tests were recorded for later analysis of the sequence of the steps performed.

First, the users loaded the dataset. Once loaded, the users inspected the column types on the right-hand menu. Some of them started to make a simple scatter plot. To do this, one user opened the floating menu, repositioning it until it fitted his/her needs. The users selected the desired columns and plotted the result, as shown in Fig. 12.

Most of the users analyzed the overview of the data by walking through the 3D graph. Thanks to the different views of the stereoscopic 3D visualization, the majority of them commented that could see a separation between the classes. That is, based solely on the observation of geometric data, they were able to begin to differentiate between malignant and benign cancer samples. Some users began to take notes and visually separate the data using their virtual hands.

Next, a user tried to check whether the separation of the samples could be seen more accurately. Up to this point, he had only used three columns. He used the VRDR menu to select all data and applied a dimensionality reduction technique. He selected LDA on the advice of the expert and chose the classification class to better separate the dataset. He was able to see a much sharper separation between the classes (see Fig. 13). Other users selected the PCA technique, with less interpretation success.

Users commented that the visualization achieved reinforced the theory of building a robust model for detecting malignancy in cancer from these data, as the class was separable. Some users tested the nonlinear t-SNE technique selecting different parameters. Afterwards, the separation

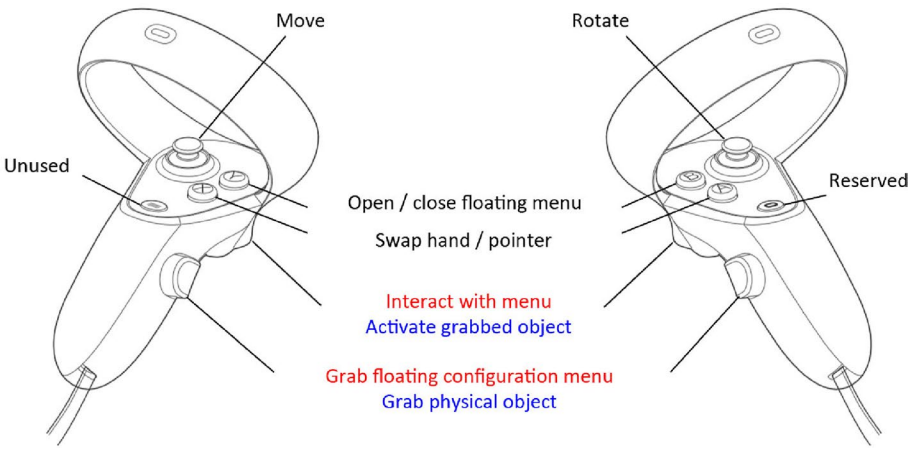


Fig. 11 Controller bindings

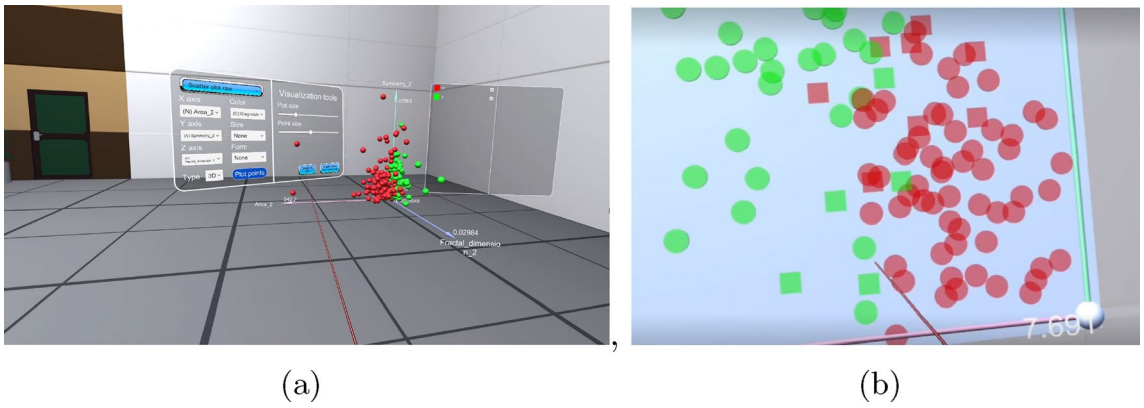


Fig. 12 Users performed different actions to understand the dataset. **a** Four users chose selection by color in 3D. **b** One user preferred to use the 2D option to visualize the data

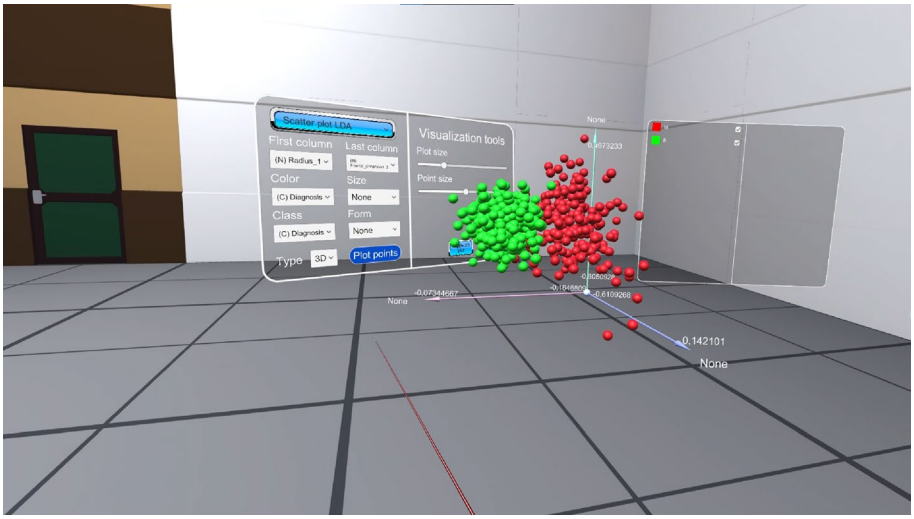


Fig. 13 The users had the freedom to select the best classification technique they considered. In this example, LDA method is shown

was even clearer. One user ran through the data to view the resulting plot from different perspectives (see Fig. 14).

However, some records were still mixed up in the class separation visualization. Another user drew a circle around one of the values, which he believed to be outliers, with the pencil tool and indicated that more information would be needed to correctly classify these samples (see Fig. 15).

After exploring the different options, some users tried to reclassify the identified outliers as a third class, change the perplexity value for t-SNE, or even move the axes in star coordinates to understand its operation.

After the task of handling VRDR with the dataset, users indicated that they found it easy to operate the menus and interact within the virtual world. Four of the five users indicated that they found it very useful to get into the data and interact with it with a natural interaction through their hands. One of the users indicated that they would prefer not to have to scroll through the virtual world to select the different options, being more usable to have everything without scrolling. Another user indicated that he found the 2D whiteboard option very useful to have two points of view to visualize the data. All users stated that the collaborative mode of being able to be inside the data with another person was very useful, especially for academic purposes.

Subsequently, we conducted the system usability scale (SUS) questionnaire (Brooke 1996) to a diverse group of 16 participants, with an age range between 20 and 41 years, half of them with knowledge of VR and half of them without. Users who had no previous experience with virtual reality HMD were given an additional 5 min to adapt to the immersive device.

We provided several datasets and allowed subjects to freely explore and interact with the tool for 10 min to provide valuable information for the VRDR usability evaluation. Subjects were then asked to complete the SUS test questions, and, later, a conversation was held with each of them.

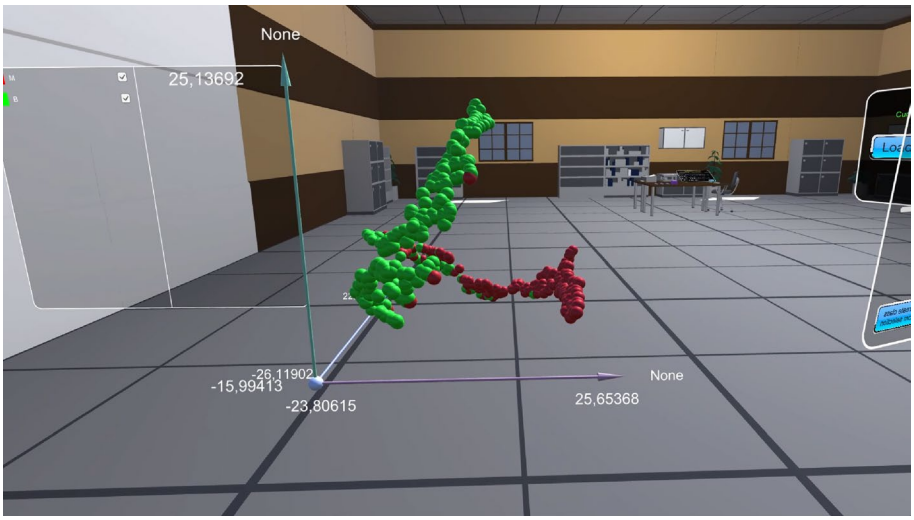


Fig. 14 Users can move through the virtual world to see different perspectives of the data and interact with it

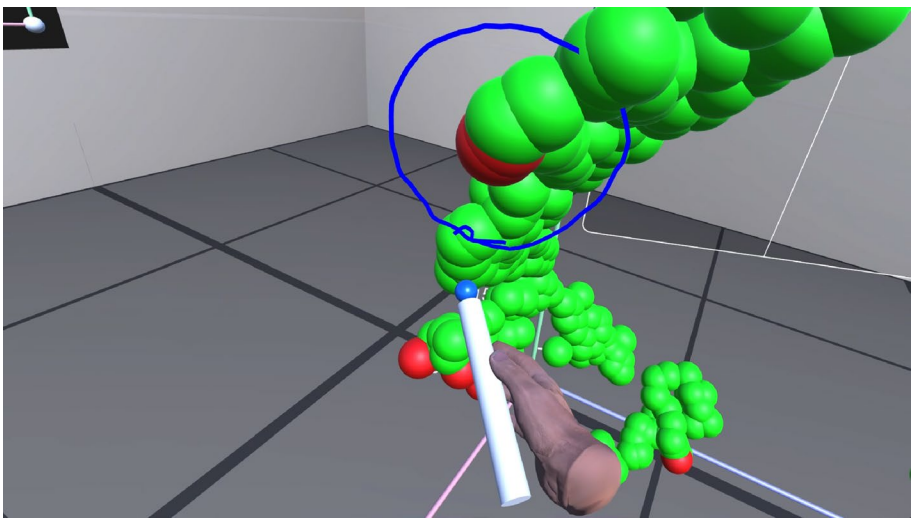


Fig. 15 Example of one of the users selecting an outlier

The SUS scores ranged from 87.65 to 100, indicating a high level of user satisfaction with the application's usability. These results complement the qualitative feedback obtained from the subjects, confirming the effectiveness and user-friendliness of the proposed immersive VR data visualization tool.

Specifically, several subjects commented that the application was remarkably intuitive and easy to use, even for those who had not previously experienced working with virtual reality applications. They found the three-dimensional visualization of data particularly useful, as it allowed them to rotate the visual representation, customize axes directly with hand gestures, make selections, and immerse themselves within the graphs. Additionally, the collaborative note-taking feature was regarded as an advantage compared to other dimension reduction and data visualization software.

Interestingly, none of the subjects used the option to switch from 3D to 2D visualization after becoming accustomed to the 3D graphs. However, several subjects mentioned that the true potential of utilizing virtual reality in data visualization would be realized by incorporating sensory channels beyond just visual and stereoscopic cues. They proposed integrating other sensory dimensions, such as sound or touch, to complement visual cues and provide additional stimuli for domain experts to interpret the dataset. This could involve using vibrations, forces, or sounds to convey certain values, allowing experts to engage multiple senses in the data exploration process.

These insights highlight the positive reception of the application's usability, with subjects emphasizing its intuitiveness, immersive 3D visualization capabilities, and collaborative note-taking functionality. Furthermore, the suggestions for incorporating additional sensory dimensions demonstrate the potential for further enhancing the application's effectiveness in data analysis and interpretation.

5 Conclusions and future work

In this paper, we have presented an immersive and collaborative VR data visualization tool, VRDR. The tool allows dimensionality reduction through different techniques as LDA, t-SNE, and so on. It is equipped with several interactions such as data marks scaling, coloring, class highlighting, or creation, among others, to facilitate the analysis of different datasets. The stereoscopic visualization in an immersive environment allows the user to perform the analysis collaboratively with other users, facilitating joint investigation. The ability to walk through the data and get 360 degrees of perspective reduces occlusions and data reduction. The tool enables annotations to be made on the data itself, which can be viewed by all users.

Following a case study, it has been shown that the tool offers data analysis possibilities within a virtual reality world. Users have been able to obtain information from the data without showing problems in the use of the tool. In a second evaluation with the SUS usability test, the users have indicated a high level of satisfaction with the tool's usability. It is important to indicate that our focus has been on improving usability rather than the sense of presence.

Note that the tool can handle datasets with a substantial number of dimensions, but it may not be optimized for large-scale datasets with a huge number of records. The processing and rendering capabilities required to visualize and interact with large-scale datasets can lead to potential performance issues and decreased user experience. Furthermore, it is worth noting that VRDR is primarily designed for information visualization (InfoVis) and is not specifically tailored for volumetric data with multiple values per voxel, such as those derived from 3D scans or simulations. These types of datasets, commonly encountered in scientific visualization (SciVis), require specialized techniques and algorithms to effectively represent and analyze the data.

As future work, we are considering the inclusion of multiple simultaneous graphs within the virtual environment, each representing different datasets. This enhancement would enable analysts to compare and contrast various datasets simultaneously, fostering a more comprehensive understanding of the underlying patterns and relationships. Additionally, we plan to incorporate parallel coordinates, a powerful visualization technique, into the tool to further enhance its versatility and enable analysts to visualize high-dimensional data effectively. Furthermore, while the initial evaluation involved pilot groups in a case study setting, we recognize the value of a larger subject pool and a concrete real-world scenario. This broader evaluation would provide deeper insights into the tool's usability, effectiveness, and practical application. It would enable us to gather more robust feedback and identify any potential areas for refinement and optimization. Finally, we intend to explore additional dimensionality reduction algorithms beyond the ones already implemented to enhance the tool's flexibility and applicability to various data types and domains. This expansion would offer analysts a broader range of options to transform their data into low-dimensional spaces, catering to diverse analytical needs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10055-024-00939-8.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The datasets analyzed during the current study are available in the UCI Machine Learning repository, https://archive.ics.uci.edu/ml, mainly Breast Cancer Wisconsin (Diagnostic) Data Set https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Bertin J (1983) Semiology of graphics. University of Wisconsin Press, Madison

Betella A, Bueno E, Kongsantad W, Zucca R, Arsiwalla X, Omedas P, Verschure P (2014) Understanding large network datasets through embodied interaction in virtual reality. In Proceedings of the 2014 virtual reality international conference, vol 2014, pp 23:1–4

Brooke J (1996) SUS: a 'quick' and 'dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability evaluation in industry, chapter 21. Taylor and Francis, pp 189–194

Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A (2021) Benchmarking joint multiomics dimensionality reduction approaches for cancer study. Nat Commun 12(1):124

de Haan G, Koutek M, Post FH (2002) Towards intuitive exploration tools for data visualization in VR. In Proceedings of the ACM symposium on virtual reality software and technology, VRST '02, pp 105–112, New York, Association for Computing Machinery

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language

technologies, NAACL-HLT 2019, pp 4171–4186. Association for Computational Linguistics
Donalek C, Djorgovski SG, Cioc A, Wang A, Zhang J, Lawler E, Yeh S, Mahabal A, Graham M, Drake A, Davidoff S, Norris JS, Longo G (2014) Immersive and collaborative data visualization using virtual reality platforms. In Proceedings of the 2014 IEEE international conference on big data (Big Data), pp 609–614
Ferguson AL, Panagiotopoulos AZ, Kevrekidis IG, Debenedetti PG (2011) Nonlinear dimensionality reduction in molecular simulation: the diffusion map approach. Chem Phys Lett 509(1):1–11
Helbig C, Bauer H-S, Rink K, Wulfmeyer V, Frank M, Kolditz O (2014) Concept and workflow for 3D visualization of atmospheric data in a virtual reality environment for analytical approaches. Environ Earth Sci 72(10):3767–3780
Huang B, Jiang B, Lin H (2001) An integration of GIS, virtual reality and the internet for visualization, analysis and exploration of spatial data. Int J Geogr Inf Sci 15:439–456, 07
Johansson J, Forsell C (2015) Evaluation of parallel coordinates: overview, categorization and guidelines for future research. IEEE Trans Visual Comput Graph 22:1–1, 11
Kandogan E (2001) Star coordinates: a multi-dimensional visualization technique with uniform treatment of dimensions. In Proceedings of the IEEE information visualization symposium
Kingsley LJ, Brunet V, Lelais G, McCloskey S, Milliken K, Leija E, Fuhs SR, Wang K, Zhou E, Spraggon G (2019) Development of a virtual reality platform for effective communication of structural data in drug discovery. J Mol Graph Model 89:234–241
Lawonn K, Smit NN, Bühler K, Preim B (2018) A survey on multimodal medical data visualization. Comput Graph Forum 37(1):413–438
Mangasarian OL, Wolberg WH, Street WN (2022) Breat cancer wisconsin (diagnostic) data set. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic), Visited 08 Feb 2022
McInnes L, Healy J, Saul N, GroBberger L (2018) UMAP: uniform manifold approximation and projection. J Open Source Softw 3(29):861
Mishra S, Sarkar U, Taraphder S, Datta S, Swain D, Saikhom R, Panda S, Laishram M (2017) Principal component analysis. Int J Livestock Res
Mohedano-Munoz MA, Alique-García S, Rubio-Sánchez M, Raya L, Sanchez A (2021) Interactive visual clustering and classification based on dimensionality reduction mappings: a case study for analyzing patients with dermatologic conditions. Expert Syst Appl 171:114605
Montes J, Sánchez A, Valdés JJ, Pérez MS, Herrero P (2008) The grid as a single entity: towards a behavior model of the whole grid. Lecture notes in computer science, on the move to meaningful internet systems: OTM 2008. Springer, Berlin. pp 1611–3349, 5331:886–897
Montes J, Sánchez A, Valdés JJ, Pérez MS, Herrero P (2010) Finding order in chaos: a behavior model of the whole grid. Concurrency and computation: practice and experience. Wiley vol 22, pp 1386–1415
Nanome. Nanome: virtual reality for drug design and molecular visualization. https://nanome.ai/, Visited 2021-05-05
Nolte H, Macvicar TD, Tellkamp F, Krüger M (2018) Instant clue: a software suite for interactive data visualization and analysis. Sci Rep 8(1)
Ratamero EM, Bellini D, Dowson CG, Römer RA (2018) Touching proteins with virtual bare hands. J Comput Aided Mol Des 32(6):703–709
Reinhard E (2008) Color imaging: fundamentals and applications. CRC Press, Boca Raton
Rubio-Sánchez M, Raya L, Díaz F, Sanchez A (2016) A comparative study between radviz and star coordinates. IEEE Trans Visual Comput Graphics 22(1):619–628
Ruder H, Weiskopf D, Nollert H-P, Müller T (2008) How computers can help us in creating an intuitive access to relativity. New J Phys 10(12):125014
Rzeszotarski JM, Kittur A (2014) Kinetica: naturalistic multi-touch data visualization. In Proceedings of the SIGCHI conference on human factors in computing systems, CHI '14, pp 897–906, New York. Association for Computing Machinery
Schwabish JA (2014) An economist's guide to visualizing data. J Econ Perspect 28(1):209–34
Shneiderman B (2000) The eyes have it: a task by data type taxonomy for information visualizations. Proceedings of IEEE symposium on visual languages, 03
Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017) Linear discriminant analysis: a detailed tutorial. AI Commun 30:169–190, 05
Todorovic D (2008) Gestalt principles. Scholarpedia 3(12):5345 (**revision 91314**)
Valdes JJ, Barton AJ (2006) Virtual reality spaces for visual data mining with multiobjective evolutionary optimization: Implicit and explicit function representations mixing unsupervised and supervised properties. In 2006 IEEE International conference on evolutionary computation, pp 1442–1449
Valdés JJ, Barton AJ (2007) Visualizing high dimensional objective spaces for multi-objective optimization: a virtual reality approach. In Proceedings of the IEEE congress on evolutionary computation, CEC 2007, 25–28 September 2007, Singapore, pp 4199–4206. IEEE
Valdés JJ, Romero E, Barton AJ (2012) Data and knowledge visualization with virtual reality spaces, neural networks and rough sets: application to cancer and geophysical prospecting data. Expert Syst Appl 39(18):13193–13201
van Dam A, Laidlaw DH, Simpson RM (2002) Experiments in immersive virtual reality for scientific visualization. Comput Graph 26(4):535–555
van der Maaten L, Hinton G (2008) Viualizing data using t-sne. J Mach Learn Res 9:2579–2605, 11