Eurographics Workshop on Visual Computing for Biology and Medicine (2023)
C. Hansen, J. Procter, and R. G. Raidou (Editors)

Visual Analytics for the Integrated Exploration and Sensemaking of Cancer Cohort Radiogenomics and Clinical Information

Sarah El-Sherbiny[1], Jing Ning[2,3], Brigitte Hantusch[2], Lukas Kenner[2,3,4,5,6], and Renata Georgia Raidou[1]

[1]TU Wien, Austria [2]Department of Pathology, Medical University of Vienna, Austria [3]Christian Doppler Laboratory for Applied Metabolomics, Austria [4]Comprehensive Cancer Center, Medical University of Vienna, Austria [5]Unit of Laboratory Animal Pathology, University of Veterinary Medicine Vienna, Austria [6]Center for Biomarker Research in Medicine, Austria

Abstract

We present a visual analytics (VA) framework for the comprehensive exploration and integrated analysis of radiogenomic and clinical data from a cancer cohort. Our framework aims to support the workflow of cancer experts and biomedical data scientists as they investigate cancer mechanisms. Challenges in the analysis of radiogenomic data, such as the heterogeneity and complexity of the data sets, hinder the exploration and sensemaking of the available patient information. These challenges can be answered through the field of VA, but approaches that bridge radiogenomic and clinical data in an interactive and flexible visual framework are still lacking. Our approach enables the integrated exploration and joint analysis of radiogenomic data and clinical information for knowledge discovery and hypothesis assessment through a flexible VA dashboard. We follow a user-centered design strategy, where we integrate domain knowledge into a semi-automated analytical workflow based on unsupervised machine learning to identify patterns in the patient data provided by our collaborating domain experts. An interactive visual interface further supports the exploratory and analytical process in a free and a hypothesis-driven manner. We evaluate the unsupervised machine learning models through similarity measures and assess the usability of the framework through use cases conducted with cancer experts. Expert feedback indicates that our framework provides suitable and flexible means for gaining insights into large and heterogeneous cancer cohort data, while also being easily extensible to other data sets.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** • ***Applied computing*** → ***Life and medical sciences;***

1. Introduction

We investigate and present the design and development of a visual analytics (VA) framework for the comprehensive exploration and integrated analysis of radiomic and genomic data with regard to clinical information in a cohort of cancer patients. Our framework supports the workflow of cancer experts and biomedical data scientists for the investigation of cancer mechanisms, on the basis of high-dimensional and heterogeneous cohort data.

Radiogenomics refers to the combined study of imaging-derived features, called *radiomics*, and gene sequencing data, called *genomics*. Combining these information channels is anticipated to be more expressive of the mechanisms of cancer. However, challenges arise regarding the size, heterogeneity, and complexity of the involved data sets. These challenges make the analysis of the available information space tedious for cancer experts and hinder the exploration and sensemaking of patient information. This is further hampered when additional clinical information is included in the analysis. In the context of radiogenomics analysis combined with clinical data, visual analytics (VA) approaches offer promises for tumor profiling. However, VA approaches bridging radiogenomic and clinical data in an interactive visual framework are lacking.

Our VA framework enables cancer experts to highlight correlations and patterns in the data, supporting the interactive stratification of patient cohorts. Additionally, it facilitates the integrated analysis of radiomic features together with genetic mutation and clinical data. This is anticipated to help experts identify mechanisms that may have an impact on the treatment process. We follow a user-centered strategy and integrate domain knowledge into a semi-automated analytical approach based on unsupervised machine learning. Our interactive visual interface further supports the exploratory and analytical process in a *free* and a *hypothesis-driven* manner. Although we showcase the capabilities of our framework on a prostate cancer scenario, our approach is not bound to this scenario and is extensible to other data sets.

In contrast to previous approaches, which focus either on imaging-derived features (e.g., [RvdHD*15, MWH*20, GDKB17]) or on genomic information (e.g., [LSKS10, LSS*12]), we support the holistic exploratory analysis of radiogenomic features with respect to clinical data in a *unified, extensible framework*. Our framework allows users to examine different methods to reduce the feature space and recluster the data or subsets of the data on demand. We expect that these data set combination and flexible analysis capabilities support domain experts to gain new insights into can-

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/vcbm.20231220

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

cer mechanisms. To the best of our knowledge, radiomics and genomics have never been bridged in a visual interactive framework for knowledge discovery and hypothesis confirmation before.

2. Clinical Background: Radiogenomic Analysis

The mechanisms and risk factors of prostate cancer are still not fully understood. Besides age, risk factors of prostate cancer comprise family history, ethnicity, obesity, and environmental factors [PCKA*17], while inherited gene mutations are also reported as a frequent cause [Don06]. Diagnosing the disease includes examination, biopsies of the prostate and imaging tests such as computer tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). A definitive diagnosis of prostate cancer can currently only be made by histological analysis, which requires invasive procedures such as biopsy or surgery.

Radiomics refers to quantitative features extracted from medical images, obtained from radiological imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET) [KGB*12, ZLVL20]. These features are derived using advanced image analysis techniques and are used to characterize the tumor phenotype, assess treatment response, and predict clinical outcomes. Radiomic features include shape-based, intensity-based, and texture-based features, among others. These features capture aspects of the tumor's spatial distribution, shape, intensity variations, and texture patterns.

Genomics involves the analysis of the exact DNA sequence in the genome of cells or tissues to determine changes in the encoded proteins that affect their biological function [Chr12]. It is related to genetics that considers individual genes and their inheritance throughout the generations but deals with the complete set of genes in the cell or an organism. Genomic analysis of tumors aims to provide information on the behavior of cancer that affects the growth and, hence, the treatment process. For prostate cancer, it is performed on a sample of prostate tissue gained from needle biopsy or from the tissue of the whole prostate, when it is removed from the patient by surgery. As a result, genomic data is retrieved through DNA sequencing for further analysis [NHG19].

Clinical data includes demographic data such as the age, weight, or Body Mass Index (BMI) of patients. Moreover, it consists of scores for patient management such as the Prostate-Specific Antigen (PSA) or the Gleason score (GS). The former is derived from blood tests, while the latter through visual inspections of the tissue morphology of biopsies by pathology experts. Clinical data also comprise prognosis information such as the Biochemical Recurrence (BCR), a rerise of the PSA value that might be an indication of the disease progression [LS09]. Finally, tumor staging determines whether cancer cells have developed or spread within the prostate or to other parts of the human body. All these values (tumor staging values, BCR, GS, and PSA) are combined to determine the D'Amico Risk stratification score [HNHP07].

Each of the three datasets of radiomic, genomic, or clinical data includes indications of cancer [LXNR19, SRY*21]. Therefore, a combined analysis of these three datasets opens the potential to support clinical experts in understanding their complex and heterogeneous data. The analysis of radiogenomic data together with clinical data is expected to improve clinical decision support and to assist the diagnosis and prognostic assessment for diseases, including cancer treatment [LXNR19, SRY*21].

3. Related Work

Visual analytics for radiomics involves the analysis and visualization of radiomic data. Different tools and techniques have been developed to analyze radiomic and imaging-derived data. These include RadEx [MWH*20], which identifies and visualizes relations between radiomic tumor profiles and clinical or histological markers, IIComPath [CCW*21], which employs imaging-derived features to support hypothesis generation through interactive analysis of patient groups, and the system by Raidou et al. [RvdHD*15] to explore and analyze the features space of imaging-derived data of tumor tissue characteristics. Other approaches, such as iVAR [YJY*17], are used to analyze radiomic data of patient cohorts through filtering [BBJ*17]. Contrary to these approaches, Gutenko et al. [GDKB17] use radiomic features to support the alignment of temporal organ data. All existing approaches analyze radiomic data either separately or in relation to clinical data. Several techniques are employed for dimensionality reduction, clustering methods are used to identify patterns in the data, and statistical or machine learning techniques help compare subclusters. Visualization plays a crucial role in these approaches, with scatterplots, heatmaps, parallel coordinate charts, and bar plots being commonly used. Interactive features like data filtering, selection, and zooming allow users to explore the data and gain insights.

Visual analytics for genomics involves the analysis and visualization of genomic data. The most prominent challenges include the visualization of long sequences and sparse distributions, the interaction between distant sequences, and the diversity of data types. Several approaches have been developed to address these challenges. ClinOmicsTrail combines genomics with clinical data for breast cancer decision support [SKT*19], while Caleydo integrates gene expression data with biological pathway models to interpret individual effects and identify disease subtypes [LSKS10]. StratomeX analyzes genomic data sets in combination with clinical data to identify subtypes [LSS*12]. Caleydo StratomeX focuses specifically on breast cancer patients and overcomes limitations in identifying characteristic genes of cancer subtypes [TLS*14]. Nguyen et al. [NNH*14] allow a patient-to-patient analysis by providing an overview of the patient population in the similarity space. These approaches mainly focus on genomic data combined with clinical data and do not incorporate radiomic data or specifically target prostate cancer data sets. The analysis of genomic data involves so far filtering genes, applying clustering algorithms, and performing statistical tests to identify significant differences. Dimensionality reduction techniques are used to reduce the complexity of the data, and visualization revolves around the use of scatterplots, matrix heatmaps, genomic coordinates, and networks.

Visualizing clinical data is crucial in research and healthcare, whether these range from demographic information to disease registries and clinical trials. Bernhard et al. [BSM*15] visualize patient histories in prostate cancer research, Müller et al. [MSO*20] work with patient-specific data for laryngeal cancer decision support, and Angelelli et al. [AOH*14] analyze brain measurements

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

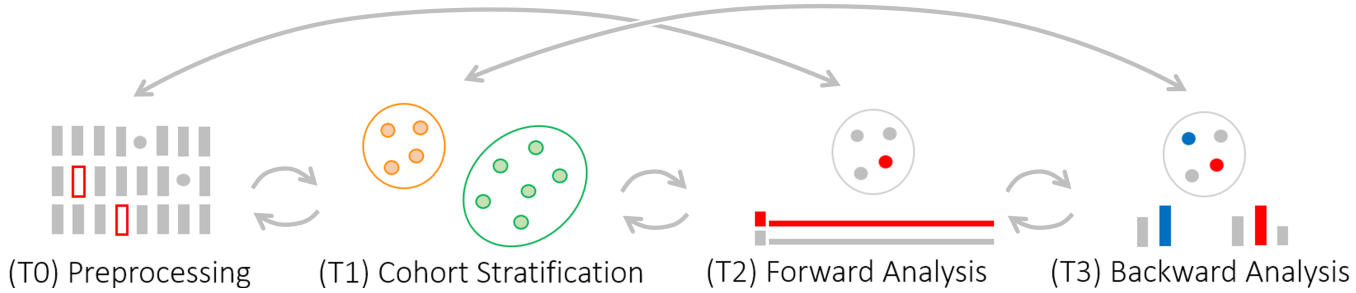

Figure 1: *Main tasks of our analytical workflow.* ***(T0) Preprocessing*** *to enhance the data quality and facilitate automated analysis and visualization.* ***(T1) Cohort stratification*** *to identify and visualize patterns in the high-dimensional and complex data.* ***(T2) Forward analysis*** *to freely explore the data for knowledge discovery.* ***(T3) Backward analysis*** *to assess and refine the hypothesis on the underlying data.*

for cognitive aging. They employ various visualizations such as data cubes, Bayesian networks, scatterplots, and bar charts to offer insights and support decision-making for different scenarios.

Although radiogenomic data analysis has shown potential in identifying correlations, predicting cancer, and providing precise prognoses, the complexity and high dimensionality of radiogenomic data require advanced frameworks and algorithms. *Combining radiomic, genomic, and clinical data* in an integrated visual analytics system has not been proposed before. Zanfardino et al. [ZCP*21] presented a framework for combining and analyzing radiogenomic data in breast cancer patients—yet, their approach lacks interactive exploration and knowledge discovery features. Conversely, we provide an interactive interface that allows users to perform a selection and repeated analysis of patients or features of interest. In our interface, the user is encouraged to freely interact with visual representations to support data sensemaking.

4. Data — Users — Tasks Analysis

Data characteristics: The available prostate cancer data sets—hereby employed to showcase our framework—consist of 153 radiomic PET/MRI-derived features, 10 307 gene mutations, and 18 clinical parameters from a cohort of 89 patients. Challenges in the data include the high dimensionality, and also the missingness in patient values and different data types, which hampers automated data analysis. The clinical data consist of demographic information and patient management scores determined through clinical assessments. These include categorical data such as the pre-operative therapy methods or the International Society of Urological Pathology (ISUP) grade that represents a disease grouping based on the derivation of extracted cells from normal cells [SDE*16].

Target users: Our target audience comprises *cancer experts* and *biomedical data scientists*. Cancer experts include pathologists, biochemists, and nuclear medicine physicians aiming to investigate and understand the mechanisms behind cancer data. They desire to gain new knowledge from the data to identify clinical markers relevant to cancer research, diagnosis, and treatment. Cancer experts also strive to assess the correctness of any hypothesis or biological mechanism they have in mind. Biomedical data scientists and informaticians have, on the other side, different goals. Primarily, they ought to attain insight into the underlying algorithms by interactively comparing different analysis components to investigate and understand their suitability and impact on the analysis results.

Tasks: Through structured interviews with our target users, we defined collaboratively their goals and summarized them in an analytical workflow with four tasks **(T0–T3)**. These are illustrated in Figure 1. The data sets need to be prepared to enhance their quality and facilitate automated analysis and visualization. Therefore, our first task represents the **data preprocessing (T0)**. This includes identifying and resolving inconsistencies in the data [GGAM12], handling mixed data types and missing data values [ANI*20], detecting outliers, and scaling the data. Subsequently, we enable the identification of patterns in the high-dimensional and complex data through **cohort stratification (T1)**. This entails data analysis to identify groups of patients with similar radiomic, genomic, or clinical profiles. It requires a dimensionality reduction and clustering step of the data to reduce the high-dimensional data into two dimensions that we visualize on screen. We support users in understanding the cohort stratified patterns through the **forward analysis (T2)** step that allows users to freely explore the data for knowledge discovery. Users can interact with the data by selecting and processing patient subsets on the visualization depending on their radiogenomic or clinical profiles. Furthermore, users gain insights through identifying and interacting with characterizing and differentiating features, patient distribution values, and the most frequent gene mutations. The last step of the workflow represents the **backward analysis (T3)** to assess and refine any present hypothesis on the underlying data. This includes interactive filtering of patient features on the visualization and the comparison of features impacting the clustering based on a specified condition.

5. Visual Radiogenomics Analysis

Figure 1 illustrates the steps of our visual radiogenomics analysis workflow. The workflow starts with data processing **(T0)**, which serves as an input to the cohort stratification **(T1)**. Users can repeatedly apply the cohort stratification on patient subsets or examine different parameters on data subsets. Through brushing the scatterplot points, users can compare patient subsets and their characteristics, differences, and distributions in the different dimensionality reduced and clustered spaces. The result of the cohort stratification **(T1)** can be used as an input to the forward analysis **(T2)** or backward analysis **(T3)**. The forward analysis **(T2)** can be applied to the

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

data selected through a hypothesis that results from the backward analysis **(T3)**, on the automatically created data clusters resulting from the cohort stratification **(T1)**, or on the unstratified data **(T0)**. Similarly, the backward analysis **(T3)** can be applied to the result of the cohort stratification **(T1)** or a processed subset resulting from the forward analysis **(T2)**.

5.1. Preprocessing (T0)

The data analysis requires automated processing of the data sets through statistical measures or unsupervised machine learning algorithms. This comprises data cleansing, encoding, imputation, outlier detection, and scaling as part of the preprocessing.

Data cleansing: Our data can be considered *dirty* due to missingness in feature values, and inconsistencies in numerical and qualitative features that need to be handled properly for a correct data interpretation [GGAM12, KCH*03]. Therefore, we apply data cleansing to identify and correct errors and inconsistencies in the data [RZ19, GGAM12]. We perform cleansing by using replacement rules that we define and confirm together with our cancer experts [KCH*03].

Data encoding: The data sets are composed of *mixed data* types of qualitative and numerical values, which hamper the automated data analysis and visualization. To encode qualitative values as numbers with minimal processing, we apply the one-hot encoding algorithm on categorical features [CV22]. This indicates that for each categorical variable, we convert each value into a new column and assign a binary value of 1 or 0 to those columns.

Data imputation: We predict and replace *missing data* values through imputation techniques [van18, ANI*20]. Single imputations replace the missing values with the mean or median of the respective feature. This reduces the variability of the distribution and leads to a biased estimation [ANI*20]. Advanced options that deliver an appropriate level of accuracy and bias use multiple imputation algorithms. In contrast to single imputations, multiple imputations consider dependency between variables [ANI*20, van18]. To identify the most suitable imputation strategy for each feature in our data, we test and compare the error metrics of different imputation methods on the basis of complete data. We compare single imputation methods by using the mean, median, or most frequent value of the feature. Furthermore, we apply linear regression, and multiple imputation methods, such as Multivariate Imputation by Chained Equations (MICE) and k-Nearest Neighbors (KNN).

To assess the robustness of the algorithms and assess the generalizability of our approach, we simulate missingness percentages ranging from 5 % to 95 % in the complete data subset that we divide into a training and test set. We apply all imputation methods directly on the original data, not on partly imputed values, to compare them with each other. As an evaluation metric of the imputation results, we use the Root Mean Square Error (RMSE) between the imputed and the true value in the training set. Figure 2 illustrates the RMSE (on the y-axis) for the continuous *POST-PSA* feature. The simulated missingness percentages range from 5 % to 95 % (on the x-axis). The plot compares the RMSE of KNN and MICE applied on all features with the BEST method that represents the imputation method with the smallest RMSE of all imputation methods we tested for the specific features. In the case of the *POST-PSA* feature, using the most frequent value is identified as the best suitable for the given feature. For each feature, we set the best option as the default imputation method in our interface. This is an informed choice that users can change on demand—if an imputation alternative should be used.

Outlier detection: Outliers affect the accuracy and stability of automated data analysis and influence clustering outcomes [LLWF21]. Therefore, we allow users to identify, highlight, remove outliers, or compare them with the remaining data points *on demand*, when combined with the upcoming forward **(T2)** and backward analysis steps **(T3)**. We deploy the unsupervised machine learning method *isolation forest* for global outlier detection and removal, given its linear time complexity and low memory requirement [LTZ08]. For local outliers, we provide the density-based *local outlier factor* algorithm [CZD19]. Per default, no outlier detection is applied to the data. Users can choose through the interface to apply local or global outlier removal on demand.

Data scaling: Values of data sets include measurements in different units. Analyzing each measurement in its data-dependant scale affects the outcome of the analysis process as values dominate over others [THFM14]. Data standardization is preferable for data with a Gaussian distribution and outliers, as it improves the signal-to-noise ratio and the discrimination power of the data sets [Ng17]. If the data distribution is not known or non-Gaussian, data normalization is advantageous [Ase22]. It eliminates bias in features with high values compared to features with low values [Ase22]. Based on the characteristics of our data, we set normalization as the default option and allow users to change it on demand.

5.2. Cohort Stratification (T1)

We identify and visualize patterns in the preprocessed data sets through cohort stratification. This process divides patients into meaningful groups based on similarities in their radiogenomic and clinical profiles. It requires dimensionality reduction, clustering, and visualization of the data.

Dimensionality reduction: Extracted radiogenomic features contain redundant and unnecessary information that lead to overfitting

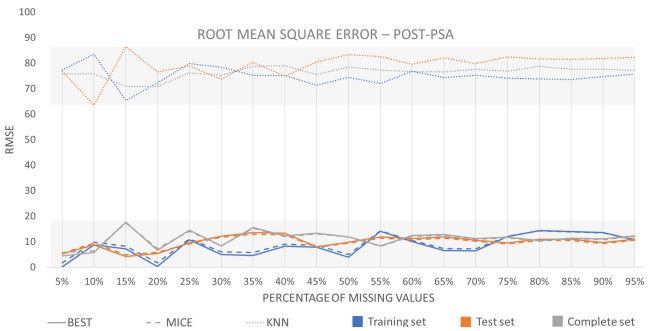

Figure 2: *Comparison of the RMSE of the KNN, MICE, and BEST imputation methods for missingness percentages ranging from 5 % to 95 %, exemplified on the continuous POST-PSA feature.*

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

and hamper the generalizability of the machine learning models for new data [SRY*21, LXNR19]. Eliminating features that lack robustness against variability sources avoids overfitting [LLD*17]. This is performed by applying dimensionality reduction algorithms on the data [LLD*17]. A reduced dimensionality of the data helps to maintain imaging characteristics that strongly correlate with clinical features [SRY*21, LXNR19, SJG*22].

We test and compare the algorithms Principal Component Analysis (PCA), Factor Analysis of Mixed Data (FAMD), Multidimensional Scaling (MDS), Uniform Manifold Approximation and Projection (UMAP), and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA is a linear method that focuses on the global data structure and highlights interclass differences [XWY*21, EMK*21]. FAMD combines PCA with Multiple Correspondence Analysis (MCA) and is applicable to complex data of mixed types. Therefore, FAMD is better suitable than PCA for data sets that contain mixed datatypes of qualitative and quantitative values [GMS*15]. MDS and UMAP are non-linear methods that balance the global and local structure of the data, while UMAP assumes a uniform data distribution [XWY*21, EMK*21]. A non-linear method that focuses on the local data structure is t-SNE which minimizes the divergence between two distributions [XWY*21, VDM14]. We select t-SNE as the default dimensionality reduction technique as it leads to the best cluster separation and forms visual clusters in our data. We further allow users to test the outcome of different dimensionality reduction methods and compare the resulting patterns on the scatterplot. Users can combine methods by repeatedly applying them to patient subsets to get the advantages of their different characteristics, such as progressively exploring the local and global structure of a data subset.

Clustering: To improve the understanding of complex data sets, clustering summarizes their underlying information based on their similarities [SEK03]. Clustering supports the analysis of high-dimensional data and helps users to gain insights into the data structure [KBZ*21]. To overcome the curse of dimensionality, we apply clustering on the reduced 2D data space [SEK03, KBZ*21].

We tested distribution-based Gaussian Mixture Models (GMM), density-based Ordering Points To Identify the Clustering Structure (OPTICS), hierarchical clustering, and the centroid-based algorithms k-means and mean-shift. Distribution-based approaches, such as GMM, are suitable for Gaussian-distributed data. Density-based methods, such as OPTICS, are suitable for arbitrary-shaped distributions. However, they cannot deal with high-dimensional data or varying densities and do not assign outliers to clusters. Hierarchical clustering is time-demanding and sensitive to parameterization, e.g., to the linkage criterion [RG19]. K-means or mean-shift lead both to the same clustering result on our t-SNE reduced data that highlights the available visual clusters in the data. We determine the number of clusters for k-means with the elbow method [Tho53] and evaluate the methods based on the similarity measures and by considering distributions within clusters.

Table 1 shows the resulting cluster separation scores. We calculate the Silhouette coefficient [Rou87], the Calinski-Harabasz index [CH74], and the Davies-Bouldin index [DB79] for all clustering and dimensionality reduction methods. The higher the Silhouette Coefficient and the Calinski-Harabasz index, the better the clusters are defined. On the contrary, a lower Davies-Bouldin index is related to a better separation of clusters. While the scores of FAMD and PCA indicate a good cluster definition, these scores originate from single data points assigned to their own clusters. Besides the high separation scores of t-SNE, this method distributes the points well among the identified visual clusters. Therefore, our preset uses the k-means algorithm, but knowledgeable users can experiment with the integrated clustering alternatives.

Table 1: *Cluster separation scores: the first line for each clustering method represents the Silhouette coefficient, the second line depicts the Calinski-Harabasz index, and the third line represents the Davies-Bouldin index. Based on these scores and by considering the data distributions within clusters, the best default choices for our data are t-SNE with k-means.*

	t-SNE	MDS	FAMD	UMAP	PCA
k-means (k=2)	0.89	0.42	0.97	0.65	0.89
	414.15	37.70	104.65	111.35	118.47
	0.40	1.30	0.03	0.81	0.02
mean-shift	0.89	0.89	0.94	0.60	0.98
	414.15	414.15	3671.70	98.33	29195.39
	0.40	0.40	0.13	0.79	0.00
hierar. (4 clust.)	0.52	0.36	0.94	0.50	0.83
	255.10	38.67	3671.70	86.18	44959.24
	0.92	1.06	0.13	0.89	0.31
hierar. (6 clust.)	0.47	0.47	0.49	0.44	0.58
	219.86	45.97	4467.11	82.32	85453.22
	0.89	0.86	0.44	0.88	0.37
OPTICS	0.21	-0.39	-0.44	0.21	-0.31
	10.51	0.53	2.08	33.47	0.20
	1.41	4.32	2.01	2.84	2.09
GMM	0.50	0.36	0.94	0.56	0.81
	239.68	33.19	2408.29	90.58	44417.79
	0.88	1.13	0.21	0.82	0.32

Visualization: To present the patterns identified in the data, we visualize the reduced and clustered data through a scatterplot (Figure 9A). We additionally show kernel-density estimation contours on the scatterplot that represent an estimation of dense regions. We also shade dense areas in a blue color to provide a visual indication of the cluster density and separation. The contours are determined through the kernel density estimate (KDE) of the data points [DLH11]. Visualizing the high-dimensional data through a scatterplot matrix or a heatmap matrix of all feature correlations instead, would capture only pairwise relations and hinder the identification of complex data patterns going beyond 2D.

5.3. Forward (or Free) Analysis (T2)

After identifying and visualizing the cohort stratification, we support users in exploring and understanding these patterns for further knowledge discovery. We provide interaction possibilities with the identified patterns through patient selections, subset processing, and subset comparisons. Additionally, we propose heatmap and barplot views for the identification of features that contribute the most to the differentiation or characterization of the clustered data.

Analysis of patient stratification: To explain why the identified clusters are similar or different, we analyze the obtained patient

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

stratification. By applying Shapley Additive Explanations (Shap) to the clustering result [LL17], we predict features that impact the clustering the most and indicate the cluster characteristics. Furthermore, we need to determine pairwise differentiating features between the identified clusters. Thus, we employ Linear Discriminant Analysis (LDA) or Stochastic Gradient Descent (SGD) [OH21]. SGD works with dense or sparse data, which matches our sparse genomic or dense radiomic data. SGD further reveals more significant differences for our default preset of t-SNE, especially for the genomic data. Therefore, we set SGD as the default method and allow users to explore LDA on demand.

We determine the top 0.5% characterizing and differentiating features of each cluster and combine them in a heatmap to provide users an overview of features that impact the clustering result the most, as shown in Figure 3. In the first two rows, we employ luminance to encode the feature importance (denoted in the columns) for each of the clusters, while the third row highlights features representing the pairwise cluster differences. In both cases, darker colors indicate a higher contribution (either to importance or to difference). Figure 4 shows the heatmap view for more than two clusters. On-demand, we offer a bar chart view that allows users to filter features by the radiomic, genomic, or clinical data, as depicted in Figure 5. Users can select any feature through the heatmap or bar chart views to highlight the feature values for all patients on the scatterplot view. In Figure 6 the values of the morphological diameter feature from the radiomic data is highlighted on the scatterplot.

Patient selection: We support the understanding and exploration of patterns in the data by allowing users to process or compare patient subsets. The selection of these subsets can be performed by formulating a hypothesis as part of the backward exploration **(T3)** or through a lasso selection as exemplified in Figure 7 [BC87]. Compared to a rectangular selection, a lasso provides flexible selection that is not necessarily connected or neighbored in the scatterplot.

Top gene mutations: Cancer experts are interested in understanding and identifying relevant gene mutations. Therefore, we interactively determine and show the top gene mutations that occur for most of the patients for any cohort subset selected on the scatterplot by the user, as shown in Figure 7, using a ranked bar chart.

Navigation and tooltips: Navigating the scatterplot by zooming in or out on the scatterplot helps to open up the data points and to explore dense patient regions in more detail. To enhance the understanding of the clustering result and the features impacting the clustering the most, we show patient distributions of heatmap features on demand through a tooltip. Additionally, we display patient scores on mouse hover over a patient in the scatterplot, as demonstrated in Figure 8.

Presets and parameters: Based on our quantitative assessment of the imputations and cohort stratification options, we offer presets identifying the most suitable parameters for the underlying data sets (Figure 9E). Furthermore, users can manually change any analysis

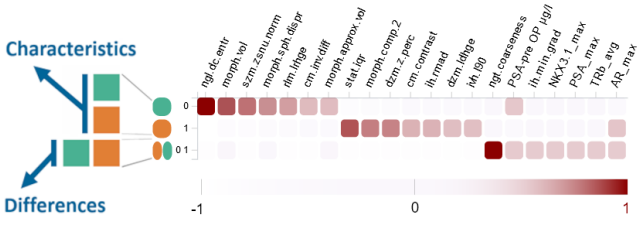

Figure 3: *Heatmap view on cluster characteristics (first two rows) and differences (last row). Most contributing features are denoted with darker red, and the resulting clusters are denoted with two distinct hues (green and orange).*

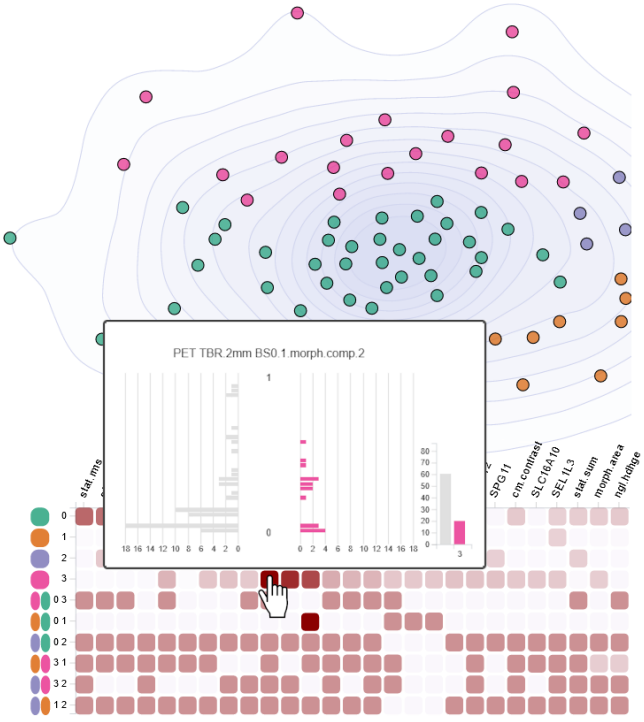

Figure 4: *Heatmap view with four clusters. It combines the cluster characteristics (first four rows) and pairwise cluster differences (last six rows) for the features that impact the clustering the most (columns). The pyramid plot on the tooltip decodes the feature distribution of patients in the selected (pink) cluster compared to patients outside this cluster. For the cluster differences, the pyramid plot illustrates the feature distribution of two cluster pairs.*

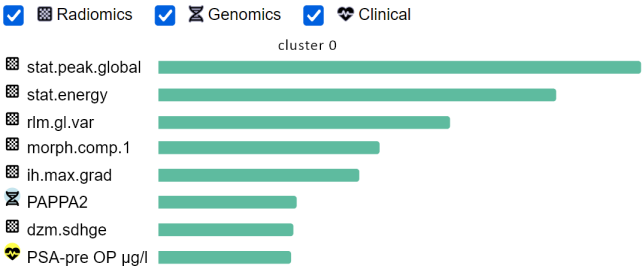

Figure 5: *Features characterizing one of the identified patient clusters, sorted by their importance in forming this cluster. The glyphs in front of each feature indicate the respective dataset from the checkboxes above that serve for filtering and as a legend.*

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

option through the interface to explore how it affects the revealed patterns in the data. We anticipate that changing the presets should be relevant only for the biomedical data scientists, while the cancer experts would proceed using the default options.

5.4. Backward (or Hypothesis-driven) Analysis (T3)

The backward analysis provides users with functionality for the assessment of the correctness of a hypothesis in mind. Users can select data subsets to include or exclude in the analysis, and interactively filter, select, and compare data subsets [Shn94]. This enables the identification of thresholds for hypothesis assessment or the determination of new hypotheses for the underlying data. The resulting data subset is selected on the scatterplot and can be used as an input for further forward **(T2)** or backward **(T3)** exploration.

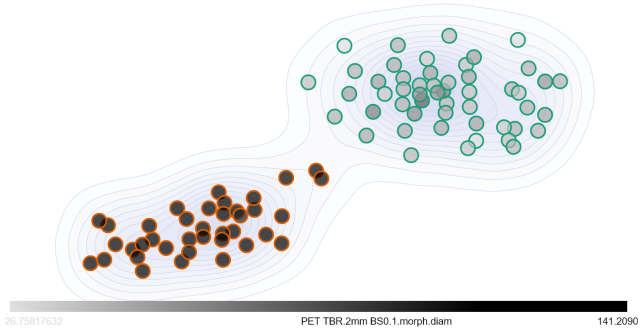

Figure 6: *Values of a radiomic feature highlighted on the scatterplot. The values on the top right cluster (indicated by light gray data points) are lower than the values on the bottom left cluster (indicated by dark gray data points).*

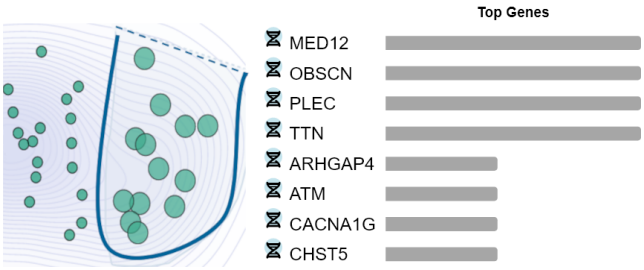

Figure 7: *Top gene mutations of an active lasso selection (enlarged points) on the scatterplot, shown in a ranked bar chart. The selected data points are enlarged upon selection for visibility reasons. In this example, the gene MED12 occurs for most of the patients in the selected cohort subset.*

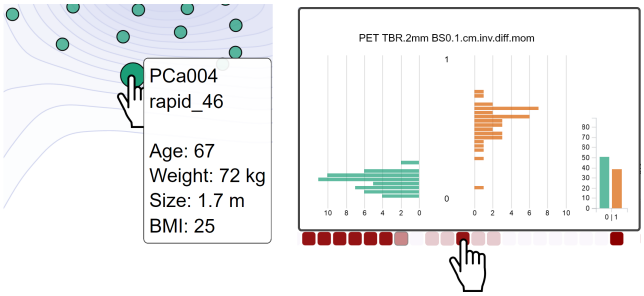

Figure 8: *On the scatterplot, the tooltip reveals clinical patient scores on mouse hover (left), while on the heatmap it shows patient distributions of a selected feature (right).*

Hypothesis assessment: Users can visually assess the correctness of any hypothesis in mind by selecting features of interest. We provide an overview of the feature distributions among all patients through the display of a histogram of the selected features. Users can specify feature ranges for the hypothesis assessment visually by interacting with sliders on the feature histograms, which leads to interactive data filtering on the scatterplot. To form a hypothesis, features can be combined through logical **`and`** or logical **`or`** operators, while the **`not`** operator can be expressed visually through the histogram bars. An example is shown in Figure 10 (Hypothesis).

Feature and patient subset processing: To identify new patterns in patient subsets, we allow users to repeat the cohort stratifications based on a hypothesis. This results in a representation of patient subsets that fulfill a specified condition. Furthermore, users can select to include or exclude any feature subsets on demand. After this selection, the previous steps of dimensionality reduction and clustering will be repeated.

Hypothesis-based feature comparison: To identify the characteristics and differences of patients based on a hypothesis, we allow users to compare patients that match a condition against patients that do not match it through the heatmap. Patients that fulfill a hypothesis are assigned to one cluster and are subsequently compared against patients not fulfilling it. The characterizing or differentiating features of the patients are updated in the heatmap (Figure 3) and barplot views (Figure 5) for further data exploration.

6. Interface and Framework Components

Figures 10 and 11 give an overview of our framework and its components. We divide the interface into three views (A–C) and employ additional tooltips on demand (D) to reveal additional information without cluttering the view (Figure 9). Furthermore, we integrate five tabs (Figure 10) to visualize feature *values*, the most significant features of the *clusters*, or the *top genes*. Our tabs also consist of a *processing* view to specify data subsets for the analysis or a *hypothesis* view to visually assess or refine hypotheses interactively.

(T0–T3) Component A shows the cohort stratification scatterplot, where each scatter point represents one patient. The color of the points indicates cluster assignment and is used consistently in all other views. On the top right of the view, three buttons are revealed when a selection of patients is made on the scatterplot. These allow users to zoom into the selection, to process and stratify only the points of this selection, or to set an active selection to further investigate its features through the heatmap (C) or in the *clusters* and *top gene* views (B).

(T2–T3) Component B consists of five tabs to visualize feature values selected on the scatterplot, a ranked list of the top clustering features, or the gene mutations that occur the most for an active selection on the scatterplot. A detailed view of them is given in Figure 10. The *values view* **(T2)** demonstrates distributions of patient

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

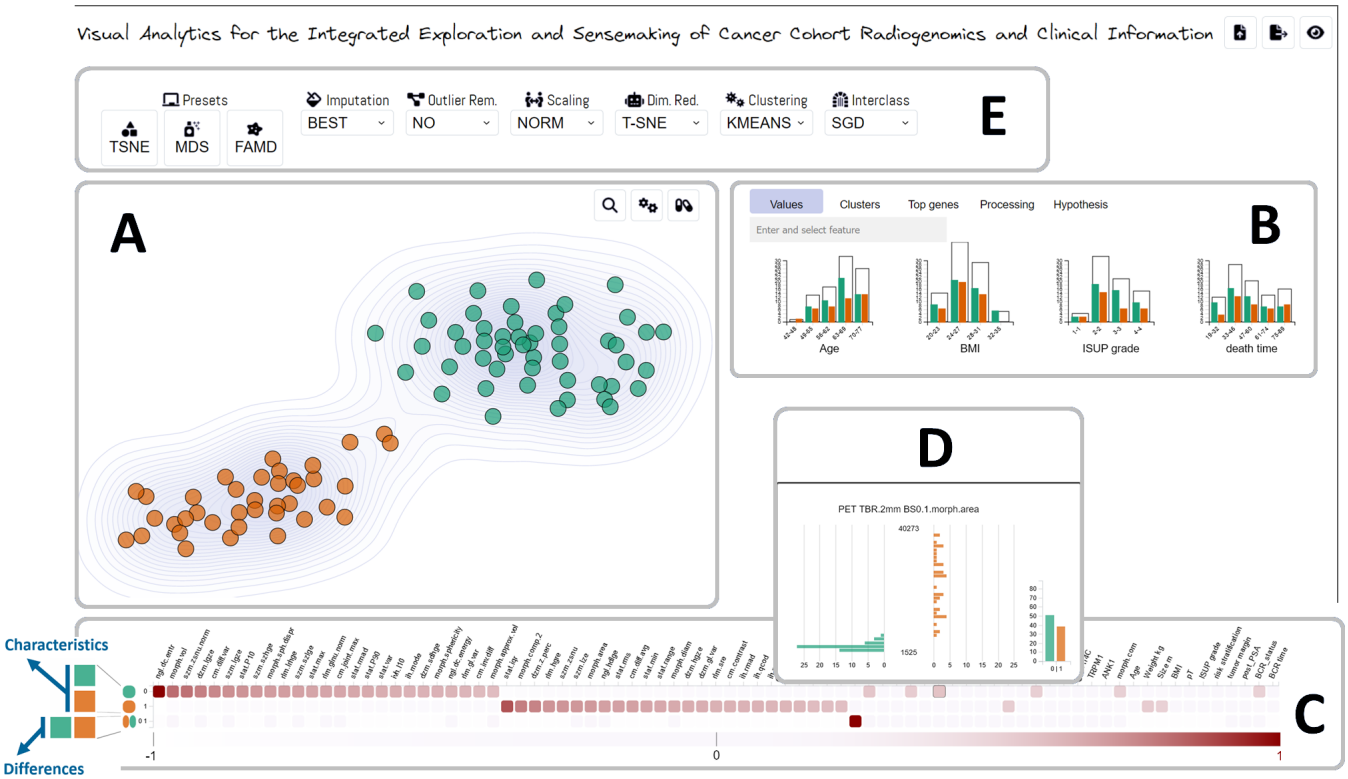

Figure 9: *Main views of our visual radiogenomics analysis framework.* ***(A) Scatterplot view*** *for the result of the cohort stratification.* ***(B) Tab views*** *for the detailed data exploration (see details in Figure 10).* ***(C) Heatmap view*** *for characterizing and differentiating clustering features.* ***(D) Tooltip view*** *with patient distributions of a selected feature.* ***(E) Advanced analysis*** *options are displayed on demand.*

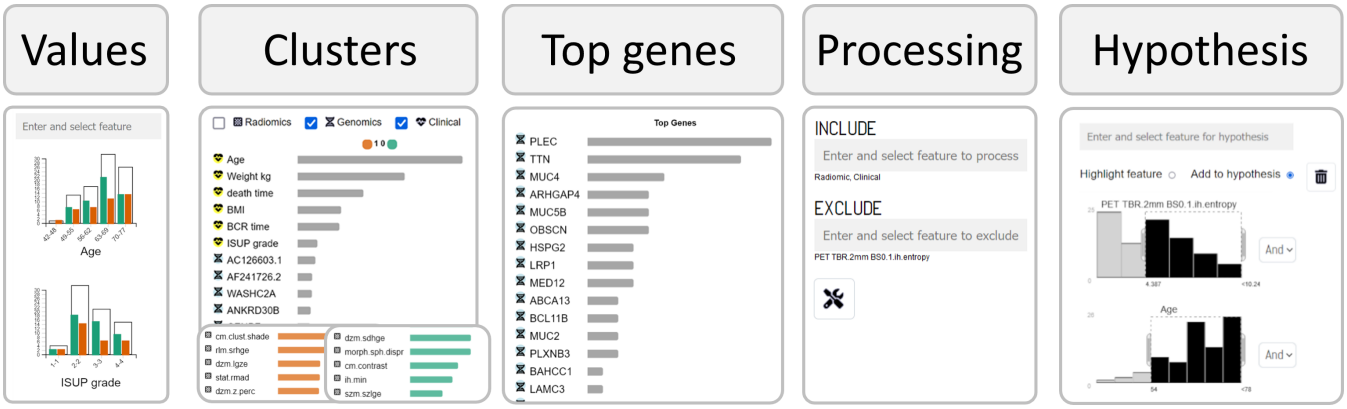

Figure 10: *Detailed view of the five tabs of our framework. These represent component (B) of Figure 9 and are composed of tab views to visualize feature* ***values****, the most predictive features of* ***clusters****, or the* ***top genes*** *of the data. The* ***processing*** *view specifies data subsets for the analysis and the* ***hypothesis*** *view provides a visual assessment and refinement of any hypothesis statements in mind.*

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

values for an active selection on the scatterplot. These distribution values are grouped per cluster, which serves as initial feedback on the data and the clustering scores. The *clusters view* **(T2)** presents a ranked list of features that pairwise differentiate between two clusters. This serves as a detailed view of the heatmap features that allows filtering the features regarding their radiomic, genomic, or clinical data. Instead of the differences, users can select to show the characteristics of one of the identified clusters and explore these features further. In the *top genes view* **(T2)** a ranked list of gene mutations is presented. It is displayed by default for the complete data set but can be filtered through any selection made on the scatterplot. The *processing view* **(T3)** allows the selection of any feature subset of the radiomic, genomic, or clinical data for the analysis process. By default, all features of all data sets are integrated that can be filtered on demand. In the *hypothesis view* **(T3)**, users can highlight any feature combination of interest to a hypothesis and assess its correctness for the underlying data. The feature ranges can be interactively and visually defined through sliders.

(T2) Component C gives an overview of the features characterizing and differentiating the clusters through a heatmap. It consists of values normalized between -1 and 1 that are colored through a linear scale ranging from white to red color to make the ranking scores comparable. The higher the value is, the higher its impact on the current clustering on the scatterplot.

(T2) Component D presents a pyramid plot as a tooltip on the heatmap to compare patient distributions between two clusters depicted on a scatterplot. It is visualized for a specific heatmap feature on demand to preserve a clean view.

(T2) Component E offers advanced options on demand. These allow users to reveal the current analysis options and adjust them to investigate how each change affects the patterns, clusters, or top features and gene mutations of the data sets. This is mainly intended for the biomedical data scientists or bioinformaticians.

7. Implementation

We implement our framework as a web application through Python and JavaScript. The *Scikit-learn* library serves for the data encoding, imputation, outlier detection, dimensionality reduction, clustering, and prediction of the characterizing and differentiating features of the identified clusters. For extended imputation capabilities, we use the *Impyute*, *AutomImpute*, *MiceForest*, and *FancyImpute* libraries. We utilize *Pandas* and *Numpy* for the data management and computations on the data. For the UMAP dimensionality reduction, we employ the *umap-learn* library. To reduce the dimensionality of the data through the FAMD algorithm, we utilize the *prince* library. On the frontend, we use the D3 plugins *d3-lasso* for the lasso selection on the scatterplot, *d3-contour* for the density-based contours on the scatterplot, and *d3-tip* to show the heatmap tooltip with the distribution plots of selected features on demand. Our code is available together with a generated toy data set on GitHub, as the prostate cancer data we used is not publicly available: https://github.com/saraheee/VACI.

8. Domain Expert Evaluation

We conducted a total of five structured workshops during the design and development phase of the tool [KGD*19], where a large number of experts (pathologists, biologists, nuclear medicine physicians, data scientists, and medical doctors) provided comments on our dashboard and raised hypotheses of interest. At the end of the design phase, we evaluated the data analysis, knowledge discovery, and knowledge management capabilities of our framework through usage scenarios [IIC*13, LBI*12] that we conducted together with two cancer experts (a biologist and a nuclear medicine physician). The two cancer experts used the tool through a first set of 50 scenarios based on their hypotheses and provided ten additional hypotheses and directions to analyze. Beyond the structured workshops and the usage scenarios, we received ongoing feedback from three cancer experts (a nuclear medicine physician, a biologist, and a pathologist) on the clinical applicability of our tool. The nuclear medicine physician also interacted with our application and provided additional comments. In the following section, we showcase a selection of scenarios to illustrate the knowledge discovery and hypothesis assessment functionalities of our framework.

8.1. Scenario 1: Features Impacting Cohort Stratification

In the scenario shown in Figure 11, the cohort stratification is based on the clinical and genomic data sets, which can be freely specified through the *processing* tab view. The result of the cohort stratification is presented with the scatterplot that reveals two clusters, denoted with orange and green (Figure 11, top). Users are presented with features that impact this clustering the most through the heatmap at the bottom-left view. The heatmap reveals that the pathological tumor (pT) stage is a strong characteristic of the orange cluster and a differentiating feature between the two clusters, as the respective heatmap cells are colored in darker red (see annotation on the heatmap). Under the *clusters* tab view, a detailed view of the cluster differences is shown by default, which offers additional filtering options. In this example, the user highlights the pT feature on the scatterplot that indicates the disease level of cancer. This is performed through a mouse click on the heatmap cell or the respective bar of the feature in the barplot (annotated in red). The values of the pT features are thereby indicated in the scatterplot in light gray color (pT=2) or in dark gray color (pT=3). This reveals that the majority of the patients in the orange cluster (91.30%) have a pT value of 3, while most of the patients in the green cluster (80.95%) have a lower pT value of 2. This indicates that our stratification regime is able to capture and reveal the two different disease levels in the given patient cohorts.

Selecting only the clinical data as an input to the cohort stratification leads to a clearer cluster division based on the pT value, as illustrated in Figure 12. In this scenario, all values in the orange cluster have a pT value of 2, while all values in the green cluster have a pT value of 3. The pT feature is the first top characteristic of the green cluster as shown through its top position on the bar plot and the heatmap view (annotated in red).

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

8.2. Scenario 2: Identification of Gene Mutations

Figure 13 illustrates a scenario for hypothesis assessment and knowledge discovery. In this example, the clinical feature for the ISUP grade and the radiomic feature for the morphological tumor sphericity are selected. The clustering on the scatterplot is performed based on the ISUP grade. Therefore, patients in the orange cluster represent patients with an ISUP grade of 4 or 5, while patients in the green cluster have ISUP grades ranging from 1 to 3. In the *top genes* tab view, the BCOR gene is identified as a feature from the genomic data set that contributes the most for patients that fulfill the specified condition of having a low tumor sphericity and an ISUP grade of 4 or 5. Mutations in BCOR are known as an indication of aggressive cancer types [AFM*19]. By clicking on the BCOR gene mutation in the bar chart, the feature values are highlighted in the scatterplot. The red annotation on the scatterplot depicts the five patients that have this gene mutation expressed. Three of them are dark gray as they have the gene mutation with a higher values, while two of them are light gray as they have the gene expressed with a low value. All of them are located in the orange and in the top right visual cluster of the scatterplot. This way, patients with an indication of aggressive cancer types can be easily identified, and their radiogenomic and clinical characteristics can be investigated.

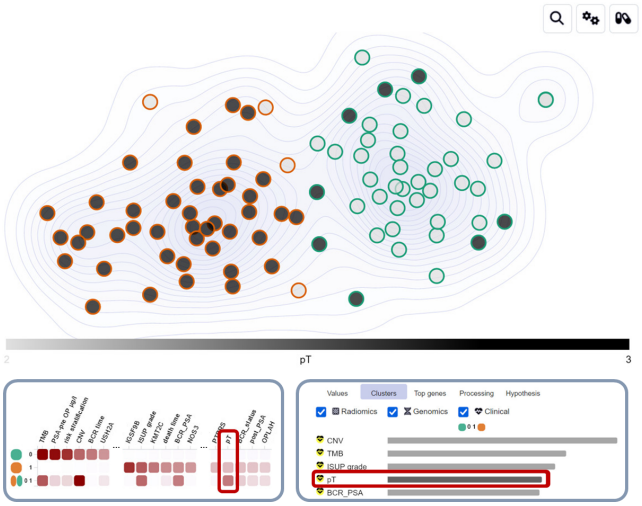

Figure 11: *Highlighting the pT feature on the scatterplot by selecting it through the heatmap or barplot (red annotation). The genomic and clinical features form the input of the cohort stratification. Light gray scatter points have a pT value of 2, while dark gray scatter points have a pT value of 3.*

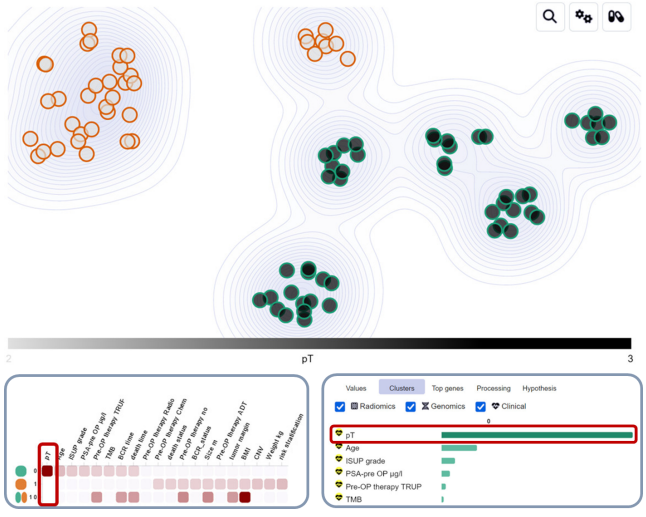

Figure 12: *Highlighting the pT feature on the scatterplot by selecting it through the heatmap or barplot (red annotation). Using the clinical features as an input to the cohort stratification leads to a clear cluster separation on the scatterplot through the pT feature. All patients in the orange cluster have a pT value of 2, while all patients in the green cluster have a pT value of 3.*

8.3. Scenario 3: Generalization to Additional Data Sets

To assess the generalizability of our approach, we extend the set of clinical features and add two new data sets as input, as shown in Figure 14. In addition to the radiomic, genomic, and clinical data, we integrate 17 extended clinical features, 25 immunohistochem-

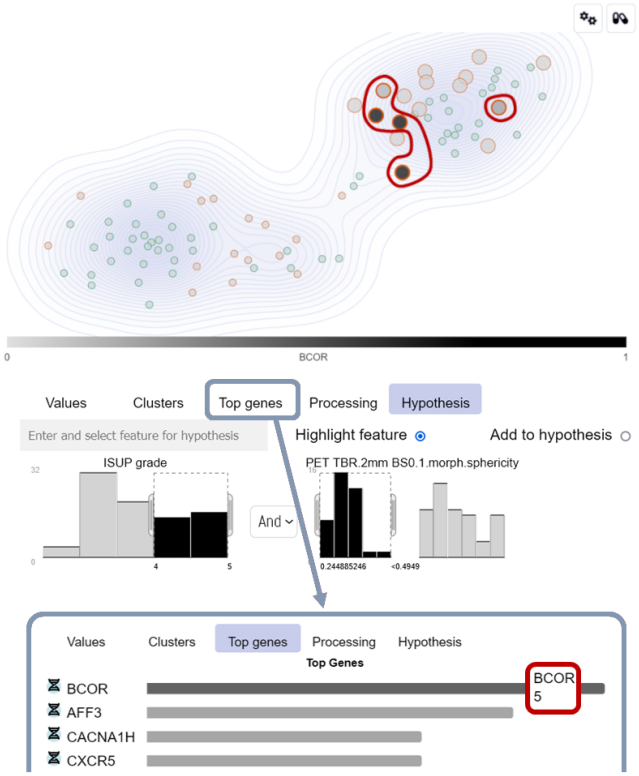

Figure 13: *Interactive hypothesis assessment is performed on the black histograms by moving the sliders to define feature ranges of interest. This reveals gene mutations that occur only for patients that fulfill a hypothesis. Delineated points represent patients with the BCOR gene mutation on a continuous range between 0 and 1.*

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

istry features, and 53 pathway-level features. This confirms that our framework is generalizable to additional data. The scenario reveals two visual clusters on the scatterplot. The heatmap indicates features interesting for the analysis that influence the clustering the most. For example, the radiomic features of the morphological tumor volume is a characteristic of the orange cluster as annotated in red color on the heatmap. Clicking on the heatmap cell of this feature highlights its values on the scatterplot. A mouse hover over this feature opens a tooltip with the distribution plot of the feature among all patients. The tooltip illustrates that patients in the orange cluster have low scores, while in the green cluster, they are almost equally distributed among patients. With this scenario, we showcase that our approach can be easily applied to other cases where additional data sets should be integrated in the analytical workflow.

8.4. Expert Feedback

Our collaborating cancer experts perceived the resulting dashboard as very clear and easy to understand. In our interviews with them, they mention that the "*framework reveals interesting insights on the data and is very helpful*" for them in assessing the correctness of their hypotheses while identifying the role of any feature combination on an interactive visual basis. The values view represents "*an important analysis*" for them to gain insight into the distribution of any active patient selection in their data. They found it "*impressive*" to compare characteristics and differences between patients who fulfill a condition and those who do not fulfill it. Furthermore, it is "*highly interesting*" for them to highlight the feature values of patients indicated through the heatmap or bar plot views on the scatterplot. This helps them to identify features that correlate with one of the visual clusters, which is "*highly interesting and meaningful*" for analyzing their data and generating new hypotheses. They commented that the top genes view of a selection or hypothesis on the scatterplot is "*really essential*" for them in the identification of relevant gene mutations. Finally, they provided us with feedback for future work, which we discuss in the upcoming section.

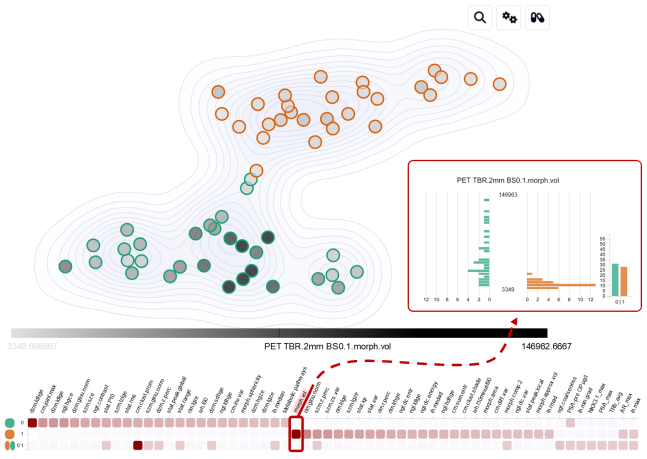

Figure 14: *Adding three additional data sets (immunohistochemistry, pathway level data, and extended clinical features) besides the radiogenomic and clinical data confirms that our approach is easily extensible and generalizable to different data sets.*

9. Conclusion and Future Work

We developed an interactive and flexible VA framework for analyzing complex radiogenomic and clinical data. It combines visualization and automated analysis to help cancer experts and data scientists to gain insights and test hypotheses. The forward and backward analysis capabilities of our framework facilitate knowledge discovery and hypothesis assessment. We validated our approach using RMSE for imputation, similarity measures for cohort stratification, and an evaluation with domain experts for usability. The feedback from cancer experts confirms the suitability of our framework for their workflow. There is potential for enhancing our approach through storing and reusing intermediate analysis results to speed up the processing time on the backend. Furthermore, the best analysis options are currently determined for the available prostate cancer data. For using different data sets, an automated identification of the best parameters would be helpful. In future work, we plan to apply our approach to larger and varied data sets to assess its scalability and clinical applicability. To enable experts to test identified correlations, we aim to include export functions for features, genes, patients, and clusters. Another possible extension is providing an option to save, store, and compare analysis sessions. Lastly, we recognize the need to link biological pathway information for a comprehensive cancer mechanism analysis.

Acknowledgment

L. Kenner acknowledges support from MicroONE, funded by the ministries BMK and BMDW, the provinces of Styria and Vienna, and managed by FFG within the COMET program. Financial support was received from the Austrian Federal Ministry of Science, Research and Economy, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, Siemens Healthineers, EU Horizon2020 Doctoral Networks (ALKATRAS, 675712; FANTOM, P101072735 and eRaDicate, 101119427), M. Hehberger Stiftung (15142), CD-Lab for Applied Metabolomics (CDL-AM), Austrian Science Fund (FWF P26011, P29251, P34781, and IPPTO 59.doc.funds), and Vienna Science and Technology Fund (WWTF LS19-018).

References

[AFM*19] ASTOLFI A., FIORE M., MELCHIONDA F., INDIO V., BERTUCCIO S. N., PESSION A.: BCOR involvement in cancer. *Epigenomics 11*, 7 (2019), 835–855. doi:10.2217/epi-2018-0195. 10

[ANI*20] ALEMZADEH S., NIEMANN U., ITTERMANN T., VÖLZKE H., SCHNEIDER D., SPILIOPOULOU M., BÜHLER K., PREIM B.: Visual Analysis of Missing Values in Longitudinal Cohort Study Data. *Computer Graphics Forum (CGF) 39*, 1 (2020), 63–75. doi:10.1111/CGF.13662. 3, 4

[AOH*14] ANGELELLI P., OELTZE S., HAASZ J., TURKAY C., HODNELAND E., LUNDERVOLD A., LUNDERVOLD A. J., PREIM B., HAUSER H.: Interactive Visual Analysis of Heterogeneous Cohort-Study Data. *IEEE Computer Graphics and Applications 34*, 5 (2014), 70–82. doi:10.1109/MCG.2014.40. 2

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

[Ase22] ASESH A.: Normalization and Bias in Time Series Data. In *Digital Interaction and Machine Intelligence* (2022), Springer International Publishing, pp. 88–97. doi:10.1007/978-3-031-11432-8_8. 4

[BBJ*17] BANNACH A., BERNARD J., JUNG F., KOHLHAMMER J., MAY T., SCHECKENBACH K., WESARG S.: Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. In *Workshop on Visual Analytics in Healthcare (VAHC)* (2017), IEEE, pp. 84–91. doi:10.1109/VAHC.2017.8387545. 2

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing Scatterplots. *Technometrics 29*, 2 (1987), 127–142. doi:10.1080/00401706.1987.10488204. 6

[BSM*15] BERNARD J., SESSLER D., MAY T., SCHLOMM T., PEHRKE D., KOHLHAMMER J.: A Visual-Interactive System for Prostate Cancer Cohort Analysis. *IEEE Computer Graphics and Applications 35*, 3 (2015), 44–55. doi:10.1109/MCG.2015.49. 2

[CCW*21] CORVÒ A., CABALLERO H. S. G., WESTENBERG M. A., VAN DRIEL M. A., VAN WIJK J. J.: Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 27*, 10 (2021), 3851–3866. doi:10.1109/TVCG.2020.2990336. 2

[CH74] CALIŃSKI T., HARABASZ J.: A Dendrite Method for Cluster Analysis. *Communications in Statistics – Theory and Methods 3* (1974), 1–27. doi:10.1080/03610927408827101. 5

[Chr12] CHRISTINE M. MICHEEL AND SHARLY J. NASS AND GILBERT S. OMENN: *Evolution of Translational Omics: Lessons Learned and the Path Forward*. The National Academies Press, 2012. doi:10.17226/13297. 2

[CV22] CERDA P., VAROQUAUX G.: Encoding High-Cardinality String Categorical Variables. *IEEE Transactions on Knowledge and Data Engineering 34*, 3 (2022), 1164–1176. doi:10.1109/TKDE.2020.2992529. 4

[CZD19] CHENG Z., ZOU C., DONG J.: Outlier Detection Using Isolation Forest and Local Outlier Factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems* (2019), Association for Computing Machinery, pp. 161–168. doi:10.1145/3338840.3355641. 4

[DB79] DAVIES D. L., BOULDIN D. W.: A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 1*, 2 (1979), 224–227. doi:10.1109/TPAMI.1979.4766909. 5

[DLH11] DAAE LAMPE O., HAUSER H.: Interactive visualization of streaming data with Kernel Density Estimation. In *Pacific Visualization Symposium* (2011), pp. 171–178. doi:10.1109/PACIFICVIS.2011.5742387. 5

[Don06] DONG J.-T.: Prevalent mutations in prostate cancer. *Journal of cellular biochemistry 97*, 3 (2006), 433–447. doi:10.1002/jcb.20696. 2

[EMK*21] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 27*, 3 (2021), 2153–2173. doi:10.1109/TVCG.2019.2944182. 5

[GDKB17] GUTENKO I., DMITRIEV K., KAUFMAN A. E., BARISH M. A.: AnaFe: Visual Analytics of Image-derived Temporal Features – Focusing on the Spleen. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 23*, 1 (2017), 171–180. doi:10.1109/TVCG.2016.2598463. 1, 2

[GGAM12] GSCHWANDTNER T., GAERTNER J., AIGNER W., MIKSCH S.: A Taxonomy of Dirty Time-Oriented Data. *International Cross-Domain Conference and Workshop on Availability, Reliability, and Security (CD-ARES)* (2012), 58–72. doi:10.1007/978-3-642-32498-7_5. 3, 4

[GMS*15] GARRISON L., MÜLLER J., SCHREIBER S., HAUSER S. O. H., BRUCKNER S.: DimLift: Interactive Hierarchical Data Exploration Through Dimensional Bundling. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 27*, 6 (2015), 2908–2922. doi:10.1109/TVCG.2021.3057519. 5

[HNHP07] HERNANDEZ D. J., NIELSEN M. E., HAN M., PARTIN A. W.: Contemporary evaluation of the D'amico risk classification of prostate cancer. *Urology 70*, 5 (2007), 931–935. doi:10.1016/j.urology.2007.08.055. 2

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A Systematic Review on the Practice of Evaluating Visualization. *Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2818–2827. doi:10.1109/TVCG.2013.126. 9

[KBZ*21] KARIM M. R., BEYAN O., ZAPPA A., COSTA I. G., REBHOLZ-SCHUHMANN D., COCHEZ M., DECKER S.: Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics 22*, 1 (2021), 393–415. doi:10.1093/BIB/bbz170. 5

[KCH*03] KIM W., CHOI B.-J., HONG E., KIM S.-K., LEE D.: A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery 7* (2003), 81–99. doi:10.1023/A:1021564703268. 4

[KGB*12] KUMAR V., GU Y., BASU S., BERGLUND A., ESCHRICH S. A., SCHABATH M. B., FORSTER K., AERTS H. J., DEKKER A., FENSTERMACHER D., ET AL.: Radiomics: the process and the challenges. *Magnetic resonance imaging 30*, 9 (2012), 1234–1248. doi:10.1016/j.mri.2012.06.010. 2

[KGD*19] KERZNER E., GOODWIN S., DYKES J., JONES S., MEYER M.: A framework for creative visualization-opportunities workshops. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 748–758. doi:10.1109/TVCG.2018.2865241. 9

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 18*, 9 (2012), 1520–1536. doi:10.1109/TVCG.2011.279. 9

[LL17] LUNDBERG S., LEE S.-I.: A Unified Approach to Interpreting Model Predictions. *Conference on Neural Information Processing Systems (NIPS)* (2017). doi:10.48550/arXiv.1705.07874. 6

[LLD*17] LAMBIN P., LEIJENAAR R. T., DEIST T. M., PEERLINGS J., DE JONG E. E., VAN TIMMEREN J., SANDULEANU S., LARUE R. T., EVEN A. J., JOCHEMS A., VAN WIJK Y., WOODRUFF H., VAN SOEST J., LUSTBERG T., ROELOFS E., VAN ELMPT W., DEKKER A., MOTTAGHY F. M., WILDBERGER J. E., WALSH S.: Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology 14*, 12 (2017), 749–762. doi:10.1038/nrclinonc.2017.141. 5

[LLWF21] LIU H., LI J., WU Y., FU Y.: Clustering With Outlier Removal. *IEEE Transactions on Knowledge and Data Engineering 33*, 6 (2021), 2369–2379. doi:10.1109/TKDE.2019.2954317. 4

[LS09] LOKESHWAR V. B., SELZER M. G.: Hyaluronidase: Both a tumor promoter and suppressor. In *Hyaluronan in Cancer Biology*, Stern R., (Ed.). Academic Press, 2009, pp. 189–206. doi:10.1016/B978-012374178-3.10011-0. 2

[LSKS10] LEX A., STREIT M., KRUIJFF E. P. C., SCHMALSTIEG D.: Caleydo: Design and Evaluation of a Visual Analysis Framework for Gene Expression Data in its Biological Context. In *Proceedings of IEEE Pacific Visualization Symposium* (2010), IEEE, pp. 57–64. doi:10.1109/PACIFICVIS.2010.5429609. 1, 2

[LSS*12] LEX A., STREIT M., SCHULZ H.-J., PARTL C., SCHMALSTIEG D., PARK P., GEHLENBORG N.: StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum (CGF) 31*, 3 (2012), 1175–1184. doi:10.1111/j.1467-8659.2012.03110.x. 1, 2

[LTZ08] LIU F. T., TING K. M., ZHOU Z.-H.: Isolation Forest. In *International Conference on Data Mining* (2008), vol. 8, IEEE, pp. 413–422. doi:10.1109/ICDM.2008.17. 4

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.

[LXNR19] LI R., XING L., NAPEL S., RUBIN D. (Eds.): *Radiomics and Radiogenomics: Technical Basis and Clinical Applications, 1st edition*. Imaging in Medical Diagnosis and Therapy. Chapman & Hall, Chemical Rubber Company (CRC), 2019. doi:10.1201/9781351208277. 2, 5

[MSO*20] MÜLLER J., STOEHR M., OESER A., GAEBEL J., STREIT M., DIETZ A., OELTZE-JAFRA S.: A visual approach to explainable computerized clinical decision support. *Computers & Graphics (CAG) 91* (2020), 1–11. doi:10.1016/j.CAG.2020.06.004. 2

[MWH*20] MÖRTH E., WAGNER-LARSEN K. S., HODNELAND E., KRAKSTAD C., HALDORSEN I. S., BRUCKNER S., SMIT N. N.: RadEx: Integrated Visual Exploration of Multiparametric Studies for Radiomic Tumor Profiling. *Computer Graphics Forum (CGF) 39*, 7 (2020), 611–622. doi:10.1111/CGF.14172. 1, 2

[Ng17] NG S. C.: Principal component analysis to reduce dimension on digital image. In *International Conference on Advances in Information Technology* (2017), vol. 111, pp. 113–119. doi:10.1016/j.procs.2017.06.017. 4

[NHG19] NUSRAT S., HARBIG T., GEHLENBORG N.: Tasks, Techniques, and Tools for Genomic Data Visualization. *Computer Graphics Forum 38*, 3 (2019), 781–805. doi:10.1111/cgf.13727. 2

[NNH*14] NGUYEN Q. V., NELMES G., HUANG M. L., SIMOFF S., CATCHPOOLE D.: Interactive Visualization for Patient-to-Patient Comparison. *Genomics & Informatics 12*, 1 (2014), 21–34. doi:10.5808/GI.2014.12.1.21. 2

[OH21] OSHO O., HONG S.: An Overview: Stochastic Gradient Descent Classifier, Linear Discriminant Analysis, Deep Learning and Naive Bayes Classifier Approaches to Network Intrusion Detection. *International Journal of Engineering and Technical Research 10* (2021), 294–308. 6

[PCKA*17] PEREZ-CORNAGO A., KEY T. J., ALLEN N. E., FENSOM G. K., BRADBURY K. E., MARTIN R. M., TRAVIS R. C.: Prospective investigation of risk factors for prostate cancer in the UK Biobank cohort study. *British journal of cancer 117*, 10 (2017), 1562–1571. doi:10.1038/bjc.2017.312. 2

[RG19] ROS F., GUILLAUME S.: A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications 128* (2019), 96–108. doi:10.1016/j.eswa.2019.03.031. 5

[Rou87] ROUSSEEUW P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20* (1987), 53–65. doi:10.1016/0377-0427(87)90125-7. 5

[RvdHD*15] RAIDOU R. G., VAN DER HEIDE U. A., DINH C. V., GHOBADI G., KALLEHAUGE J. F., BREEUWER M., VILANOVA A.: Visual Analytics for the Exploration of Tumor Tissue Characterization. *Computer Graphics Forum (CGF) 34*, 3 (2015), 11–20. doi:10.1111/CGF.12613. 1, 2

[RZ19] RIDZUAN F., ZAINON W. M. N.: A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science 161* (2019), 731–738. doi:10.1016/j.procs.2019.11.177. 4

[SDE*16] SRIGLEY J. R., DELAHUNT B., EGEVAD L., SAMARATUNGA H., YAXLEY J., EVANS A. J.: One is the new six: The International Society of Urological Pathology (ISUP) patient-focused approach to Gleason grading. *Canadian Urological Association Journal 10* (2016), 339–341. doi:10.5489/cuaj.4146. 3

[SEK03] STEINBACH M., ERTÖZ L., KUMAR V.: The Challenges of Clustering High Dimensional Data. *University of Minnesota Supercomputer Institute Research Report 213* (2003). doi:10.1007/978-3-662-08968-2_16. 5

[Shn94] SHNEIDERMAN B.: Dynamic queries for visual information seeking. *IEEE Software 11*, 6 (1994), 70–77. doi:10.1109/52.329404. 7

[SJG*22] SAXENA S., JENA B., GUPTA N., DAS S., SARMAH D., BHATTACHARYA P., NATH T., PAUL S., FOUDA M. M., KALRA M., SABA L., PAREEK G., SURI J. S.: Role of Artificial Intelligence in Radiogenomics for Cancers in the Era of Precision Medicine. *Cancers 14*, 12 (2022), 2860. doi:10.3390/cancers14122860. 5

[SKT*19] SCHNEIDER L., KEHL T., THEDINGA K., GRAMMES N. L., BACKES C., MOHR C., SCHUBERT B., LENHOF K., GERSTNER N., HARTKOPF A. D., WALLWIENER M., KOHLBACHER O., KELLER A., MEESE E., GRAF N. M., LENHOF H.: ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification. *Bioinformatics 35* (2019), 5171–5181. doi:10.1093/bioinformatics/btz302. 2

[SRY*21] SHUI L., REN H., YANG X., LI J., CHEN Z., CHENG C.-Y., ZHU H., SHUI P.: The Era of Radiogenomics in Precision Medicine: An Emerging Approach to Support Diagnosis, Treatment Decisions, and Prognostication in Oncology. *Frontiers in Oncology 10* (2021). doi:10.3389/FONC.2020.570465. 2, 5

[THFM14] TREBUŇA P., HALČINOVÁ J., FIL'O M., MARKOVIC J.: The importance of normalization and standardization in the process of clustering. In *International Symposium on Applied Machine Intelligence and Informatics (SAMI)* (2014), vol. 12, IEEE, pp. 381–385. doi:10.1109/SAMI.2014.6822444. 4

[Tho53] THORNDIKE R. L.: Who belongs in the family? *Psychometrika 18*, 4 (1953), 267–276. doi:10.1007/BF02289263. 5

[TLS*14] TURKAY C., LEX A., STREIT M., PFISTER H., HAUSER H.: Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. *IEEE Computer Graphics and Applications 34*, 2 (2014), 38–47. doi:10.1109/MCG.2014.1. 2

[van18] VAN BUUREN S.: *Flexible Imputation of Missing Data (2nd Edition)*. Interdisciplinary Statistics. Chapman & Hall, Chemical Rubber Company (CRC), 2018. doi:10.1201/9780429492259. 4

[VDM14] VAN DER MAATEN L.: Accelerating T-SNE Using Tree-Based Algorithms. *Journal of Machine Learning Research 15*, 93 (2014), 3221–3245. 5

[XWY*21] XIANG R., WANG W., YANG L., WANG S., XU C.: A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics 12* (2021). doi:10.3389/fgene.2021.646936. 5

[YJY*17] YU L., JIANG H., YU H., ZHANG C., MCALLISTER J., ZHENG D.: iVAR: Interactive visual analytics of radiomics features from large-scale medical images. In *International Conference on Big Data* (2017), IEEE, pp. 3916–3923. doi:10.1109/BigData.2017.8258398. 2

[ZCP*21] ZANFARDINO M., CASTALDO R., PANE K., AFFINITO O., AIELLO M., SALVATORE M., FRANZESE M.: MuSA: a graphical user interface for multi-OMICs data integration in radiogenomic studies. *Scientific Reports 11*, 1 (2021), 1550. doi:10.1038/s41598-021-81200-z. 3

[ZLVL20] ZWANENBURG A., LEGER S., VALLIÈRES M., LÖCK S.: Image biomarker standardisation initiative. *Radiology 295*, 2 (2020), 328–338. doi:10.1148/RADIOL.2020191145. 2

© 2023 The Authors.
Proceedings published by Eurographics - The European Association for Computer Graphics.