TU WIEN Informatics

Visual Analytics zur korrelativen Erforschung und des Erkenntnisgewinns in der Radiogenomik-Analyse

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Visual Computing

eingereicht von

Sarah El-Sherbiny, BSc
Matrikelnummer 01126592

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr. Renata Georgia Raidou

Wien, 4. Mai 2023

Sarah El-Sherbiny

Renata Georgia Raidou

Informatics

Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Visual Computing

by

Sarah El-Sherbiny, BSc
Registration Number 01126592

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr. Renata Georgia Raidou

Vienna, 4th May, 2023

Sarah El-Sherbiny

Renata Georgia Raidou

Erklärung zur Verfassung der Arbeit

Sarah El-Sherbiny, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 4. Mai 2023

Sarah El-Sherbiny

Danksagung

Dieses interessante und bedeutende Thema im Rahmen meiner Diplomarbeit zu erarbeiten, erfährt meine volle Wertschätzung. Zusätzlich auch von einer äußerst unterstützenden und professionellen Betreuung auf dem gesamten Weg begleitet zu werden, hat diese einzigartige Erfahrung und das Ergebnis nochmals auf ein ganz anderes Level angehoben. Meine tiefste Dankbarkeit geht an meine Betreuerin Renata Raidou für ihr kontinuierliches Feedback, ihre konstruktiven Kommentare, sowie ihre inspirierenden Ideen. Ich schätze all ihre Bemühungen sehr, mir die nötige Orientierung und Freiheit zu geben, um Neues auszuprobieren, während sie mir das bestmögliche und effektivste Arbeitsumfeld in jeder Hinsicht ermöglicht hat. Eine derart hervorragende Betreuung und Erfahrung während der gesamten Diplomarbeit zu erhalten, geht weit über jede Selbstverständlichkeit hinaus und hatte einen erheblichen positiven Einfluss auf mich.

Ich danke unseren kooperierenden Expert_innen Lukas Kenner, Brigitte Hantusch, und Jing Ning von der Medizinischen Universität Wien, für die Bereistellung der Daten die dieses Thema möglich gemacht haben, für ihre regelmäßige Begeisterung für das Thema, sowie ihre hilfreichen Erklärungen über die Daten und ihr ausführliches Feedback zu unseren Ergebnissen. Weiters danke ich allen, mit denen sie uns durch Meetings oder Präsentationen in Kontakt gebracht haben, für ihr Interesse und ihr Feedback. Ich bin auch Clemens Spielvogel für seine frühen Kommentare und Vorschläge zum Machine Learning Teil dieser Arbeit dankbar.

Darüber hinaus möchte ich mich bei allen bedanken mit denen ich studiert oder gearbeitet habe, die mein Studium auf diese wunderbare Weise vervollständigt haben, und selbst die Pandemiezeit in etwas Besonderes verwandelt haben.

Mein herzliches Dankeschön geht an meine Familie, die immer an mich glaubt und eine wahre Freude in meinem Leben darstellt.

Acknowledgements

Having this interesting and meaningful topic for my thesis gets my full appreciation. In addition to the topic, I sincerely appreciate the highly supportive and professional supervision throughout the entire thesis that raised this unique experience and outcome to a different level. My deepest gratitude goes to my supervisor Renata Raidou for her continuous feedback, constructive comments, and inspiring ideas. I highly appreciate all her efforts in giving me the guidance and freedom needed to test new ways while providing me the best and most effective work environment by all means. Such an outstanding supervision and experience throughout all steps of the thesis is way beyond any norm and had a tremendously positive impact on me.

I want to thank our collaborating experts Lukas Kenner, Brigitte Hantusch, and Jing Ning from the Medical University of Vienna for providing us the data that made this topic possible, for their regular excitement on the subject, as well as their helpful explanations on the data and extensive feedback on the outcome. I also want to thank everyone they connected us with through meetings or presentations for their interest and feedback. I am thankful to Clemens Spielvogel for his early comments and suggestions on the machine learning part of the thesis.

Furthermore, I want to thank everyone I studied or worked with who complemented my study in this marvelous way and even changed the pandemic time into something special.

My heartfelt thank you goes to my family, who always believes in me and is an absolute pleasure in my life.

Kurzfassung

Unter Radiogenomik versteht man die kombinierte Erforschung von bildgebenden Merkmalen namens Radiomik und Gen-Sequenzierungsdaten, der sogenannten Genomik. Zu den Herausforderungen der Datenanalyse zählen die Größe, Heterogenität und Komplexität der Datensätze. Diese Herausforderungen machen die Analyse des verfügbaren Informationsraums für Krebsexpert_innen zu einer mühseligen Aufgabe und behindern die Erforschung sowie den Erkenntnissgewinn von Informationen. Dies wird zusätzlich erschwert, wenn klinische Informationen in die Analysen miteinbezogen werden müssen. Visual Analytics (VA) kombiniert automatisierte Analysetechniken, wie maschinelles Lernen oder Statistik, mit interaktiven visuellen Schnittstellen. VA ermöglicht es Einblicke in komplexe Daten zu gewinnen, um effektive Entscheidungen zu treffen. Im Kontext der Radiogenomik-Analyse in Kombination mit klinischen Daten bieten VA-Ansätze vielversprechende Ergebnisse für die Erstellung von Tumorprofilen. Allerdings wurden VA-Ansätze, die radiogenomische und klinische Daten in einem interaktiven, flexiblen, visuellen Tool vereinen, bisher nicht untersucht. In dieser Arbeit ermöglichen wir die integrierte Erforschung und Analyse von Radiogenomik-Daten und klinischen Informationen zur Wissensfindung und Hypothesenbewertung in einer großen Kohorte von Prostatakrebs-Patient_innen. Wir behandeln fehlende Daten durch Imputationstechniken und wenden unüberwachtes maschinelles Lernen für die Dimensionalitätsreduktion und das Clustering der Daten an, um die Datenverarbeitung und Visualisierung zu vereinfachen. Als Ergebnis präsentieren wir ein interaktives visuelles Tool für zwei Zielgruppen: Krebsexpert_innen, sowie biomedizinischen Datenwissenschaftler_innen. Unser Tool ermöglicht es Krebsexpert_innen, Einblicke in die Daten zu gewinnen, indem neue Muster oder Korrelationen in den Datensätzen aufgedeckt werden. Diese können Hypothesen, die ihnen zu den zugrunde liegenden Datensätzen vorschweben, interaktiv bewerten und verfeinern. Für biomedizinische Datenwissenschaftler_innen bietet unser Framework die Möglichkeit, die Analysekomponenten zu verstehen und ihre Auswirkungen auf das Ergebnis interaktiv zu erforschen. Wir bewerten die unbeaufsichtigten maschinellen Lernmodelle anhand von Ähnlichkeitsmaßen wie dem Silhouettenkoeffizienten. Um die Funktionalität des Frameworks zu evaluieren, führen wir Anwendungsszenarien durch, die von Krebsexpert_innen bestätigt werden. Das Feedback unserer Fachpersonen zeigt, dass unser Tool flexibel und geeignet ist, um Einblicke in große heterogene radiogenomische Daten in Kombination mit klinischen Daten zu erlangen. Es fördert den Wissensgewinn und unterstützt bei der Aufstellung, Überprüfung, sowie Verfeinerung von Hypothesen. Unser Tool umfasst die

Integration von interaktiver Visualisierung und automatisierten Analysekomponenten. Es unterstützt die mit uns kooperierenden Fachexpert_innen der Medizinischen Universität Wien dabei, neue Einblicke in ihre Daten zu erhalten und gleichzeitig ihre Hypothesen zu untersuchen.

Abstract

Radiogenomics refers to the combined study of imaging-derived features, called radiomics and gene sequencing data, called genomics. Challenges in the analysis of radiogenomic data include the size, heterogeneity, and complexity of the datasets. 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 needs to be included in the analyses. Visual Analytics (VA) combines automated analysis techniques, such as machine learning or statistics, together with interactive visual interfaces. It allows users to gain insights into complex data and make effective decisions. In the context of radiogenomics analysis with respect to clinical data, VA approaches offer promising directions in tumor profiling. However, VA approaches that bridge radiogenomic and clinical data in an interactive and flexible visual framework have not been investigated before. In this work, we enable the integrated exploration and analysis of radiogenomic data and clinical information for knowledge discovery and hypothesis assessment in a large cohort of prostate cancer patients. We handle missingness in the data through imputation techniques and apply unsupervised machine learning for the dimensionality reduction and clustering of the data to facilitate data handling and visualization. As a result, we present an interactive visual interface for two target audiences: cancer experts and biomedical data scientists. Our framework enables cancer experts to gain insights into the data by revealing new patterns or correlations in the datasets. It allows them to interactively assess and refine any hypothesis in mind for the underlying datasets. For biomedical data scientists, our framework offers the possibility to understand the analysis components and interactively explore their impact on the outcome. We evaluate the unsupervised machine learning models through similarity measures such as the silhouette coefficient. To assess the usability of the framework, we perform usage scenarios that we confirm by our cancer experts. The feedback from our domain experts reveals that our framework is a suitable and flexible technique to gain insights into large and heterogenous radiogenomic data with respect to clinical data. It promotes knowledge discovery as well as hypothesis creation, assessment, and refinement. Interacting with the different visualization and analysis components enhances the understanding of the data and the resulting visual representations. Our approach incorporates the integration of interactive visualization and automated analysis components. It supports our collaborating domain experts at the Medical University of Vienna to obtain new insights into their data, while investigating hypotheses at hand.

Contents

CHAPTER 1

Introduction

Radiogenomics refers to the combined study of imaging-derived features, called *radiomics*, with gene sequencing data, called *genomics* [SRY+21]. *Visual Analytics (VA)* combines automated data analysis approaches with interactive visual interfaces that allow the user to gain insight into complex data and make effective decisions [KMSZ09]. In this work, we apply visual analytics to support the correlative exploration of radiogenomic data in a prostate cancer cohort and the sensemaking process of cancer experts trying to understand these data or biomedical data scientists working on the analysis of these data. Our cancer experts consist of pathologists, biologists, biochemists, or nuclear medicine physicians. Our biomedical data scientists comprise bioinformaticians or data scientists working with these data. We present the motivation and problem statement in Section 1.1, while we summarize the goals of our work in Section 1.2. We provide an overview of the research question and tasks in Section 1.3. Finally, we outline the methodological steps and evaluation of our approach in Section 1.4.

1.1 Motivation and Problem Statement

Radiomic features are extracted from medical imaging data that show tumor characteristics as indicators of metabolic activity or metastasis. Genomic data decode functional information of Deoxyribonucleic Acid (DNA) or Ribonucleic Acid (RNA) sequences. The analysis of radiogenomics with respect to *clinical data*, such as the age or Body Mass Index (BMI) of patients is recently investigated as a potential enabler of prostate cancer risk stratification [SRY+21]. In this process, each patient is assigned a risk status that supports a better understanding of the tumor aggressiveness and is an indication of the treatment process. However, the size, heterogeneity, and complexity of radiogenomic data [SRY+21] make the analysis of the available information space tedious for cancer experts or biomedical data scientists working with these data. These challenges hinder the exploration and sensemaking of patient information. In the context of radiogenomics analysis with regard to clinical data, visual analytics approaches have not been applied before, although they offer promising directions. We provide a visual analytics strategy that supports the correlative exploration and analysis of radiogenomic data in a cohort of 89 prostate cancer patients.

1.2 Goals of this Work

In this thesis, we investigate, design, and implement a visual analytics approach that enables our collaborating domain experts of cancer experts and biomedical data scientists from the Medical University of Vienna to gain insights into radiogenomic data with regard to clinical data. Furthermore, we highlight correlations and patterns in the data that support the stratification of patient cohorts and have an impact on the treatment process. We propose an integrated analysis of radiomic features with mutation data of genomes and clinical data. We follow a user-centered strategy and integrate domain knowledge into a semi-automated analytical approach based on unsupervised machine learning to identify patterns in the complex, heterogeneous data that is provided by our collaborating domain experts. These data consist of 153 radiomic features and 10 307 gene mutations for each of the 89 patients. The clinical data comprises 18 features for 144 patients. We visualize the identified correlations and patterns through an interactive interface that supports the exploratory process in a free and a hypothesis-driven manner.

The most related visual analytics techniques of the state of the art investigate imaging-derived features [RvdHD+15], radiomic [GDKB17, MWH+20], or genomic data [LSKS10, LSS+12] only separately. Gutenko et al. [GDKB17] and Mörth et al. [MWH+20] provide visual analytics approaches for the analysis of radiomic features. Gutenko et al. [GDKB17] analyze the changes in the spleen organ and enable the comparison of features in a linked view. They use clinical data, radiomic features, and surface meshes to visualize the change of the organ over time. Similar to this approach, Mörth et al. [MWH+20] implement multiple linked views and combine radiomic features with clinical cohort data. In contrast to the approach of Gutenko et al. [GDKB17] and Raidou et al. [RvdHD+15], Mörth et al. [MWH+20] identify and visualize relations between the radiomic tumor profile and clinical and histological markers. Furthermore, the approach of Raidou et al. [RvdHD+15] does not use radiomic features but imaging-derived features for tumor tissue characterization from pharmacokinetic modeling. Different to these approaches, Lex et al. [LSS+12, LSKS10] do not analyze radiomic features, but genomic data alone [LSKS10] or with respect to clinical data [LSS+12].

In contrast to these approaches, we support the exploratory analysis of radiomic features and genomic data with respect to clinical data in a unified framework. We expect that this combination leads to new insights into the data for domain experts. To the best of our knowledge, radiomics and genomics were never bridged together in a visual interactive framework for knowledge discovery and hypothesis confirmation before. This poses significant challenges with respect to the size, heterogeneity, and complexity of radiogenomic data [SRY+21]. Other challenges include the missingness of patient values or the mixed data types. The clinical data contains missingness in patient scores that must be adequately handled through data imputations. Besides quantitative numerical values, the clinical data also consist of qualitative ordinal and nominal values that must be encoded into numerical values to allow an automated data processing through unsupervised machine learning approaches. On top of this, domain experts do not work on a single dataset. Each available dataset may contain diverse radiogenomic or clinical information from various tumor localizations, where different clinical hypotheses are of relevance. We strive to propose a unified, generalizable solution that is applicable to varying cancer scenarios and contexts.

1.3 Research Question and Tasks

The main goal of this work corresponds to the design and implementation of a visual analytics framework for the exploration and analysis of radiomic features and genomic data with respect to clinical data for a cohort of prostate cancer patients. This aims to support the sensemaking process of the high-dimensional and heterogeneous radiogenomic and clinical data. Cancer experts and biomedical data scientists want to explore and understand the complex and heterogenuous radiogenomic data with respect to clinical parameters for knowledge discovery and hypothesis confirmation. This leads to our main research question and the tasks stemming from this research question.

Research question:

(R) How can visual analytics support domain experts to gain insight into the large and complex radiogenomic and clinical data of prostate cancer patient cohorts?

Tasks:

(T0) **Preprocessing:** The data needs to be prepared to enhance its quality and facilitate the automated analysis and visualization. This task requires identifying and resolving inconsistencies in the data, handling mixed data types, imputing missingness in data values, detecting outliers, and scaling feature ranges.

(T1) **Cohort stratification:** Domain experts want to get insight into the high-dimensional and complex data of prostate cancer patients. Cohort stratification entails the 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. These dimensions are then visualized on screen to identify patterns in the data.

(T2) **Forward analysis:** The forward analysis supports the discovery of new knowledge from the data. It allows the user to freely explore the data without having a specific hypothesis in mind. The user gets the possibility to interact with the data by selecting subsets of patients on a visualization depending on interesting radiogenomic or clinical profiles. Furthermore, the user is presented characterizing and differentiating features, patient distribution values, and the most frequent genes, which provides insights into the stratified data.

(T3) **Backward analysis:** The backward analysis allows domain experts to confirm or reject a present hypothesis. After formulating a hypothesis, the user can filter the radiomic, genomic, or clinical data to verify the correctness or incorrectness of the hypothesis for the underlying data. This task includes processing any subset of features in the dimensionality reduction and clustering of the data. Furthermore, the user can highlight patient values or specify conditions to highlight, process, and explore the matching subset of patient data.

1.4 Methodological Steps and Evaluation

Our main *contribution* in this work includes the design and development of an interactive, flexible visual analytics approach to bridge three datasets of radiomic, genomic, and clinical features together in a visual interactive interface. This allows domain experts to gain insights into the high-dimensional and heterogenuous datasets. It supports the sensemaking process of users in discovering new knowledge or confirming hypotheses within cohorts of prostate cancer patients.

Answering our research question **(R)** requires the design and development of four **methodological task components:**

1. **Data preprocessing and imputation:** Preparation of the data and substitution of missing values in patient scores to enhance the data quality and enable an automated data processing and analysis.

2. **Dimensionality reduction:** Transformation of the data into a low-dimensional feature space to facilitate the data handling, visualization, and analysis.

3. **Cluster analysis:** Division of the patient data into groups with similar properties by an unsupervised machine learning approach.

4. **Visualization:** Interactive visual exploration of the data to support the discovery of correlations and patterns in multiple linked views.

Specifically, data imputation is required to substitute missing patient scores in the clinical data. Furthermore, we require a dimensionality reduction and clustering step of the data by an unsupervised machine learning approach to identify groups of patients with similar or different characteristics. We determine and visualize the characteristics and differences of the identified groups to support the understanding and sensemaking of the data. To allow the user to interact with the data, for example, by selecting a subset of the data or by filtering the data, we represent the processed data in an interactive visual interface. We assess the clustering results and the usability of the application by evaluating it with cluster metrics and by continuous feedback from domain experts. The data preprocessing and imputation component (1) is applied to Task (T0), while the methodological components (2)–(4) are applied to the tasks (T1)–(T3). While the imputation, dimensionality reduction, and cluster analysis are unified for all tasks, the visualization components adapt to the specific problem domain and subtask.

Evaluation of our visual analytics framework: We evaluate the unsupervised machine learning model by the Silhouette Coefficient [Rou87] for the analysis of the cluster definition. Furthermore, we use the Calinski-Harabasz index [CH74], and the Davies-Bouldin index [DB79] that indicate the separation of clusters. The higher the Silhouette Coefficient and the Calinski-Harabasz index is, the better are the clusters defined. In contrary, a lower Davies-Bouldin index is related to a better separation of clusters. To assess the usability of the application and its efficiency for clinical research, we perform a qualitative evaluation with domain experts [Mun09, LBI+12].

CHAPTER 2

Clinical Background

In this chapter, we present the clinical background of the thesis. As we work with the data of prostate cancer patients, we give a short introduction into prostate cancer, and its causes in Section 2.1. The diagnosis and treatment process of prostate cancer leads to the three datasets of radiomics, genomics, and clinical data. We present the radiomic data in Section 2.2, the genomic data in Section 2.3, and the clinical data in Section 2.4. While radiomic data is acquired from imaging scans, genomic data is obtained from prostate tissue, and clinical data is retrieved from patient screening. The analysis of these complex and heterogenuous datasets enables cancer experts to gain insights into the data to support the decision and treatment process.

2.1 Prostate Cancer

Cancer is a malignancy that leads to genetic abnormalities [SRY+21, SRM+22]. Body cells affected by cancer grow uncontrolled and spread over the body [NCI21, Bro08], as shown in Figure 2.1. As a result, patients suffer from a decreased quality of life [Su10]. The mortality rate of cancer is low if medical experts detect it in patients at an early stage and treat it curatively [Su10, SRM+22].

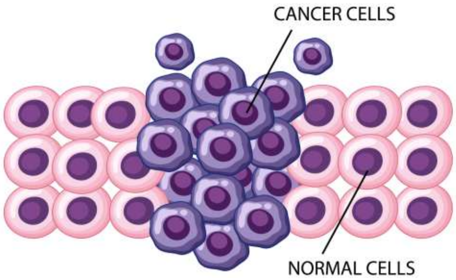


Figure 2.1: Malignant cancer cells compared to normal cells [brg22].

Prostate cancer affects the prostate organ in the reproductive system of the biological male human body [NCI21]. It is after breast cancer the second most frequent cancer type worldwide [WHO22]. Figure 2.2 from the World Health Organization (WHO) visualizes the age-standardized incidence rates of cancer types worldwide for all ages and genders in 2020. Due to the aging population and economic growth, the incident rate is supposed to increase further in the upcoming years [WHO22, WLH+22]. Moreover, prostate cancer is the fifth leading death cause among biological males worldwide [WLH+22, SRM+22].

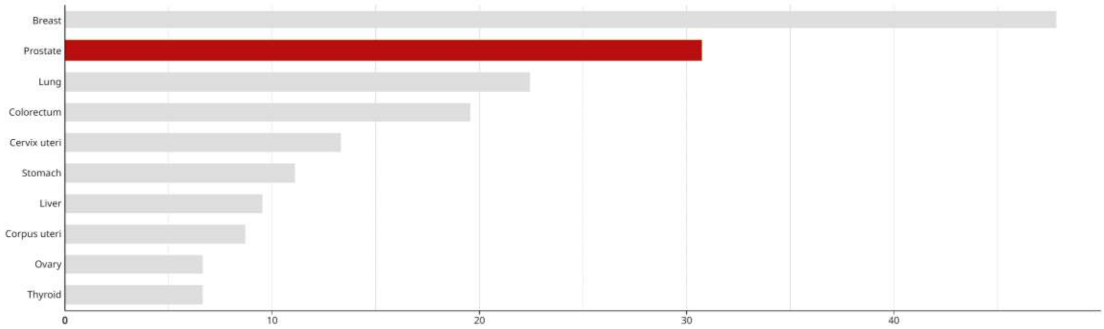


Figure 2.2: Incidence rates of cancer types world wide in 2020 [WHO22].

As the body has a lower ability at an increased age to eliminate cells with damaged DNA before cancer is developed, prostate cancer often occurs when the biological male population gets older [NCI21]. The highest incident rates are reached at an age starting from 65 years [Raw19], while the disease mainly affects biological males starting from the age of 45 years [SRM+22]. Besides the age, the risk factors of prostate cancer comprise the family history, ethnicity, obesity, and environmental factors [SRM+22, NCI21]. The family history includes inherited gene mutations as a frequent cause of cancer [SRM+22].

Prostate cancer is a heterogeneous disease concerning the epidemiology and genetics [SRM+22]. It is often asymptomatic in an early stage, which raises the importance of an early screening [Su10, NCI21]. The diagnosis of the disease includes imaging scans from which radiomic data is retrieved. Additionally, DNA or RNA sequencing leads to genomic data, while health screening determines the clinical scores [SRM+22, NCI21, LXNR19]. Radiomic data indicate, for example, the shape characteristics of the tumor, while genomic data represent changes in the DNA or RNA sequence of genes. Clinical scores embrace, for example, the Prostate Specific Antigen (PSA) that is measured in the blood and could indicate cancer.

The analysis of radiomics, genomics, and clinical data supports the diagnosis and treatment process of prostate cancer and improves clinical decision-making [SRY+21, LXNR19]. It shows potential for precision medicine that aims to enable customized approaches in patient care by considering individual patient characteristics [LXNR19, SRY+21].

2.2 Radiomic Data

The analysis of radiomic data supports the evaluation of disease characteristics and helps clinical experts to understand biological processes [LXNR19]. As a result, the discovery of characteristics for diagnostic or predictive values of diseases is encouraged [LXNR19, SRY+21]. These insights into the data lead to the potential to improve clinical decision support and aid the diagnosis and treatment of diseases, including cancer [LXNR19].

As part of the diagnosis and treatment process of diseases, including prostate cancer, medical imaging is applied [SRM+22, NCI21, LXNR19]. Imaging techniques are divided into anatomical methods and functional methods [MV98]. While anatomical imaging measures structural information, functional imaging also captures temporal information. Anatomical methods involve Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). Functional imaging methods include Positron Emission Tomography (PET). MRI generates an image by sending radio waves through the body. This imaging technique has a high spatial resolution and no limit on tissue penetration. A further advantage is that it does not make use of harmful radiation. Its downside is the high magnetic force, besides the low sensitivity and contrast [LY15]. CT is a nuclear imaging method that uses X-rays to creates cross-sectional images of the body. It has a high spatial resolution and depth penetration. Its disadvantages reveal in the radiation risk and low contrast [LY15]. PET uses radioactive tracers, which are injected into the body and absorbed by organs and tissues. It captures functional information and has a high sensitivity and penetration depth [LY15]. Multimodal imaging such as PET/MRI acquire and combine functional PET information with structural MRI information simultaneously to gain the combined advantages of both modalities [PWKJ08]. Figure 2.3 shows a PET/MRI scan of the prostate with a tumor roughly located within the white drawn area. From these imaging scans, radiomic features are retrieved for further analysis.

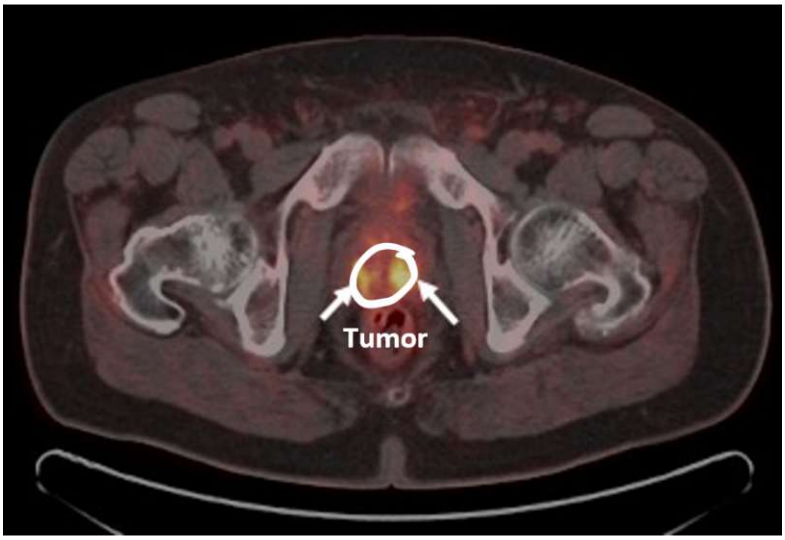


Figure 2.3: PET/MRI scan of a prostate tumor within the drawn area [DoNM22].

Radiomic data represents quantitative features extracted from medical imaging [ZLVL20, LXNR19]. These features characterize the content of a Region of Interest (ROI) such as a tumor region on the scan non-invasively [ZLVL20, ABCA22]. They describe the characteristics of the tumor, including its size, shape, volume, or texture on the image.

Figure 2.4 shows the radiomic pipeline to acquire these features. First, the raw imaging data is converted by using the Standardised Uptake Values (SUV) [HDH+18, ROC+17] or the Target to Blood Pool Ratio (TBR) [CD15]. The TBR is derived from the SUV divided by a constant that represents the venous blood pool to correct the values for the blood uptake [CD15]. Then, the data is processed and segmented to determine the ROI. The ROI is interpolated to the same grid as the image and then split into an intensity and morphological mask. Finally, radiomic features are calculated from the whole image, the extracted ROI, or the ROI discretized by intensity values [ZLVL20].

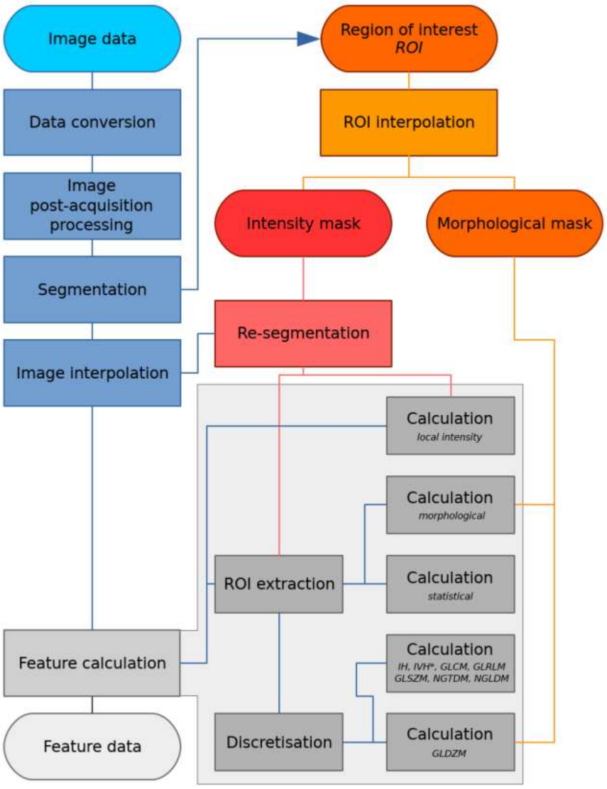


Figure 2.4: Radiomic pipeline to calculate quantitative features of image data [ZLVL20].

To support the reproducibility and validation of radiomic features, the Image Biomarker Standardisation Initiative (IBSI) standardized the nomenclature and definitions of radiomics [ZLVL20]. The standardized feature families by IBSI are shown in Table 2.1. While the statistical (STAT) feature family describes first-order statistics, these families also include features of higher-order statistics to characterize the shape or texture of the tumor on the image [SGB+21, VVTT20, MML+20].

Table 2.1: Radiomic feature families standardized by IBSI [ZLVL20].

Abbreviation	**Full name**
STAT	Intensity-based statistical features
LOC	Local intensity
IH	Intensity histogram
IVH	Intensity-volume histogram
GLCM	Grey level co-occurrence matrix
GLDZM	Grey level distance zone matrix
GLRLM	Grey level run length matrix
GLSZM	Grey level size zone matrix
NGLDM	Neighbouring grey level dependence matrix
NGTDM	Neighbourhood grey tone difference matrix
MORPH	Morphological features

First-order features are based on single pixel or voxel analysis [MML+20]. In contrast, higher-order features capture spatial relationships of pixel or voxel pairs in different directions [MML+20, SGB+21, ABCA22]. Figure 2.5 shows four examples of higher-order features that capture spatial relations of pixels.

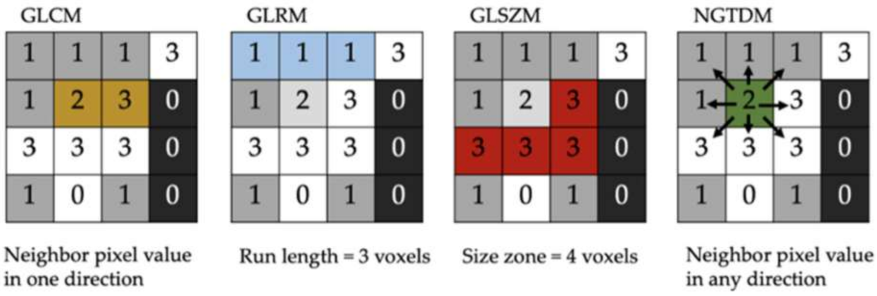

GLCM
GLRM
GLSZM
NGTDM
Neighbor pixel value in one direction
Run length = 3 voxels
Size zone = 4 voxels
Neighbor pixel value in any direction

Figure 2.5: Radiomic feature families that capture spatial relations [SGB+21].

Intensity-based statistics (STAT) features describe the intensity distributions within a ROI [ZLVL20]. Examples of this feature family include, the mean, variance, or skewness of the intensity distribution. These features indicate whether the intensity distribution within a ROI is continuous [ZLVL20]. They are calculated based on the pixel intensities and do not consider the relationship between neighborhood pixels [ABCA22].

Grey level co-occurrence matrix (GLCM) features represent statistical information about the distribution of pixel pairs and operate on discretized intensities [SGB+21, ZLVL20]. They allow the assessment of the surface texture in images [ZLVL20].

Grey level distance zone matrix (GLDZM) features count the number of linked voxels with a common discretised grey level value that have the same distance to the ROI edge [ZLVL20]. They represent the relation between the voxel location and its discretized intensities [ZLVL20].

Grey level run length matrix (GLRLM) features represent the length of consecutive voxels with the same intensity in one direction in the image [SGB+21]. This is called the run length and represents the distribution of discretised intensities [ZLVL20].

Grey level size zone matrix (GLSZM) features quantify connected voxels with the same gray level intensity [SGB+21, ZLVL20]. These features represent the homogeneity of the image ROI [SGB+21].

Neighbouring grey level dependence matrix (NGLDM) features consider the number of connected voxels that depend on the center voxel [SGB+21]. These features represent the coarseness of the texture and are rotationally invariant [ZLVL20].

Neighbourhood grey tone difference matrix (NGTDM) features describe the difference between a gray value from the average value of its neighbors [SGB+21]. It operates on the discretized intensities [ZLVL20].

Local intensity (LOC) features represent local intensity features within a neighborhood around a center voxel. The center voxel has to be inside the ROI, while the neighborhood considered also comprises pixels outside the ROI [ZLVL20]. This family includes two features for the local and global intensity peaks. While the local peak is calculated for the voxel with maximum intensity inside the ROI, the global intensity peak is calculated for all ROI pixels and represents the highest peak value. These features aim to reduce the variance of SUV values [ZLVL20].

Intensity histogram (IH) features are calculated from the discretized intensity distribution into bins. Besides the mean, max, or the skewness of the discretized intensity bins, they also include the maximum histogram gradient intensity [ZLVL20].

Intensity-volume histogram (IVH) features of voxel intensities in the ROI describe the relation between the discretized intensity and the volume fraction that contains at least the discretized intensity [ZLVL20]. It is a measure to analyze the heterogeneity of the image [NGA+09].

Morphological features (MORPH) features represent the geometry of the ROI [ZLVL20]. They include, for example, the tumor area, volume, diameter, or sphericity and describe the shape characteristics of a tumor that is an indication of its malignancy [VVTT20].

Analyzing these quantitative radiomic features can lead to new insights into the data in a non-invasive way [AVL+14]. These imaging features can reveal prognostic information in diseases that have an association with genomic and clinical data [AVL+14].

2.3 Genomic Data

Cancer changes the genome of cells, which leads to mutations in the DNA or RNA [Bro08]. If these cell mutations affect the cell growth and occur at a high rate, they can be dangerous [Bro08]. Genomics deals with the behavior of a set of genes in the genome to analyze and understand the molecules of the biological system [Chr12, oPC23]. It is related to genetics that considers individual genes and their inheritance throughout the generations, but deals with a complete set of genes in the cell or an organism [oPC23, Chr12]. Genomic testing aims to provide information on the behavior of cancer that affect the treatment process. In case of prostate cancer, it is performed on a sample of prostate tissue gained from biopsy or on the tissue of the whole prostate, when the prostate is extracted from the patient after surgery [oPC23]. As a result, the genomic data is retrieved through DNA or RNA sequencing for further analysis.

Figure 2.6 shows the DNA and RNA in comparison, which represent molecules in cell biology responsible for genetic information [Mac22]. The DNA is made of two strands in the shape of a double helix. In contrary, the RNA consists of only one strand. These strands are made of subunits called nucleotids [Mac22]. The DNA consists of the four bases Adenine (A), Cytosine (C), Guanine (G), and Thymine (T), while the RNA entails the base Uracil (U) instead of Thymine (T). Apart from the bases, the DNA and RNA consist of sugar and phosphates [Mac22]. The correct ordering and pairing of the molecule bases is essential for their biological function [Mac22]. DNA or RNA sequencing determines the order of the nucleotides or bases of the molecule [Ada23, NHG19]. From these sequences, the genomic data is retrieved, which represents the human genome in a readable form [Ada23]. A challenge is understanding the meaning of this information for the human health [Ada23]. This makes further analysis of the data necessary.

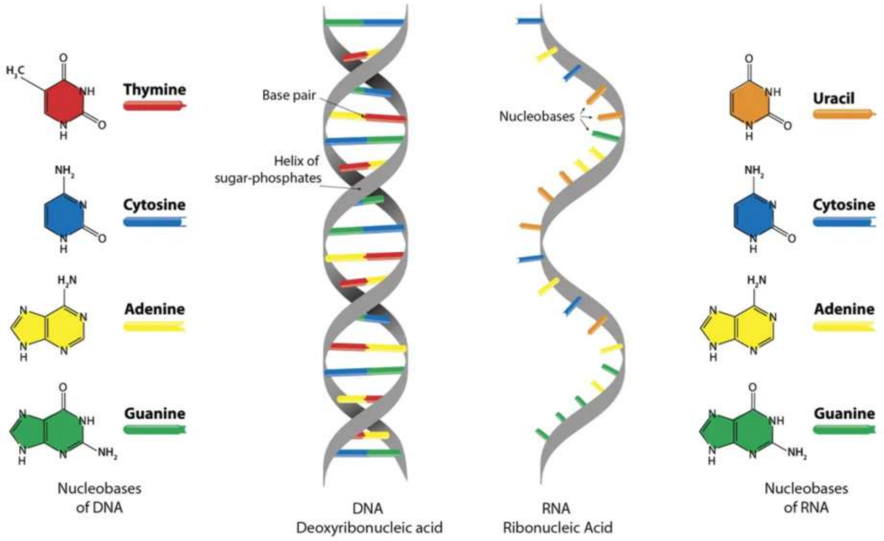


Figure 2.6: DNA and RNA molecules and their bases [Mac22].

2.4 Clinical Data

The collection and analysis of clinical data has an impact on the decision-making process in the healthcare domain [Gro20]. Clinical data includes demographic data such as the Age, Weight, or BMI of patients. In addition, it consists of scores for patient management such as the Prostate-Specific Antigen (PSA). The PSA represents a substance of the prostate that is measured in the blood [fDCP22]. Patients with higher values have a higher probability of having a problem in the prostate [fDCP22]. However, the PSA value is also affected by medications or infections without having prostate problems [fDCP22]. This makes the investigation of other diagnosis methods necessary.

A further indication of prostate disease is the Gleason score (GS). It is determined through tissue that is extracted from the prostate and investigated under the microscope [fDCP22, NCI21]. This score describes the derivation of the extracted cells from normal cells and represents how likely a present tumor will spread [fDCP22, NCI21]. Figure 2.7 shows a schematic representation of the Gleason grades. The higher the values are, the more the analyzed tissue differs from normal tissue. The GS score is grouped further through the International Society of Urological Pathology (ISUP) grade into five categories based on the Gleason patterns [SDE+16].

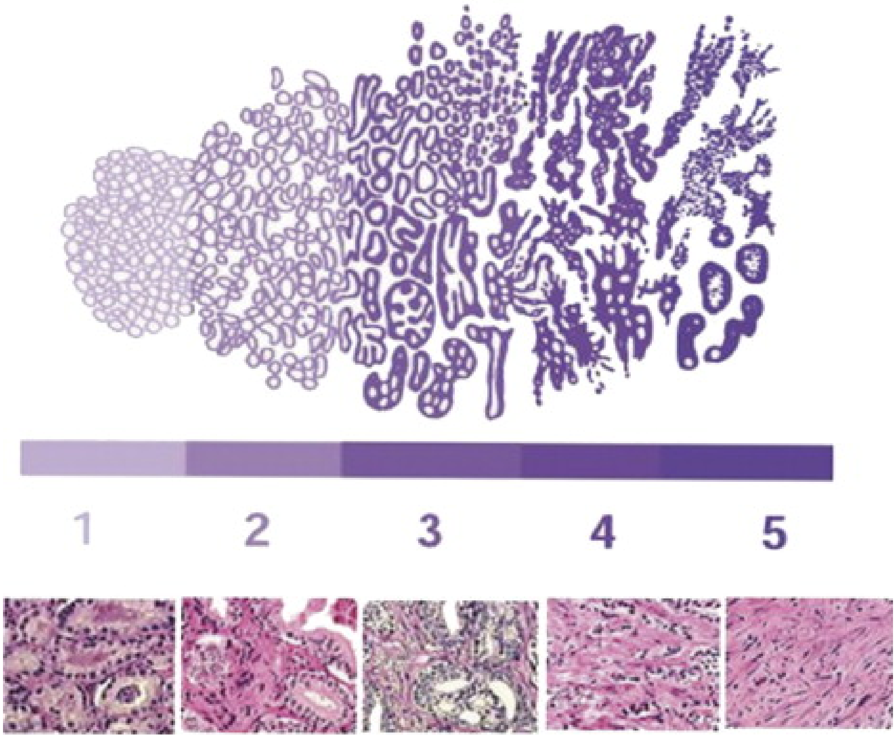


Figure 2.7: Schematic representation of the gleason grades [HSC+07].

Clinical data also comprise prognosis data such as the Biochemical Recurrence (BCR) that might be an indication of the disease progression [LS09]. BCR is defined as the increase of the PSA levels [LS09] and is followed by an operation called radical prostatectomy [LS09]. In this operation, the prostate is removed from the patient as part of the treatment process [Med23].

Tumor staging is the process of determining whether cancer cells have developed or spread within the prostate or to other parts in the human body [fDCP22, oSA23]. The staging process also includes determining the location of the primary tumor, the size, and extent of tumors, or the involvement of lymph nodes in the disease [oSA23]. Clinical staging determines the amount of cancer cells available in the body. It is performed through physical examination, imaging tests, or biopsies [oSA23]. Pathological staging is retrieved when a patient has an operation for tumor removal as it combines clinical staging with surgical results [oSA23]. The clinical staging values, the GS, and the PSA values are combined to determine the D'Amico Risk Statification score [DWM+98, HNHP07]. This score defines three risk groups based on the assessment of 5 years of treatment failure in prostate cancer [DWM+98, HNHP07]. However, it represents a basic stratification scheme with limited clinical relevance for patients with multiple risk factors [HNHP07, Sch22].

2.5 Radiogenomic Data

Each of the three datasets of radiomic, genomic, or clinical data includes indications of prostate cancer but is not expressive alone. Therefore, a combined analysis of these three datasets opens the potential to support clinical experts in understanding the 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]. An essential application is precision medicine, where radiogenomics supports the therapy process and the evaluation of clinical outcomes [SRY+21]. Radiogenomics analysis has the advantages of being cost-effective and repeatable [SRY+21]. It further allows the detection of continuous changes and can be applied as a replacement for invasive interventions [SRY+21].

CHAPTER 3

State of the Art

Visual Analytics (VA) combines automated analysis techniques, such as machine learning or statistics, together with interactive visual interfaces. It allows the user to gain insight into complex data and make effective decisions [KMSZ09]. In this work, we introduce a VA application that enables the integrated exploration and analysis of radiogenomic data together with clinical information in one framework. None of the existing approaches bridges radiomics and genomics together in a visual interface for flexible knowledge discovery and interactive hypothesis confirmation. However, VA approaches exist that analyze radiomics or genomics alone or with respect to clinical data. In this chapter, we summarize the state of the art of related VA approaches for radiomic, genomic, and clinical data analysis. We group these approaches based on their main focus in employing radiomics in Section 3.1, genomics in Section 3.2, or clinical data in Section 3.3.

The combined analysis of radiogenomic data with respect to clinical data opens the potential to understand complex and heterogeneous datasets to support data sensemaking. Therefore, we summarize analysis approaches for radiogenomic data in Section 3.4. Although studies exist that conduct a combined analysis of the three datasets, they lack interactivity and flexibility that allows the user to explore the data to gain insights and knowledge from it. This represents the main difference to our visual interactive interface.

3.1 Visual Analytics for Radiomics

Radiomic data analysis encourages the discovery of characteristics for diagnostic or predictive disease values [LXNR19, SRY+21]. It has potential to improve clinical decision support and to aid the diagnosis and treatment of cancer diseases [LXNR19]. Existing VA approaches analyze radiomic data either separately or in relation to clinical data. Mörth et al. [MWH+20] present an application named *RadEx* to identify and visualize relations between the radiomic tumor profile and clinical and histological markers. RadEx supports the generation of a hypothesis by interactively exploring the data in multiple linked views.

Raidou et al. [RvdHD+15] present a technique for tumor tissue characterization that also supports the generation and confirmation of hypotheses. In contrast to the approach of Mörth et al. [MWH+20], Raidou et al. [RvdHD+15] explore and analyze the features space of imaging-derived data of tumor tissue characteristics. Similarly, Corvo et al. [CCW+21] present the tool *IIComPath* to employ imaging-derived features instead of radiomics and support the hypothesis generation through an interactive selection and analysis of patient groups. Moreover, they provide a provenance mechanism to ensure the reproducibility of defined subcohorts. Yu et al. [YJY+17] adopt and extend the radiomic features defined by Aerts et al. [AVL+14] with first order statistics, shape, size, and texture features to describe tumor phenotype characteristics. Similarly, Tautz et al. [TZH+20] extend radiomics to capture morphological and dynamic heart characteristics. Yu et al. [YJY+17] present an application named *iVAR* to explore the radiomics feature space and analyze relations between the features. Contrary to these approaches, Gutenko et al. [GDKB17] use radiomic features to support the alignment of temporal organ data. They present an application named *AnaFe* that allows the observation of trends in the radiomic data over time to identify predictors for the organ change. Bannach et al. 2017 [BBJ+17] present the tool *VA4Radiomics* to analyze the radiomic data of patient cohorts. They allow a refinement of the cohorts through data filtering.

These presented radiomics approaches utilize multiple linked views to analyze and visualize the radiomic data in an interactive visual interface. They aim to support the exploratory process to gain insights into the data or identify potentially interesting features. An overview of the disease, datasets, and the underlying imaging modalities utilized in the presented approaches is given in Table 3.1. Tautz et al. [TZH+20] analyze radiomic data only separately, while the other approaches link the imaging-derived or radiomic features together with clinical data. Gutenko et al. [GDKB17] and Raidou et al. [RvdHD+15] test the applicability of their approach to prostate cancer that is investigated through MRI data, while the other approaches are validated only on different cancer or disease types. For extracting features from imaging scans, MRI and CT are mainly used. In general, none of these approaches exploits genomic datasets.

3.1.1 Imaging-Derived Datasets and Radiomics

Radiomic features are a set of quantitative image features [BBJ+17]. These features are extracted from imaging data, including multimodality imaging such as PET/CT or PET/MRI. The analysis of these features offers a promising role in diagnosing diseases and predicting the patient outcome [CLTC+21]. In this section, we summarize the different imaging modalities from which radiomic features or imaging-derived features are extracted in related work. Table 3.1 shows an overview of the datasets and imaging modalities employed.

Raidou et al. [RvdHD+15] present a visual analytics approach for tumor tissue characterization. Instead of radiomic features, they use imaging-derived features from Dynamic Contrast Enhanced (DCE) and Diffusion-Weighted (DW) MRI images. Their approach can also be generalized to CT and PET images. They extract per-voxel features by phar-

Table 3.1: Disease, data, and modality used in related radiomics approaches.

Paper	Disease	Data	Modality
Raidou et al. [RvdHD+15]	Prostate and cervical tumor	IDF + C	MRI
Bannach et al. [BBJ+17]	Head and neck cancer	R + C	CT
Gutenko et al. [GDKB17]	Spleen, prostate cancer	R + C	CT, MRI
Yu et al. [YJY+17]	Lung cancer	R + C	MRI
Mörth et al. [MWH+20]	Gynecological cancer	R + C	MRI
Tautz et al. [TZH+20]	Cardiac diseases	R	CT
Corvo et al. [CCW+21]	Breast cancer	IDF + C	WSI

Data: Imaging-derived features (IDF), radiomics (R), clinical data (C)
Modality: Magnetic resonance imaging (MRI), computed tomography (CT), whole slide image (WSI)

macokinetic models that describe tumor tissue characteristics. Corvo et al. [CCW+21] also use imaging-derived data that they extract from a vast number of tissue samples called Whole-Slide Images (WSI). They mention that their methodology can be applied on radiomic features. Gutenko et al. [GDKB17] use radiomic features extracted from CT or MRI images. They group these features into four categories. These categories describe measurements of the organ, shape descriptors, intensity represented by density values, and texture features that describe cluster prominence. Bannach et al. [BBJ+17] use CT based radiomic features and group them into five classes. These classes consist of statistical, geometry, texture based, Laplacian of Gaussian (LoG), and wavelet radiomic features. Statistical features include the mean, median, or standard deviation of tumor intensity values. Geometric features characterize the tumor shape, such as the surface area or volume of the tumor. Texture features are related to statistic features, but analyze the tumor on a voxel level by considering only the nearest neighbors. LoG features focus on areas with significant intensity changes by applying the laplace operator on an image. Wavelet features decompose the image into low and high frequencies using a low- or high-pass filter in x- and y-direction. Similar to Bannach et al., Yu et al. [YJY+17] analyze CT based radiomic features, while Tautz et al. [TZH+20] use radiomic features from Cardiac Magnetic Resonance (CMR) imaging to explore diseases affecting the heart or blood vessels. Mörth et al. [MWH+20] extract radiomic features from multiparametric MRI images to identify tumor characteristics responsible for a possible outcome. According to Cutaia et al. [CLTC+21], the future direction in extracting radiomic data moves towards multimodality imaging. This leads to an integration of image information of different scales of the anatomical and molecular level that helps to overcome the limitations of single techniques in identifying tumor properties [CLTC+21].

3.1.2 Radiomic Data Analysis

For the analysis of radiomics, dimensionality reduction methods are applied to reduce the high-dimensional data to a low-dimensional space. This reduces the complexity of the data, mitigates the curse of dimensionality, and facilitates data visualization [SRY+21, LXNR19]. Clustering methods divide the data into groups with similar intra-characteristics and different inter-characteristics to highlight the underlying patterns in the data [ESA+21, SGTB13]. Statistics and machine learning techniques are utilized to compare subclusters of the data and identify their similarities or differences.

Dimensionality reduction Raidou et al. [RvdHD+15] reduce the dimensionality of the radiomic data by a 2D t-Distributed Stochastic Neighbor Embedding (t-SNE) to preserve local structure in the features space and analyze intrinsic feature characteristics. In contrast, Mörth et al. [MWH+20] apply 1D t-SNE to keep one axis for a clinically meaningful feature selection. They also test Principal Component Analysis (PCA), but t-SNE leads to more suitable results for their scenario and data. They state that this choice may vary dependent on the problem domain. Corvo et al. [CCW+21] rely on Principal Component Analysis (PCA) as a simple method to reduce the feature space. Differently, Bannach et al. [BBJ+17], Gutenko et al. [GDKB17], and Tautz et al. [TZH+20] select a series of carefully abstracted attributes or robust radiomic features that are of interest for the physicians or clinicians. According to Gutenko et al. [GDKB17] this avoids preferring a particular feature type in the data analysis. Yu et al. [YJY+17] enable an interactive selection of the feature dimensions of interest to ensure the interpretability of these features for domain experts without requiring statistics or machine learning knowledge. However, considering only selected features does not ensure that these features include the most indicative ones in the data.

Clustering For data clustering, Corvo et al. [CCW+21] provide k-means and hierarchical clustering as basic clustering methods but claim that considering different methods could overcome limitations with the assumption of data distribution. Yu et al. [YJY+17] apply hierarchical clustering on heatmap values to identify outliers and features with similar radiomic data that might be of interest. Gutenko et al. [GDKB17], Tautz et al. [TZH+20], and Bannach et al. [BBJ+17] do not apply automated clustering techniques on the data. Tautz et al. [TZH+20] mention that clustering correlated features or a filtering functionality could be helpful in the exploration process. Clustering could support the user in choosing features to display, while a filtering functionality could reveal remarkable data ranges to highlight. Raidou et al. [RvdHD+15] and Mörth et al. [MWH+20] achieve visual clusters through the t-SNE dimensionality reduction.

Cluster comparison Raidou et al. [RvdHD+15] compare two selected clusters through Linear Discriminant Analysis (LDA) to show features that differentiate between them. This leads to a vector that maximizes the linear separation of cluster means, while minimizing the variance within clusters [RvdHD+15]. Gutenko et al. [GDKB17] compute the similarity of temporal sequences by the cosine similarity of feature vectors in combination with the distance measure Dynamic Time Warping (DTW) that is based on the cosine similarity. Bannach et al. [BBJ+17] compare distributions of a selected cohort with the patient population through the Goodness-of-Fit measure to determine relations between the two groups.

3.1.3 Visualization of Radiomic Data

The results of the data analysis are visualized in an interactive VA framework through multiple linked views. These views mainly include scatterplots, heatmaps, parallel coordinate charts, or bar plots to highlight patterns or relations in the data. Interaction with the views is reached through data filtering, selections on the visualizations, hovering, clicking, or zooming in visualizations, updating sliders, or linking and brushing.

Mörth et al. [MWH+20], Raidou et al. [RvdHD+15], and Gutenko et al. [GDKB17] visualize radiomic or imaging-derived data through a scatterplot. Mörth et al. [MWH+20] allow the selection of patients on the scatterplot through unit charts that depict clinical features. They sort the unit charts based on the influence of the clinical features on the outcome. Raidou et al. [RvdHD+15] use a density plot to select points on the scatterplot that are not well defined. They indicate density regions through a heated body colormap. Gutenko et al. [GDKB17] use a scatterplot to visualize measurement metrics over time points, while they represent feature vectors for each subject through heatmaps. Figure 3.1 shows the similarity comparison of Gutenko et al. [GDKB17] based on all radiomic features. They visualize the data that is similar to a selected subject through scatterplots of the measurement progression over time, temporal 3D organ data, and heatmaps.

Yu et al. [YJY+17] provide a heatmap and correlation matrix for all features to interactively select the feature dimensions of interest. Parallel coordinate plots are used by most of the approaches. Corvo et al. [CCW+21] integrate a parallel coordinate plot for an overview of the data distributions. Tautz et al. [TZH+20] visualize seven pre-selected radiomic features in a parallel coordinate plot. Mörth et al. [MWH+20] use parallel coordinate plots to show correlations between clinical features. Raidou et al. [RvdHD+15] employ a parallel coordinate plot to highlight relations beyond two dimensions.

Gutenko et al. [GDKB17] visualize clinical data through bar charts, while Corvo et al. [CCW+21] employ bar charts to display distributions of values. Bannach et al. [BBJ+17] rely on multiple linked bar charts to visualize single patients, patient cohorts, and the distribution of a selected patient subset of a cohort. Raidou et al. [RvdHD+15] use stacked bar charts to visualize feature combinations that contribute the most to the cluster separation, in addition to the separate and joint cluster distributions. The cluster analysis and comparison view of Raidou et al. [RvdHD+15] is shown in Figure 3.2. They abstract each cluster into a sphere. Then, they visualize the cluster cohesion by the area and opacity of a sphere, the cluster separation by a glyph between two spheres, and the average silhouette coefficient that combines both scores through a luminance color scale.

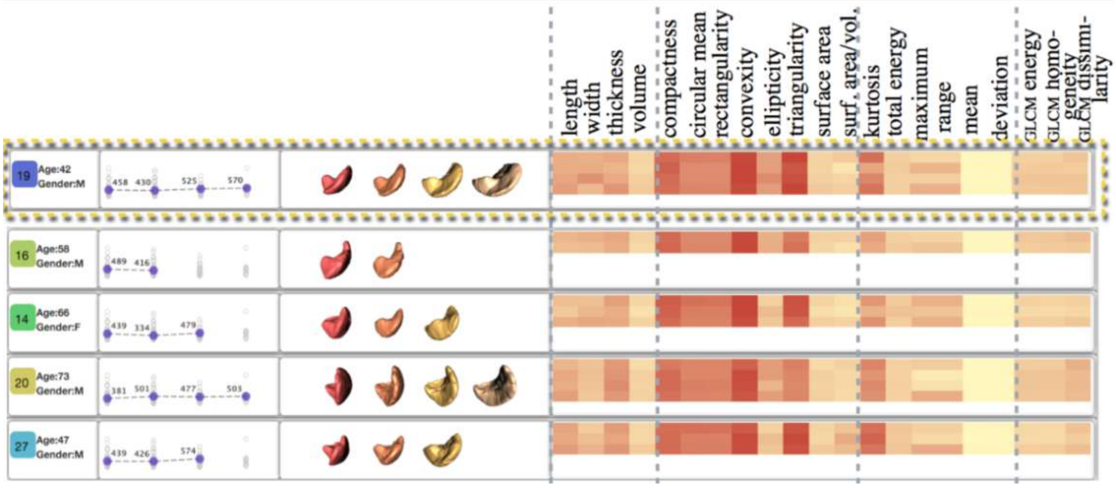


Figure 3.1: Similarity comparison of a selected subject based on all features [GDKB17].

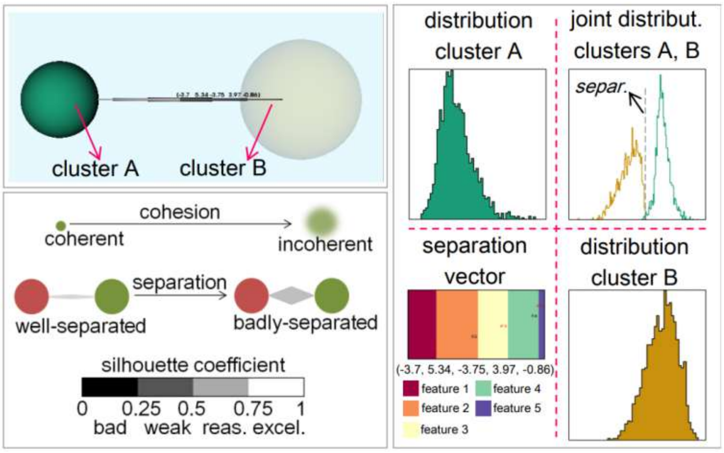


Figure 3.2: Cluster analysis and comparison [RvdHD+15].

3.2 Visual Analytics for Genomics

Sequencing technologies lead to extensive amounts of genomic data. Analyzing and visualizing these data aims to acquire a clear and understandable data representation for clinicians and medical experts interacting with these data [QLN+19]. Getting insight into genomic data and uncovering hidden patterns in these data supports clinical experts in interpreting the data, formulating a hypothesis, and determining the treatment needs of patients [QLN+19, NHG19]. Visualizing genomic data faces challenges due to the long sequences, sparse distributions, interaction between distant sequences, and diverse data types [NHG19]. These challenges must be addressed in the design of a visual interface and make the data analysis through algorithmic tools necessary [NHG19].

Schneider et al. [SKT+19] present an application called *ClinOmicsTrail* for breast cancer decision support to identify clinical markers by combining genomics with clinical data. In contrast to this approach, Lex et al. [LSKS10] present the tool *Caleydo* to combine gene expression data with models of biological processes called *pathways*. They focus on the relationship between gene expression data with pathways and between multiple pathways to interpret individual effects and identify different subtypes of diseases. When significant differences exist between cancers from the same tissue, these differences are characterized by biomolecular properties and are called *cancer subtypes* [TLS+14]. Lex et al. [LSS+12] analyze genomic datasets to identify subtypes in combination with clinical data in their application called *StratomeX*. Turkay et al. [TLS+14] present the combined application *Caleydo StratomeX* to analyze and discover relations between genomic and clinical data. In contrast to the approaches of Lex et al. [LSKS10, LSS+12], *Caleydo StratomeX* focuses on breast cancer patients instead of combining different cancer datasets. However, it solves the limitation of *StratomeX* in identifying characteristic genes of cancer subtype candidates, which could be the target of a therapy or diagnosis process [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. They show detailed views of selected genes and patients on demand.

An overview of the diseases and datasets used in the most related genomic approaches is given in Table 3.2. While Lex et al. [LSKS10] analyze only genomic data, the other approaches combine genomic data with clinical data. None of these approaches employs radiomic data or tests the applicability of their interface specifically on prostate cancer datasets. Qu et al. [QLN+19] mention that analyzing genomic data has potential in the personalized therapy of any cancer type.

Table 3.2: Disease and data utilized in genomics approaches.

Paper	**Disease**	**Data**
Lex et al. [LSKS10]	Cancer subtypes	G
Lex et al. [LSS+12]	Cancer subtypes	G + C
Nguyen et al. [NNH+14]	Leukemia	G + C
Turkay et al. [TLS+14]	Breast cancer	G + C
Schneider et al. [SKT+19]	Breast cancer	G + C

Data: Genomics (G), clinical data (C)

3.2.1 Genomic Data Analysis

Lex et al. [LSKS10] perform the data analysis by filtering out inconspicuous genes in the data and applying clustering algorithms. They cluster the data by k-means clustering, affinity propagation, or hierarchical clustering to assign genes with similar functions into groups [LSKS10]. Similarly, Nguyen et al. [NNH+14] apply k-means++ clustering as an extension of the k-means algorithm. This extension chooses initial values that avoid a poor clustering outcome. In their approach, the number of clusters is defined through the interface. Turkay et al. [TLS+14] select the most significantly underexpressed and overexpressed genes through statistical properties. They perform a dual analysis as shown in Figure 3.3 by applying statistics, such as the calculation of the mean, median, or standard deviation, on the dataset rows and the columns separately. For identifying significant differences in the data, they perform a *two-sample Welch's t-test* [Rux06], which does not require an equal variance between subsets and is therefore suitable for genomic datasets [TLS+14]. The user performs the patient stratification manually by selecting subclusters of interest. Their application does not allow the comparison of more than two clusters and does not employ dimensionality reduction methods [TLS+14]. In contrast, Schneider et al. [SKT+19] reduce the dimensionality of the genomic data through Principal Component Analysis (PCA) and 2D t-SNE.

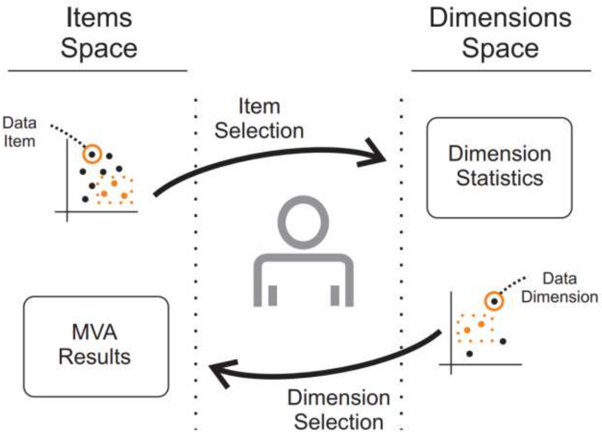


Figure 3.3: Dual analysis of the dimensions space through statistics and of the items space through multivariate analysis (MVA) by iteratively selecting items and dimensions [TFH11].

3.2.2 Visualization of Genomic Data

Genomic data is mainly visualized through scatterplots, matrix heatmaps of clustered data, or genomic coordinates [QLN+19]. Genomic coordinates visualize the data based on their physical location mapped to a reference genome to indicate their function [QLN+19, SGPLB13]. Genomic coordinates can be visualized as a genome browser or a circular plot and represent various alternation types in tumor samples [SGPLB13]. Figure 3.4 shows a heatmap (left) and genomic coordinates (right) in comparison. The genomic coordinates are visualized as a genome browser, while the clinical data is displayed vertically to allow sorting and grouping of the genomic data. In the heatmap example, the columns represent the tumor samples, while the genes are shown in the heatmap rows. To bring distant genomes of interest together, the rows or columns of the heatmap are clustered by clinical features. The color in the heatmap indicates a measurement of the mutational status or expression level of genomic data [SGPLB13]. Furthermore, network visualizations, as depicted in Figure 3.5, represent functional relationships between genes or pathways linked with clinical data [SGPLB13]. Emerging methods for genomic data visualizations include Artificial Intelligence (AI) for predicting and evaluating models or Virtual Reality (VR) to make the data interaction more intuitive [QLN+19].

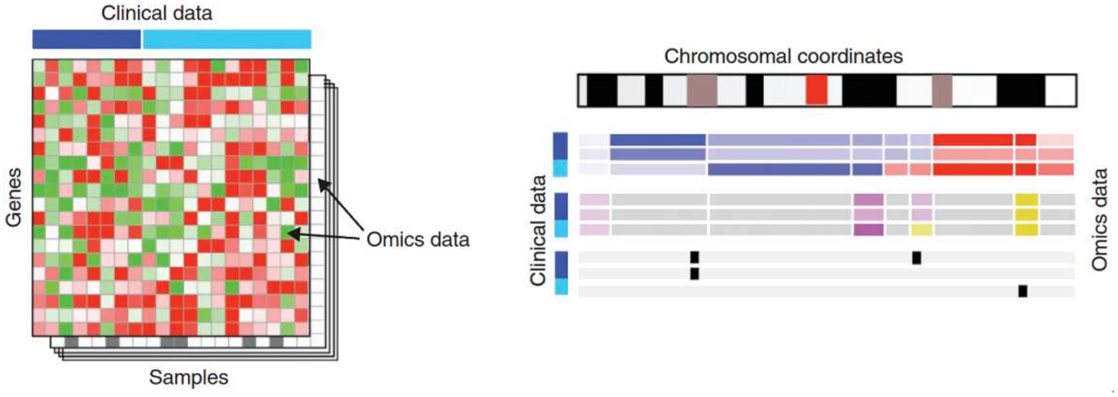


Figure 3.4: Matrix heatmap of genomic and clinical data (left) and genomic coordinates to map the gene location on its chromosomes (right) [SGPLB13].

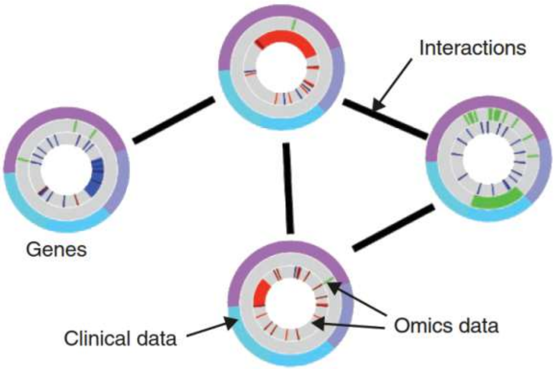


Figure 3.5: Network to visualize cluster nodes representing high interconnections of genes or pathways and the interactions between them [SGPLB13].

Lex et al. [LSKS10] use hierarchical heatmaps to visualize the clustered genes, and parallel coordinate plots to link the genes to experiments. They combine these views in an open bucket in 2.5 dimensional space to visually link relations between the visual representations as shown in Figure 3.6. In a subsequent work of Lex et al. [LSS+12], they split the patient stratification result into visual bricks that represent candidate subtypes, clusters, or categories based on the genomic data loaded into their framework. They allow the user to switch the visual representation between heatmaps, parallel coordinates plots, or histograms on demand. For the default visual representation, they choose a heatmap as it is the most suitable for their primary goal of characterizing cancer subtypes. A schematic representation of the comparison of datasets with different patient stratifications and their subtypes is shown in Figure 3.7.

Turkay et al. [TLS+14] visualize the two spaces of their dual-analysis statistics through two scatterplots linked to a heatmap. They allow a selection of samples through the heatmap and highlight differences in genes between selected and unselected samples on the scatterplot. Similarly, Nguyen et al. [NNH+14] visualize genes of interest through a heatmap and the clustered data as a scatterplot. Schneider et al. [SKT+19] provide an overview of tumor characteristics through a sunburst chart to place pathways and their genes on a circle, as depicted in Figure 3.8. They show up to eight circles in the chart with genomic, clinical data, and sample-specific measurements.

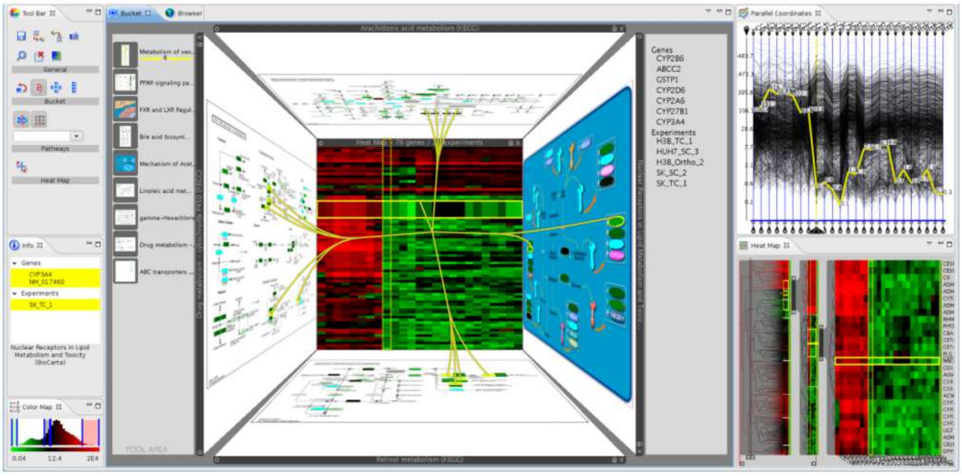


Figure 3.6: Visual representations are placed in an open bucket view in the 2.5 dimensional space. They are linked together through the yellow marks to highlight their relations [LSKS10].

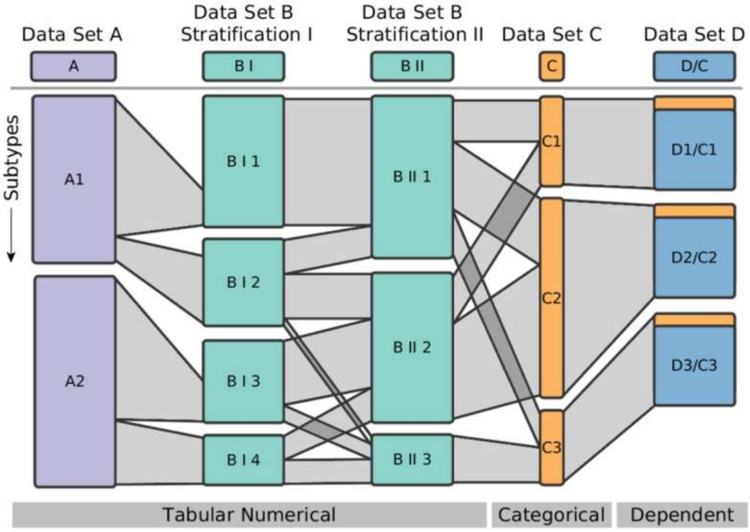


Figure 3.7: Schematic representation of dataset comparison [LSS+12].

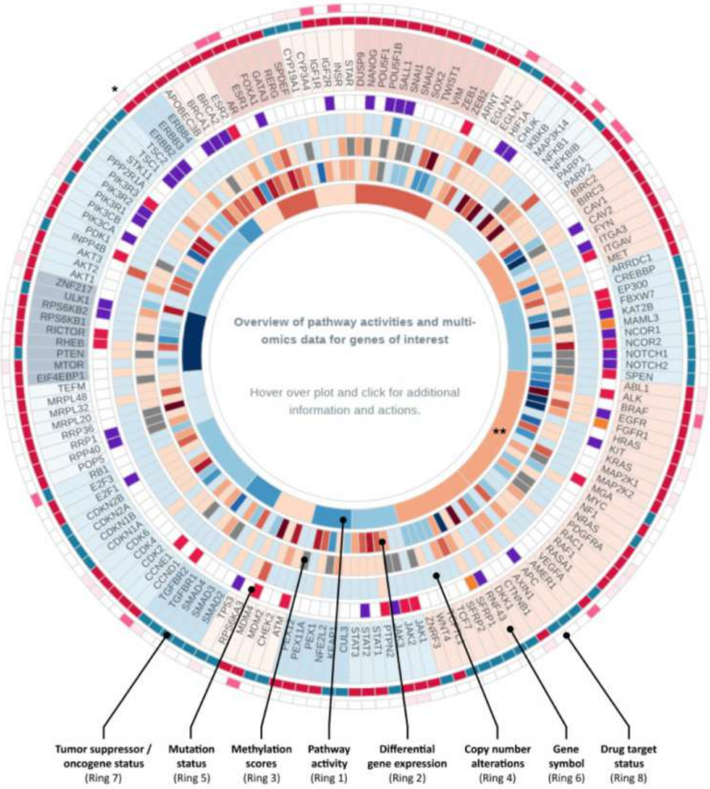


Figure 3.8: Sunburst chart to provide an overview on tumor characteristics [SKT+19].

3.3 Visual Analytics for Clinical Data

Clinical data is essential for research and the healthcare domain [Gro20]. It ranges from demographic data to disease registries or clinical trials [SRY+21, Gro20]. In this section, we summarize three approaches to highlight possible directions of clinical data analysis. Interaction with the data is reached in these approaches through sorting scores, dragging and dropping datasets to the view of choice, brushing on the views to filter the data, or hovering over views to display additional information on demand. Bernhard et al. [BSM+15] visualize histories of prostate cancer patients, while Müller et al. [MSO+20] work with patient-specific data of laryngeal cancer for decision support. Differently, Angelelli et al. [AOH+14] analyze brain measurements for cognitive aging. Angelelli et al. [AOH+14] present a data-cube-based model to handle heterogeneous data of a longitudinal study by linking spatial and nonspatial views. Each person is examined twice to acquire three datasets of brain measurements that they visualize in a cube. Bernhard et al. [BSM+15] focus on visualizing single and multiple patient histories to support physicians in medical research. They design an interactive visualization to define patient cohorts and compare their histories. Müller et al. [MSO+20] assist physicians in clinical decision support by providing transparent recommendations for the diagnosis or therapy plan. These recommendations are generated based on examination findings and clinical knowledge modeled by experts or acquired through machine learning. An overview of analyzed diseases in the presented clinical work is given in Table 3.3.

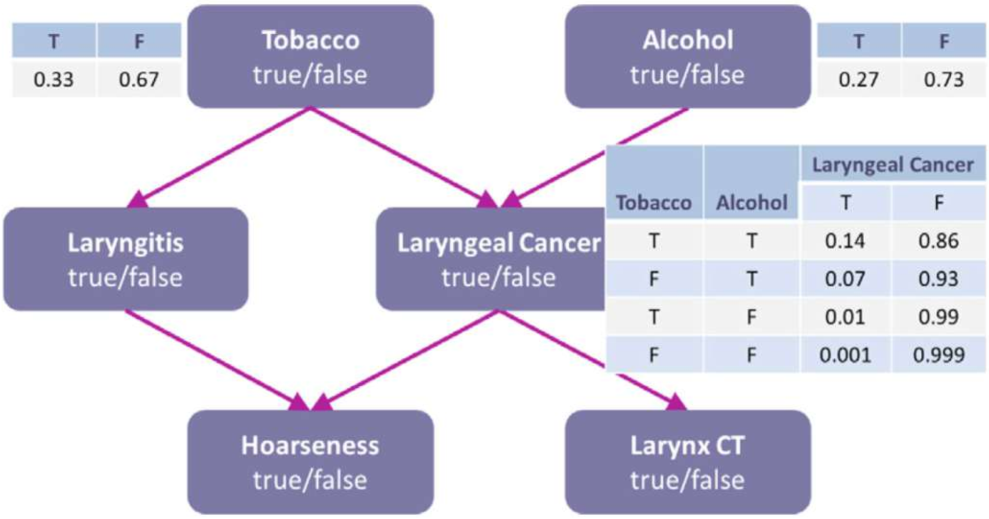
Table 3.3: Disease and data of clinical data analysis approaches.

Paper	**Disease**	**Data**
Angelelli et al. [AOH+14]	Cognitive aging	C
Bernhard et al. [BSM+15]	Prostate cancer	C
Müller et al. [MSO+20]	Laryngeal cancer	C

Data: Clinical data (C)

3.3.1 Clinical Data Analysis

Müller et al. [MSO+20] analyze the data through a causal Bayesian network for decision making. Figure 3.9 shows a network for differential diagnosis of laryngitis and laryngeal cancer. Each node uses the variable values of its parent nodes as an input and outputs a probability distribution saved in a table. The conditional probability table depicts, for example, how tobacco and alcohol influence laryngeal cancer development [MSO+20]. Differently, Angelelli et al. [AOH+14] create data cubes with categorical data as dimensions and quantitative numerical values as measures that they access through dimension coordinates. They work with heterogeneous brain measures with only a partial overlap in the dimensions. Therefore, they use multiple normalized data cubes and a statistical aggregator, such as the average score, to combine uncommon cube dimensions. Figure 3.10 shows this aggregation process of the dimensions.



T	F
0.33	0.67

T	F
0.27	0.73

		Laryngeal Cancer	
Tobacco	Alcohol	T	F
T	T	0.14	0.86
F	T	0.07	0.93
T	F	0.01	0.99
F	F	0.001	0.999

Figure 3.9: Bayesian network for differential cancer diagnosis [MSO+20].

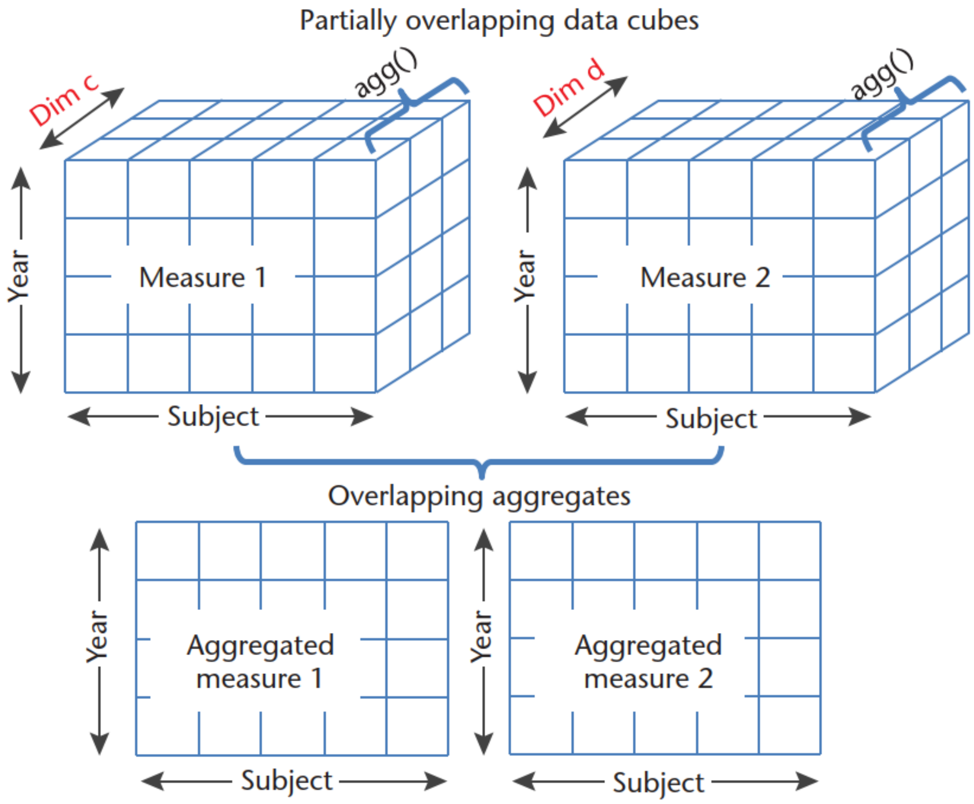


Figure 3.10: Data cubes with categorical data as dimensions and quantitative numerical values as measures accessed through dimension coordinates. The uncommon cube dimensions marked in red are processed through a statistical aggregator. [AOH+14].

3.3.2 Visualization of Clinical Data

Bernhard et al. [BSM+15] visualize large sets of patient histories through a list-based visualization. They present each patient history by a list and allow users to sort these lists by the well-being status of patients that they encode in green, yellow, or red color. To visualize multiple patients at once, they bundle patient information of the list-based visualization through a box plot or area chart. For dynamic cohort queries, they show distributions of patient attributes through bar charts. Müller et al. [MSO+20] employ bar charts to visualize observed evidence items. To present the probability distributions of the Bayesian network, they use donut and pie charts. They encode increased and decreased probabilities through a line texture on the charts to compare distributions resulting from different evidence items. Differently, Angelelli et al. [AOH+14] provide a scatterplot, curve view, and histogram for brain measures and allow the user to select the view of choice.

3.4 Radiogenomics Analysis

Studies and technical surveys demonstrate that radiogenomic data analysis has the potential to identify correlations [IAI+17, SRY+21], predict cancer [IAI+17, SRY+21, SJG+22, HHD+21], and has diagnostic performance comparable to expert knowledge [HLY+23]. It supports understanding cancer biology and behavior, and provides a precise prognosis [SJG+22, SRY+21]. Moreover, it allows clinical experts to discover new knowledge from the data [SJG+22, PPL+19, SRY+21]. Patient stratification entails a meaningful grouping of patients based on similarities or differences in their profiles [SRY+21]. For radiogenomic data, patient stratification through machine learning approaches presents the state-of-the-art in precision medicine [SJG+22, PPL+19].

However, radiogenomics data is diverse, complex, and high-dimensional, which makes the analysis in advanced frameworks, techniques, and algorithms necessary [SRY+21, IAI+17]. Machine learning is suitable for managing and analyzing large and complex data [SJG+22, SRY+21, PPL+19]. Features extracted are often redundant and contain unnecessary information, which leads to overfitting [SRY+21, LXNR19]. This makes the machine learning model not generalizable for new data. Therefore, a dimensional limitation or feature selection is essential to maintain imaging characteristics that have strong correlation with clinical data [SRY+21, LXNR19, SJG+22]. The relevance of features can be evaluated through rank criterion or the use of a weighted sum to maximize relevance and minimize redundancy [SRY+21]. To identify patient groups with highly correlated features, clustering methods are applied. This corresponds to an unsupervised analysis that divides the data into subgroups based on the similarity between samples that is determined through distance measurements such as the Silhouette Coefficient [Rou87].

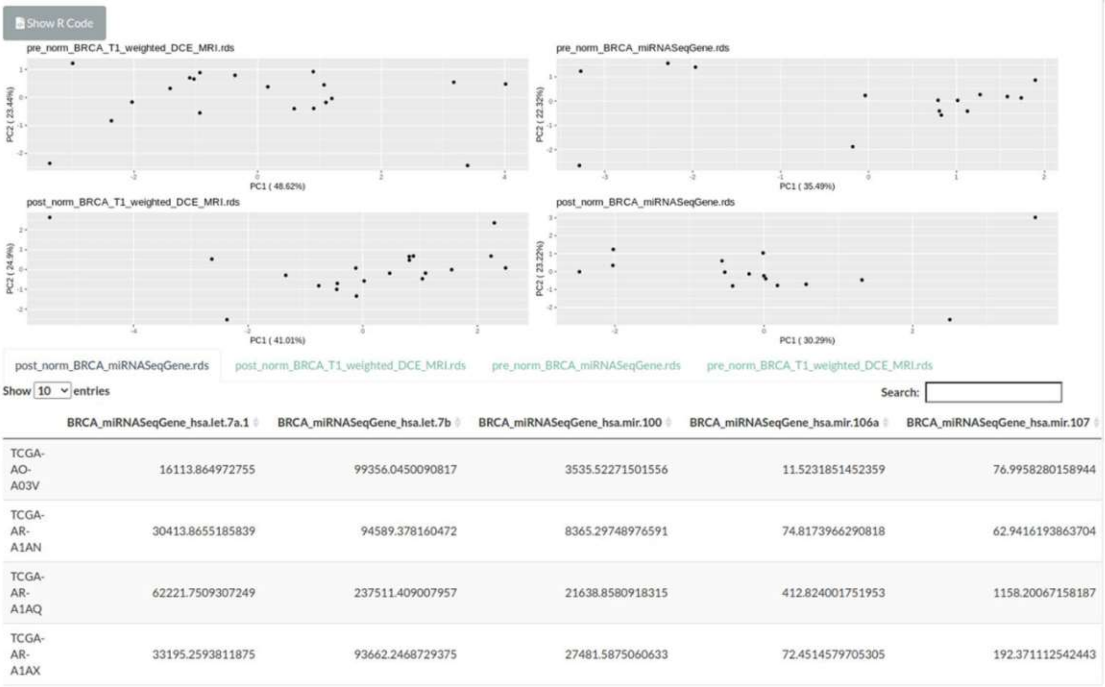
Unsupervised machine learning does not require knowledge of a clinical label and has therefore broad application areas. In contrast, supervised machine learning is used when the treatment endpoints, such as tumor control or toxicity grades, are known [SRY+21]. However, supervised machine learning approaches require a large number of training samples to avoid overfitting [SRY+21]. When large amounts of labeled cohort samples are available, deep learning is applicable and the preferred method as it performs better on unstructured data [SRY+21, LXNR19].

Radiogenomics studies can be exploratory or hypothesis-driven [SRY+21]. In exploratory studies, hierarchical cluster analysis is widely applied to gene expression data [SRY+21]. It groups similar objects into distinct clusters and shows hierarchical relationships between the clusters through a dendrogram. Hypothesis-driven approaches are based on imaging phenotypes collected by medical researchers that they investigate with a specific hypothesis in mind [SRY+21, LXNR19].

Future directions of radiogenomic data analysis go towards the creation of interpretable models [SRY+21]. Without interpretability, radiogenomics analysis is inconvenient and not applicable in clinical practice [SRY+21, HLY+23]. Furthermore, radiogenomic approaches must be validated on independent cohorts to assess their clinical transformability [HLY+23, LXNR19]. This faces challenges in the heterogeneity of cancer diseases [SJG+22] and requires the identification of signatures of intra- and inter-tumor heterogeneity in an anatomic context [IAI+17]. Clinical applicability requires robust approaches based on multi-institutional data that offer generalizable and cross-validated results [SJG+22]. An inter-disciplinary collaboration between clinicians and technical researchers helps in combining technical perspectives of analysis approaches with clinical relevance [BTNK+19].

A first step in combining and analyzing radiogenomic data with respect to clinical data in one framework is performed by Zanfardino et al. [ZCP+21] for breast cancer patients. They present a framework named *MuSA* that is based on the *MultiAssayExperiment* data structure of the *R* package to store and handle multiple heterogeneous data types and facilitate the selection of features in the dataset. They allow the user to filter the data by features of interest or by selecting the patient subset with complete data without invalid values. They do not offer an imputation of missing data, but exclude missing values from the analysis, which leads to a loss of patient information. After filtering the data, they optionally perform a data normalization on the resulting subset and apply PCA to compare different data normalization methods as depicted in Figure 3.11. They do not offer other dimensionality reduction methods than PCA. Their interface focuses on the data preprocessing steps and performs a preliminary data analysis through statistical methods [ZCP+21]. They offer a clustering analysis, a correlation analysis, and feature selection. For the feature selection and clustering analysis, the user chooses the features of interest to display them in a heatmap. The distance method, the cluster linkage, and the number of clusters are specified through the interface. The correlation analysis is performed through a correlation plot between selected radiomic and genomic features for different tumor stages. Using this framework requires the set-up of a configuration file to tag radiomic and genomic datasets. A further limitation is that it does not support the use of gene mutation data [ZCP+21]. Although *MuSA* combines radiogenomic and clinical data in one framework, it focuses on data preprocessing and offers limited analysis, visualization, and interaction capabilities. It does not allow free exploration of the data to gain new insights for knowledge discovery, which is one of the main goals of our work.

In comparison, we provide a visual interactive interface that allows users to perform a free selection and repeated analysis of patients or features of interest. We determine and visualize characteristics and differences of patient clusters, and encourage the user to freely interact with visualizations to support the sensemaking of the data and the identified patterns. These interaction capabilities include, for example, the investigation of distribution plots or highlighting feature combinations on the visualization. We test and compare different imputation, dimensionality reduction, and clustering methods, and allow the identification and removal of outliers on demand. The outcome of these options can be freely explored through the interface, while we provide presets for the preferred settings based on evaluating the outcome of the different machine learning algorithms by cluster separation metrics. Moreover, we allow the users to assess any hypothesis in mind for the underlying data and to highlight the resulting patient subset on the visualization that the users can freely process and explore further.


Show R Code
pre_norm_BRCA_T1_weighted_DCE_MRI.rds
PC2 (23.44%)
PC1 (48.62%)
pre_norm_BRCA_miRNASeqGene.rds
PC2 (22.32%)
PC1 (35.49%)
post_norm_BRCA_T1_weighted_DCE_MRI.rds
PC2 (24.9%)
PC1 (41.01%)
post_norm_BRCA_miRNASeqGene.rds
PC2 (23.22%)
PC1 (30.29%)
post_norm_BRCA_miRNASeqGene.rds
post_norm_BRCA_T1_weighted_DCE_MRI.rds
pre_norm_BRCA_miRNASeqGene.rds
pre_norm_BRCA_T1_weighted_DCE_MRI.rds
Show 10 entries
Search:

	BRCA_miRNASeqGene_hsa.let.7a.1	BRCA_miRNASeqGene_hsa.let.7b	BRCA_miRNASeqGene_hsa.mir.100	BRCA_miRNASeqGene_hsa.mir.106a	BRCA_miRNASeqGene_hsa.mir.107
TCGA-AO-A03V	16113.864972755	99356.0450090817	3535.52271501556	11.5231851452359	76.9958280158944
TCGA-AR-A1AN	30413.8655185839	94589.378160472	8365.29748976591	74.8173966290818	62.9416193863704
TCGA-AR-A1AQ	62221.7509307249	237511.409007957	21638.8580918315	412.824001751953	1158.20067158187
TCGA-AR-A1AX	33195.2593811875	93662.2468729375	27481.5875060633	72.4514579705305	192.371112542443

Figure 3.11: Comparison of the datapoints on the scatterplot before and after normalization. The points are reduced through PCA. The table on the bottom shows the data values after processing [ZCP+21].

Our main contribution to the state-of-the-art is therefore the combination of radiogenomic and clinical data in a visual analytics framework to support the data exploration and sensemaking process. We investigate, design, and develop a visual interactive interface that enables cancer experts and biomedical data scientists to obtain new insights and knowledge from the data and understand the underlying analysis and visualization components. Moreover, our interface allows them to interactively confirm, reject, or refine their hypotheses within cohorts of prostate cancer patients.

CHAPTER 4

Visual Radiogenomics Analysis

In this chapter, we present our approach for the integrated exploration and analysis of radiogenomic data together with clinical information in a visual interactive interface. Figure 4.1 gives an overview on the main tasks of our application.

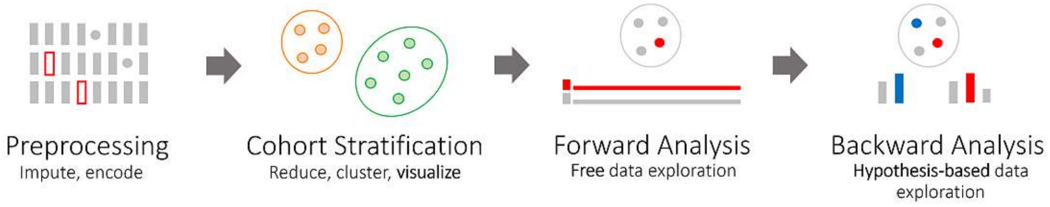


Figure 4.1: Main steps of our visual radiogenomics analysis approach.

We work with datasets that contain mixed data types and missing values. To process the data automatically, we first clean up the data by identifying and replacing any undesired symbols, encoding the categorical data values, and imputing missing values in the data (Figure 4.1, *Preprocessing*). Our data has a high dimension of 10 478 features. Therefore, we reduce its complexity by applying dimensionality reduction algorithms to it. We identify patterns in the data by clustering the data through unsupervised machine learning methods. This results in a cohort stratification, based on different or similar patient profiles, that highlights patterns in the data (Figure 4.1, *Cohort Stratification*). We support users in understanding these patterns through the forward analysis step that allows users to freely explore the data for knowledge discovery. This includes investigating the main characteristics and differences of clusters, or selecting and processing any subset of the data. Furthermore, we show feature distributions on demand to understand why features characterize or differentiate patients (Figure 4.1, *Forward Analysis*). The last step of the pipeline represents the backward analysis to confirm or reject any present hypothesis on the underlying data. This is performed through interactively filtering the data and allows the user to refine a hypothesis by identifying feature thresholds that lead to the correctness or incorrectness of a hypothesis (Figure 4.1, *Backward Analysis*).

4.1 Data – Users – Tasks Analysis

This section gives an overview of our datasets, users, and tasks. We present the radiomic, genomic, and clinical data that we work with in Subsection 4.1.1 and outline the target users of our application in Subsection 4.1.2. Our task definition follows the typology of Brehmer and Munzner [BM13] that we illustrate in Subsection 4.1.3.

4.1.1 Data Characteristics

In this work, we combine radiomic, genomic, and clinical data from a cohort of 89 prostate cancer patients. Our radiomic data is gained from PET/MRI scans. The genomic data represents gene mutation, and the clinical data consists of demographic data and patient management scores determined through clinical assessments. Table 4.1 provides an overview of the number of features per dataset and their main characteristics. We combine the data of 18 clinical parameters, 153 radiomic features, and 10 307 genomic features in the analysis process. The datasets are organized in tables, where each row represents a patient, and each column represents a feature. An integrated analysis of all datasets leads to a dimension of 10 478 features for 89 patients, as illustrated in Figure 4.2.

Table 4.1: Number of features and main characteristics of the datasets.

	Number of features	High dimensional	Missingness
Clinical data	18	–	✓
Radiomics	153	✓	–
Genomics	10 307	✓	–

When the number of features exceeds the number of patient samples, the data is characterized as high dimensional [Nar20]. Therefore, the radiomic and genomic datasets are high dimensional, but not the clinical dataset, which has a dimension of 18. However, the clinical data have their challenges in the missingness of patient scores and in including a combination of different data types. Table 4.2 gives an overview of the mixed data types of the clinical table and the number of missing values per data type. In comparison, the radiomic and genomic data consist all of quantitative continuous values without missingness. The radiomic data can be grouped into 11 feature families, while the genomic data are sparse as most of the elements are zero. A grouping of the genomic data can be performed alphabetically as they consist of gene names starting with alphabetical letters. Furthermore, the genomic data consists of repetitions in parts of the gene names as they represent different mutations. Each patient has 14 to 3 236 gene mutations, while 6 162 gene mutations occur only in one patient with a value higher than zero.

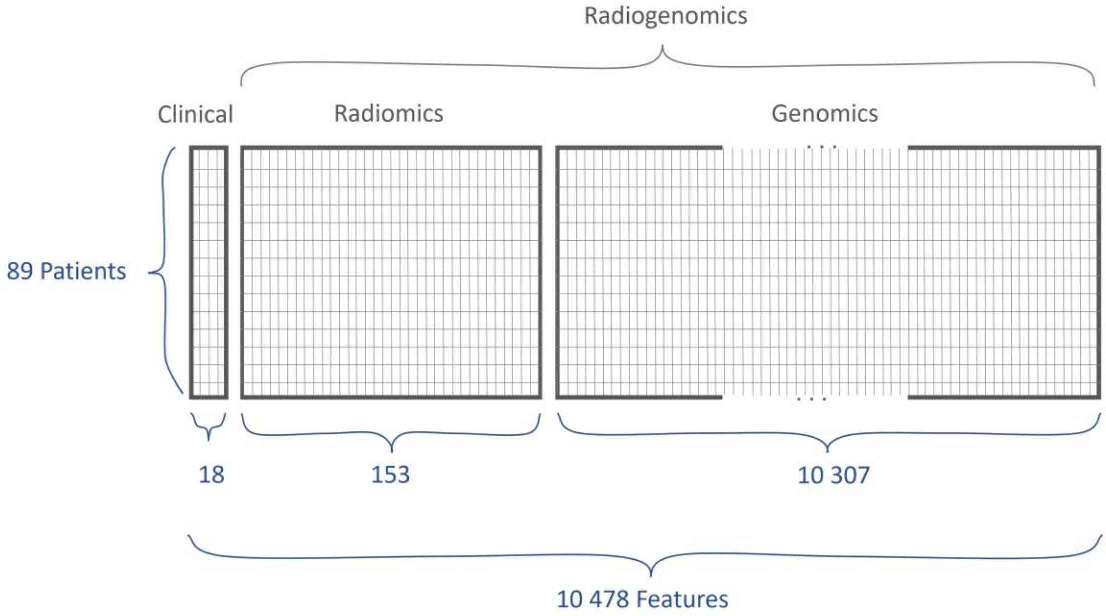


Figure 4.2: Representation of our three datasets as matrix tables.

Table 4.2: Mixed data types of the clinical table and their missingness.

	Quantitative		Qualitative	
	Discrete	**Continuous**	**Ordinal**	**Nominal**
Clinical data	3	6	3	6
Missingness	1	2	0	3

The term *big data* is used in varying contexts with no uniform definition [DMGG15, AMAK19]. De Mauro et al. [DMGG15] analyze approaches in this context and identify topics ranging from data that include complex information to data with a large feature or observation size, as well as data that require innovative information processing and visualization techniques for insight. Based on the most prominent definitions, they suggest using the term for data with high volume, velocity, and variety that require analytical methods to be transformed into valuable information, which applies to our high-dimensional datasets. *Volume* refers to the amount of data, *velocity* to the processing speed, and *variety* to different data types or sources [AMAK19]. Furthermore, our datasets must be analyzed and visualized to gain insight into them and transform the tables into valuable information.

Dirty data refers to data that consist of duplicates, missingness, inconsistencies, or errors, which lead to wrong results or misleading statistics [GGAM12, KCH^{+}03, RD00]. Based on this definition, our datasets are considered as *dirty*, as they contain missingness in feature values, and inconsistencies such as mixing dot and comma symbols in numerical features, that need to be handled properly in order to be interpreted correctly. Furthermore, an early version of the data consist of duplicates in patient information that we resolved with our domain experts. To enhance the data quality and prepare it for the analysis phase, these data needs to be cleaned by identifying and correcting its inconsistencies and missingness [RZ19, GGAM12]. The validation and verification of these corrections must involve domain experts of the data field [RZ19].

4.1.2 Target Users

Our visual analytics application targets two main user groups. On the one hand, we address cancer experts, such as pathologists, biologists, biochemists, and nuclear medicine physicians. On the other hand, we target data scientists and bioinformaticians working with cancer experts on the data. Cancer experts want to investigate and understand the mechanisms behind the data. By identifying correlations and patterns in the data and highlighting features that explain these patterns, they get deeper insights into the data that encourage knowledge discovery and data sensemaking. They further desire to check the correction of any hypothesis or biological mechanism they have in mind involving features of the different datasets or to create and refine new hypotheses on the data. These insights support them in getting a better understanding of how different tumor types work and help them to specify clinical markers that might be relevant for cancer research, diagnosis, and treatment. Furthermore, data scientists and bioinformaticians want to use our application to get insight into the data and the underlying analysis algorithms. They want to interactively change and compare different algorithms and data analysis methods to understand how these changes affect the resulting patterns, feature distributions, or hypotheses of interest. Moreover, they are interested in understanding the automated data analysis components of our framework.

4.1.3 Typologies and Tasks

Task typologies provide a model to specify aspects and goals of tasks in a unified and precise way [BM13, RAW$^+$16]. They allow the comparison and evaluation of visualization approaches and justify the creation of suitable visual representations [RAW$^+$16, SNHS13, BM13]. Schulz et al. [SNHS13] present five design dimensions to characterize the main aspects of tasks. Besides specifying *why* and *how* a task is performed, they describe the data through three dimensions representing the *characteristics*, *target*, and *cardinality* of visualization tasks. Similarly, Brehmer and Munzner [BM13] present a typology that is based on specifying *why* and *how* tasks are performed. Different than Schulz et al. [SNHS13], Brehmer and Munzner [BM13] represent the data of tasks by a *what* dimension that characterizes the task inputs and outputs. Rind et al. [RAW$^+$16] also present a three-dimensional task typology. In contrast to the approaches of Brehmer and Munzner [BM13] and Schulz et al. [SNHS13], they do not aim for an intermediate level that captures all task aspects, but describe tasks on a higher level that emphasizes diverse task concepts.

We define our tasks based on the multi-level typology of Brehmer and Munzner [BM13] that closes the gap between high- and low-level tasks. It allows the expression of complex tasks as sequences of independent simpler tasks through concise and flexible descriptions. Tasks are chained together by using the output of a prior task as an input to subsequent tasks, which enables the visualization of complex relations [BM13]. Figure 4.3 represents our main tasks and the inter-relationships between them depicted through red arrows. The cohort stratification includes dimensionality reduction, clustering, and visualization steps. These steps are also required by the forward and backward analysis as depicted by the red arrows. Furthermore, the backward analysis can be applied on the result of the cohort stratification or a processed subset resulting from the free analysis. Similarly, the free analysis can be applied on the data selected through a hypothesis resulting from the backward analysis, on the automatically created data clusters resulting from the cohort stratification, or on the unstratified data by manually selecting a subset of interest to compare its characteristics, differences, and distributions with other patients, or apply the cohort stratification on it.

In the upcoming sections, we present the tasks and their subtasks in more detail and describe how we fulfill them. We illustrate the subtasks of the preprocessing in Section 4.2, of the cohort stratification in Section 4.3, of the forward analysis in Section 4.4, and of the backward analysis in Section 4.5.

4.2 Preprocessing (T0)

Why: ***produce*** data with enhanced quality

How: ***derive*** a cleaned data representation, encode mixed values, impute missingness in the data, identify outliers

Input: data of mixed type with missingness

Output: high-dimensional numeric data without missingness

For the data analysis, we process the data tables automatically through statistical measures or unsupervised machine learning algorithms. This requires data cleansing, encoding, imputation, outlier detection, and scaling to ***produce*** data with enhanced quality by correcting errors, inconsistencies, and missingness in the data, by detecting outliers on demand and scaling all datasets to the same ranges. For each of the subtasks, we ***derive*** a data representation that matches the respective goal. The encoding and imputation step is only required for the clinical data to handle categorical values and missingness in the data. In contrast, the further preprocessing steps are applied on all datasets. We automatically process all tables separately and combine them by default to use them as an input for the cohort stratification or free analysis tasks.

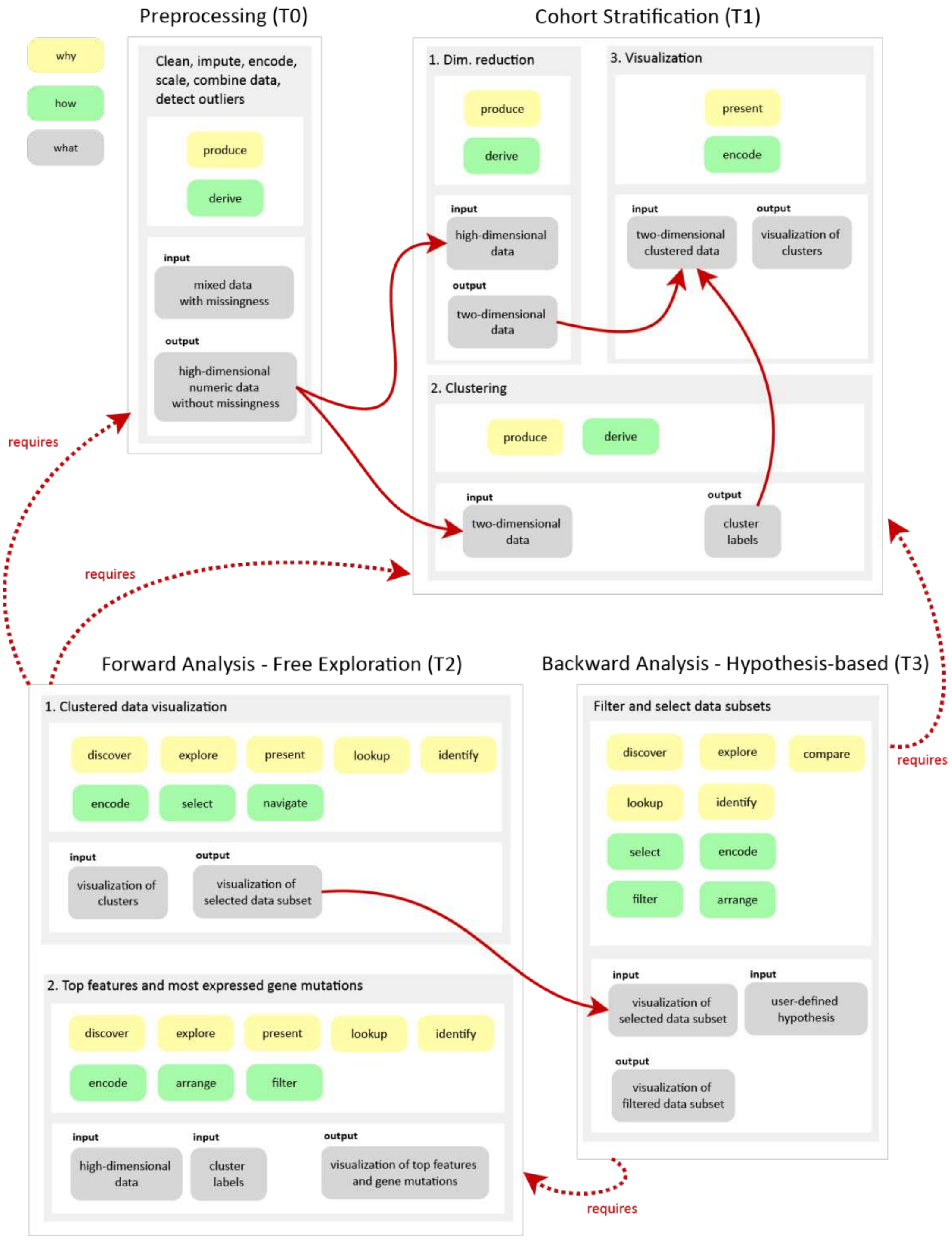

Preprocessing (T0)
Cohort Stratification (T1)
why
how
what
Clean, impute, encode, scale, combine data, detect outliers
produce
derive
input
mixed data with missingness
output
high-dimensional numeric data without missingness
1. Dim. reduction
produce
derive
input
high-dimensional data
output
two-dimensional data
3. Visualization
present
encode
input
two-dimensional clustered data
output
visualization of clusters
2. Clustering
produce
derive
input
two-dimensional data
output
cluster labels
requires
requires
requires
requires
Forward Analysis - Free Exploration (T2)
Backward Analysis - Hypothesis-based (T3)
1. Clustered data visualization
discover
explore
present
lookup
identify
encode
select
navigate
input
visualization of clusters
output
visualization of selected data subset
2. Top features and most expressed gene mutations
discover
explore
present
lookup
identify
encode
arrange
filter
input
high-dimensional data
input
cluster labels
output
visualization of top features and gene mutations
Filter and select data subsets
discover
explore
compare
lookup
identify
select
encode
filter
arrange
input
visualization of selected data subset
input
user-defined hypothesis
output
visualization of filtered data subset

Figure 4.3: Our main tasks and their inter-relationships depicted by red arrows, according to the multi-level typology of Brehmer and Munzner [BM13].

4.2.1 Data Cleansing (T0a)

Why: ***produce*** a data representation to process the data automatically

How: ***derive*** a representation with checked types, format, and symbols in data

Input: uncleaned high-dimensional data

Output: cleaned data with consistent symbols and the correct data types

Data cleansing enhances the data quality by identifying and correcting errors and inconsistencies in the data [RZ19, GGAM12]. In this subtask, we determine and remove undesired symbols in numerical fields and ***produce*** a consistent representation of all data types. Our clinical data contains, for example, the expression $< X$, where X represents a constant number. To process this field as a number, we must replace the *less-than* $<$ symbol and reduce the numeric value by a constant specified with our domain experts. Other examples include qualitative values concatenated by a *plus* $+$ symbol and not depicted in a uniform order, while their order has no clinically relevant meaning. After importing the data, we ***derive*** a cleaned data representation by resolving these cases through replacement rules that we define and confirm together with our collaborating domain experts. Kim et al. [KCH+03] provide a taxonomy to identify misspelled, inconsistent, or wrong data for replacement. Alternative methods for cleansing big data are based on machine learning techniques or knowledge-bases to discover error patterns in the data and possible solutions [RZ19]. However, the identified inconsistencies in our data are limited and we want to ensure that our domain experts are aware of these cases and confirm with the replacement rules as depicted by Ridzuan et al. [RZ19]. Therefore, the rule-based method we implemented is the most suitable for our use case.

4.2.2 One-Hot Encoding for Categorical Features (T0b)

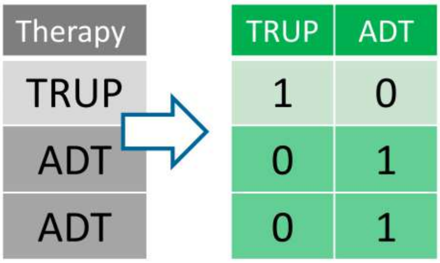
Why: ***produce*** a data representation to process the data automatically

How: ***derive*** a representation that encodes categorical values

Input: cleaned data from subtask**(T0a)**

Output: one-hot encoded data with numerical values only

Besides quantitative values, our clinical data consist of qualitative ordinal and nominal values. In our data, the ordinal values are already encoded by numbers considering their natural order. Therefore, this task addresses the encoding of qualitative nominal values. The most common technique for encoding qualitative values with minimal processing, is the one-hot encoding algorithm [HK20, Lu20, CV22]. We apply one-hot encoding on nominal values of the clinical data to ***produce*** a data representation with only numerical values. To ***derive*** this representation, we split each nominal option into an own binary-encoded feature as specified through the algorithm. Figure 4.4 illustrates this process by a *Therapy* feature with two options that is splitted into two separate binary-encoded features. One-hot encoding leads to a higher dimensionality of the data, which is a disadvantage for features with high cardinality as it increases the complextity and processing time [Lu20, CV22]. Advanced algorithms overcome this limitation through low-dimensional or quasi-orthonormal encodings [Lu20, CV22]. However, as the cardinality of the affected features in our data is limited to five options, the slightly increased dimensionality is negligible for our data.



Therapy
TRUP
ADT
ADT

TRUP	ADT
1	0
0	1
0	1

Figure 4.4: One-hot encoding of the *Therapy* feature by splitting it into separate *TRUP* and *ADT* features to transform the nominal values into numerical ones.

4.2.3 Data Imputation (T0c)

Why: ***produce*** a data representation to process the data automatically

How: ***derive*** a representation that imputes missing values

Input: cleaned, encoded data with missingness from subtask **(T0b)**

Output: imputed data without missingness

The clinical data consist of six features with missing values. Figure 4.5 shows the affected features and their percentage of missingness. Three of them are quantitative values. These are the discrete feature *BCR time*, and the continuous features *Post PSA* and *BCR PSA*. In addition to the quantitative values, the data includes three qualitative values with missingness, from which the *Tumor margin* and *BCR status* are binary, while the *Pre-OP therapy* is a nominal feature with five possible options. The *Pre-OP therapy* feature has the smallest missingness with 1.52 %. In contrast, the *BCR PSA* has the highest missingness score of 69.7 %.

Missing data can be categorized by the following three types [ANI$^+$20, van18]:

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

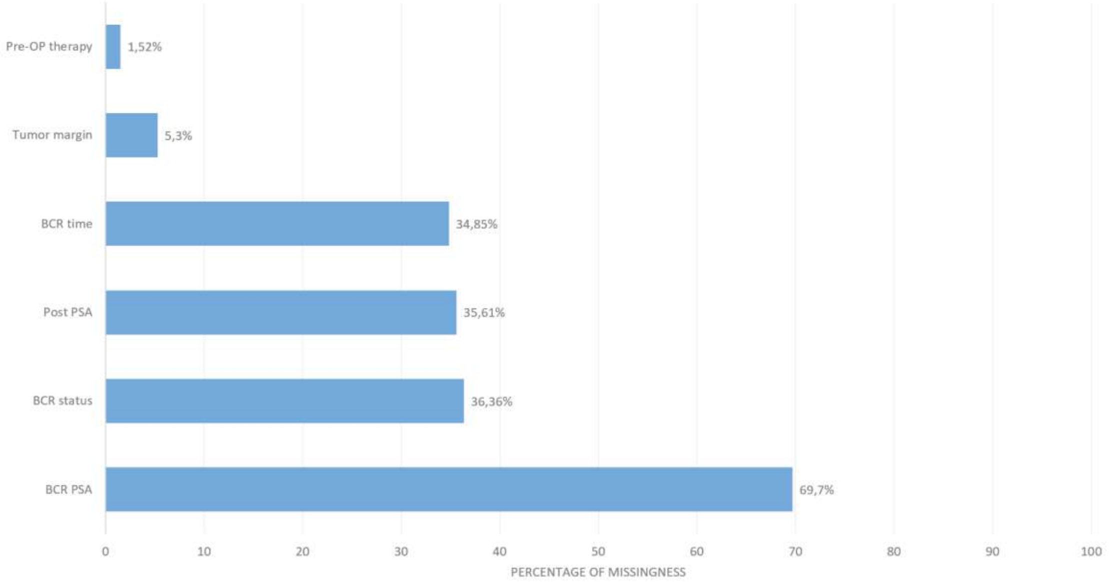


Figure 4.5: Percentages of missingness in clinical data per feature.

MCAR applies, when there is no semantic reason for the missingness in the data [ANI+20]. This could be the case when the sample tube with blood breaks [ANI+20] or the weighing scale runs out of batteries and hinders the measurement [van18]. For this missingness type the observed data are expected to be unbiased and to match the distribution of the random subset of missing scores [ANI+20, van18].

MAR occurs when the missing data depend on observed variables and can be explained through it [SWC+09]. An example is a blood measurement that is performed only for old patients, while the observed data also consist of young patients. This may cause bias in the data as the distribution of the missing and observed data is expected to be different [ANI+20].

MNAR is the case when the missing data depend on the data, but the missingness cannot be explained through the observed variables [ANI+20, van18]. In this case, the variables that could have explained the missing data are not observed. An example is a blood measurement that is performed only for old patients, while the age is not documented to make this clear. Therefore, the distribution of the missing and observed data is expected to be different, which leads to bias in the data [ANI+20].

We consider most of the missingness in our data to be MCAR as all measurements are usually performed for these patients, independent from other clinical scores. These measurements are missing as, for example, patients did examinations partially at different hospitals. However, we identify a dependency pattern between the *BCR PSA* feature, which has the highest missingness, and the *BCR status*. In our case, all values that have a *BCR status* of zero, have a missing *BCR PSA* value. Therefore, we consider this feature as MNAR and substitute it with a constant for all cases that have a *BCR status* of zero to not induce a biased distribution.

The difference between the missingness types, especially between MAR and MNAR, is not always clear and depends on background knowledge and assumptions on the data [ANI+20]. MNAR is the most complex situation that is resolved through measuring additional data and analyzing the sensitivity of the results in different scenarios [van18]. In contrast, MCAR is the easiest to handle as it allows deleting or ignoring the missing data, which is not expected to affect the distribution. However, this handling leads to reduced statistical power and a loss of valuable data samples [van18, ANI+20]. Furthermore, deletion is not a suitable solution for MAR [van18]. Therefore, it is advisable to conduct data imputation to predict and replace missing values [van18, ANI+20]. The most frequent approaches apply single imputations, which replace the missingness with the mean or median value of the feature. However, this reduces the variability of the distribution and therefore leads to a biased estimation [ANI+20]. Advanced options that deliver an appropriate level of accuracy and bias use multiple imputation algorithms, which also consider the dependency between variables [ANI+20, van18]. This includes the algorithm Multivariate Imputation by Chained Equations (MICE). It starts with a simple imputation and learns a regression model for the missing values to replace them with predictions from the regression model. This process is repeated for a maximum number of iterations [ANI+20].

Garrison et al. [GMS+15] test five different imputation methods to assess the robustness of their approach. In their work, MICE and Principal Components Imputation lead to a similar result. The general patterns in the data are preserved, which confirms the robustness of their approach. However, each imputation algorithm leads to a different outcome. They perceive most of the differences in the imputation of binary features. Stavseth et al. [SCR19] compare six methods for multiple imputation. They mention that all methods perform well for sample sizes above 1000, while for sample sizes smaller than 200 the result depends on the amount of missingness in the data.

To identify the most suitable imputation method for our data, we test, evaluate, and compare the error metrics of seven different imputation methods. Similar to Moritz et al. [MSB+15], we simulate missingness on the complete patient subset of our clinical data. We evaluate the imputation errors by the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE) between the imputed and the true value that we remove for testing. For this experiment, we simulate missingness percentages ranging from 5 % to 95 % in the data to assess the robustness of the algorithms. Then, we choose the best method and parameters that lead to the smallest errors based on the missingness percentage of the respective feature in our data and by considering the data type of the feature. We set the best option we determined as the default imputation method in our interface that the user can change on demand. In most cases, RMSE and MAPE indicate the same result. If this is not the case, we give preference to the RMSE. We compare single imputation methods by using the mean, median, or most frequent value in the non-missing data feature. Furthermore, we apply linear regression, and multiple imputation methods, such as MICE and KNN. For multiple imputations, we test imputing the values by using all feature dimensions of the clinical data, or by considering only the numeric features through excluding binary and string features. Additionally, we test the imputation using all features in the clinical data that have the same type. For example, we test imputing discrete numerical features by considering only all features with discrete numerical data type. This option leads in most cases to the lowest error rates of the multiple imputation methods, while single imputation methods create the smallest errors in our experiment. We present our results and evaluation of the imputation methods in Subsection 5.2.1.

4.2.4 Outlier Detecion (T0c)

Why: ***produce*** a data representation with enhanced quality that contains no outliers

How: ***derive*** a representation that removes outliers on demand

Input: imputed data from subtask **(T0c)**

Output: data without outliers (if desired)

Outliers represent single points or small clusters that are different or inconsistent compared to the remaining set of data [DXLL09]. Keeping and analyzing them could be of interest due to their uniqueness [BZA20, DXLL09]. However, outliers may also affect the accuracy and stability of automated data analysis and influence the cluster structure [LLWF21]. When applying clustering on the data, outliers might not belong to any of the existing groups [LLWF21]. This subtask allows users to identify, highlight, remove outliers, or compare them with the remaining data points on demand, when combined with the upcoming forward and backward analysis steps.

Outlier detection can be performed globally or locally [BZA20]. Local methods detect points based on their characteristic differences to their neighborhood, while global methods analyze differences compared to the whole dataset [BZA20]. We deploy the unsupervised machine learning method *isolation forest* for global outlier detection and removal. This algorithm has linear time complexity and low memory requirement [LTZ08]. Therefore, it is suitable for high-dimensional data and data with a large number of less prominent attributes [LTZ08]. For local outliers, we use the density-based *local outlier factor* algorithm. Basd on Cheng et al. [CZD19], it performs well in detecting local outliers. However, each method focuses on detecting outliers either globally or locally. Cheng et al. [CZD19] propose an extension by combining both algorithms in a progressive two-layer method. Figure 4.6 shows global (top right) and local (bottom right) outlier removal plots applied on the input (left) in comparison. On the global outlier plot (top right) small points represent local outliers that would have been removed by applying local outlier removal. In contrast, global outliers are highlighted by a red circle on the input (left) and the local outlier removal plot (bottom right). After removing outliers, we reduce the dimensionality of the remaining points in this figure through Multidimensional Scaling (MDS) for demonstration purposes.

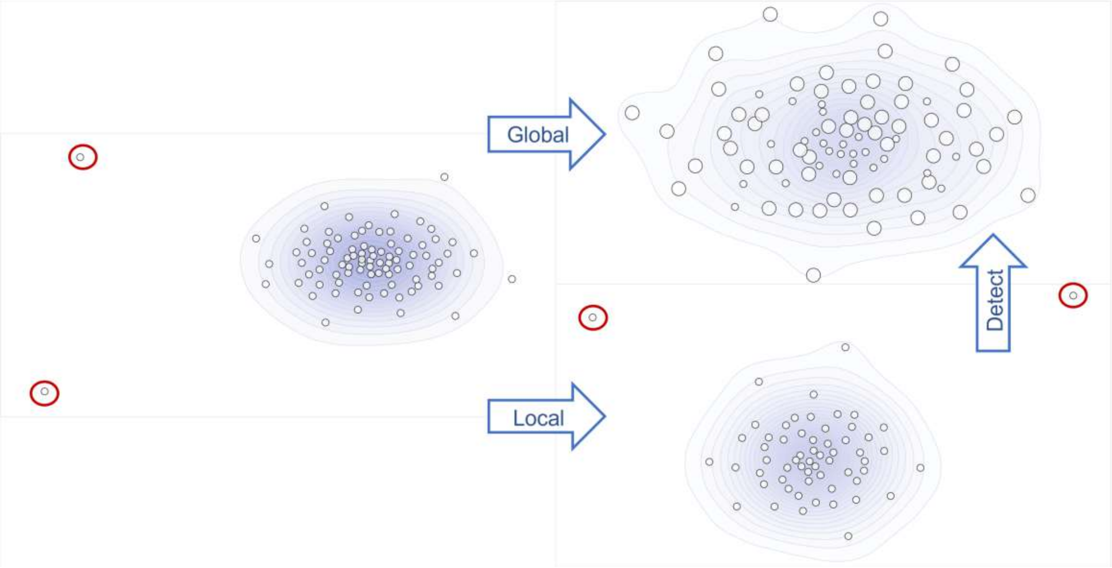


Figure 4.6: Global (top right) and local (bottom right) outlier removal applied on the input (left) in comparison. Global outliers are marked by a red circle. Small points in the top right plot represent local outliers detected on the global outlier removal plot.

4.2.5 Data Scaling (T0d)

Why: ***produce*** a scaled data representation

How: ***derive*** a representation that normalizes or standardizes data

Input: data after outlier detection from subtask **(T0c)**

Output: scaled data through normalization or standardization

Values of datasets include measurements in different units. Analyzing each measurement in its data-dependant scale affects the outcome of the analysis process as values would dominate over others [THFM14]. Data standardization improves the signal-to-noise ratio and the discrimination power of the dataset [Ng17]. In data standardization, the mean (μ) of the data is subtracted from each value (X), and it is divided through the standard deviation (σ) of the data [MAF14]. We depict this through the following formula for the standardized data point x_{stand}:

$$x_{stand} = \frac{X - \mu}{\sigma}$$

Data standardization is preferable for data with a Gaussian distribution and outliers. If the data distribution is not known or the data is not Gaussian distributed, data normalization should be preferred. Data normalization eliminates bias in features with large values compared to features with low values [Ase22]. It scales all data values to a specific range, such as the range between 0 and 1 [MAF14]. In this case, the minimum value (x_{min}) is substracted from the data values (X) and the result is divided through the differences between the maximum and minimum values ($x_{max} - x_{min}$) of data points [Ase22]. The following formula demonstrates this for the normalized data point x_{norm}:

$$x_{norm} = \frac{X - x_{min}}{x_{max} - x_{min}}$$

As the distribution of the radiomic and genomic features in our dataset varies from a Gaussian distribution, we set data normalization as the default option. However, we provide both options to enable users to examine and compare the resulting patterns of both scaling techniques on demand.

4.3 Cohort Stratification (T1)

Why: ***produce*** meaningful groups of patients with similar characteristics
present the identified clusters

How: ***derive*** a reduced and clustered dataset
encode clusters in a visual representation

Input: high-dimensional data from subtask **(T0d)**

Output: visualization of the reduced and clustered data

After the preprocessing steps, we identify and visualize patterns in the datasets 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. In the dimensionality reduction and clustering steps, we ***produce*** data in a reduced and clustered way, while in the visualization step, we ***present*** the data and its identified patterns. Therefore, we ***derive*** a data representation that matches the respective goal and ***encode*** the identified clusters in a visual representation.

4.3.1 Dimensionality Reduction (T1a)

Why: ***produce*** low dimensional data to facilitate the processing and visualization

How: ***derive*** data with low dimensionality from the high dimensional data

Input: high-dimensional data from subtask **(T0d)**

Output: two-dimensional data

Radiogenomic data is complex and high-dimensional, which hinders the identification of patterns in the data. Extracted radiogenomic features contain redundant and unnecessary information that lead to overfitting in the data analysis so that the machine learning is not generalizable for new data [SRY+21, LXNR19]. Eliminating features that lack robustness against variability sources can avoid overfitting [LLD+17]. This is performed by applying dimensionality reduction algorithms on the data [LLD+17]. A reduced dimensionality of the data also helps to maintain imaging characteristics that strongly correlate with clinical features [SRY+21, LXNR19, SJG+22].

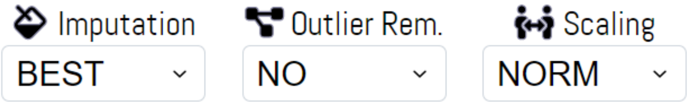
We tested and compared the following five dimensionality reduction methods:

- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Uniform Manifold Approximation and Projection (UMAP)
- Factor Analysis of Mixed Data (FAMD)
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)

PCA is the most widely used algorithm [XWY+21, EMK+21]. It is a linear method that focuses on the global structure of the data and highlights interclass differences. [XWY+21, EMK+21]. It detects patterns dominating in the data and identifies linear combinations that maximize the variance in the data. Therefore, its dimensions represent the data components with the largest variance in the data. In contrast to PCA, t-SNE is a non-linear method that reveals the local structure of the data by minimizing the divergence between two distributions [XWY+21, VDM14]. It transforms Euclidean distances between data points in the high dimensional space into conditional probabilities [XWY+21]. One distribution measures pairwise similarities of input objects, while the other distribution measures pairwise similarities of the low-dimensional points in the embedding [VDM14]. FAMD combines PCA with Multiple Correspondence Analysis (MCA). FAMD has the advantage of being applicable to complex data of mixed types. Therefore, it does not require the one-hot encoding step of task **(T0b)** and is better suitable than PCA for datasets that contain qualitative values or values of mixed datatypes [GMS+15]. In contrast, UMAP is a non-linear method that reveals the global and local structure of the data. It bases on the assumption that the data are uniformly distributed [XWY+21, EMK+21]. Similarly, MDS is a non-linear method that preserves the global and local structure of the data [EMK+21].

Xiang et al. [XWY+21] compare dimensionality reduction methods for high-dimensional and sparse genomic data. They evaluate the algorithms on 30 simulated and five real datasets and identify t-SNE as the method with the highest overall performance and accuracy. Xiang et al. [XWY+21] also mention that t-SNE leads to the highest computing cost. They further identify UMAP as the method with the highest stability, moderate accuracy, and with the second highest computing cost after t-SNE. Garrison et al. [GMS+15] apply the dimensionality reduction iteratively through a user-driven approach that allows users to create and interact with dimensional bundles. Their approach is based on the FAMD algorithm as it allows the combination of mixed data types without encoding the data beforehand.

We selected t-SNE as the default dimensionality reduction technique as it leads to the best cluster separation for our data, as we depict in Subsection 5.2.2. We further allow users to test the outcome of different dimensionality reduction methods and compare the resulting patterns on the scatterplot. Furthermore, 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. As mentioned by Xiang et al. [XWY+21], t-SNE has the highest computational costs in comparison to the other methods. Therefore, we calculate half of the components through PCA and use these for the initialization of t-SNE. Moreover, the dimensionality reduction is applied in our application only once in the beginning and on selected patient or feature subsets only on demand. We further provide the user visual feedback through a waiting indicator while the data is processed. Therefore, it does not influence the interactivity of the application. Figure 4.8–4.16 show the results of the dimensionality reduction medthods. These are created through the default settings of our framework using the *best* imputation method on the data, no outlier removal, and a data normalization as a scaling technique, as depicted in Figure 4.7. We applied MICE imputation on the t-SNE default option in Figure 4.17 to investigate its influence on the resulting patterns.


Imputation: BEST

Outlier Rem.: NO

Scaling: NORM

Figure 4.7: Default options used for the dimensionality reduction results. These are the *best* imputation method based on our evaluation in Subsection 5.2.1, no outlier removal, and data normalization as a scaling technique.

The contours on the results represent an estimation of dense regions and are determined through the Kernel Density Estimate (KDE) of the data points [DLH11]. Furthermore, dense areas are indicated through a blue color, as shown, for example, in Figure 4.10 for the PCA algorithm. Zooming into these structures reveals more details on the data. The MDS (Figure 4.8), FAMD (Figure 4.13), and PCA (Figure 4.10) algorithms produce two outliers, which present the same two patients in all representations. Applying global outlier removal on the data prior to MDS (Figure 4.8) filters these outliers and opens the points up in the available space, as demonstrated in Figure 4.9. In contrast to MDS, the PCA (Figure 4.10) and FAMD (Figure 4.13) algorithms progressively reveal two new outliers in the data when applying outlier removal algorithms prior to the dimensionality reduction. In this case, the zooming functionality can be used to open the points up in the available space and explore the details of the data representations. Figure 4.11 shows a zoomed representation of PCA, which reveals the structure of the data points. Zooming further into the PCA result leads to Figure 4.12, which provides more details on the structure of the reduced space. Figure 4.14 shows the zoomed representation of the reduced space resulting from the FAMD algorithm. The result of UMAP is illustrated in Figure 4.15, while Figure 4.16 represents the result of the t-SNE dimensionality reduction. Contrary to the other dimensionality reduction methods we tested, the t-SNE algorithm divides the data into two clear groupings, as shown in Figure 4.16.

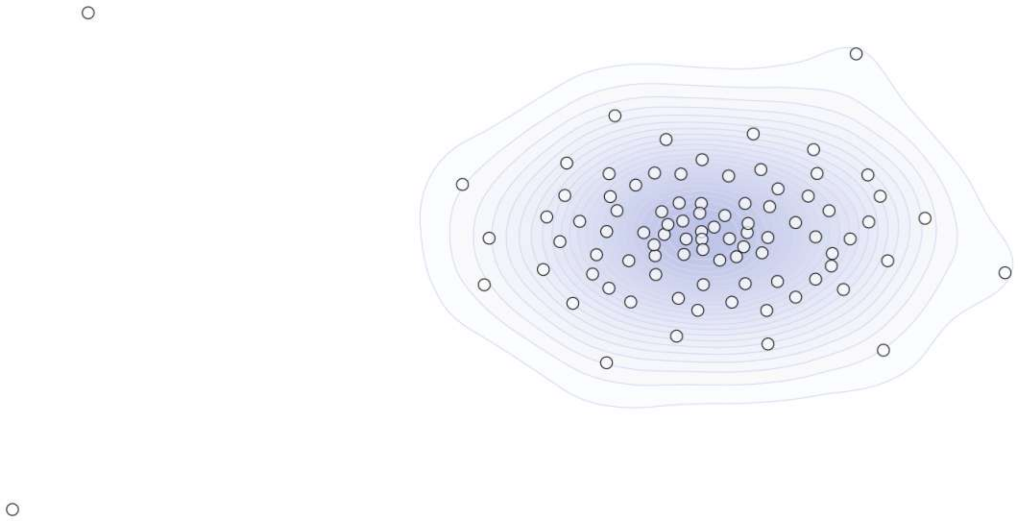


Figure 4.8: MDS dimensionality reduction is applied on the high-dimensional data. This reveals two outliers on the left side with a higher distance to the other data points.

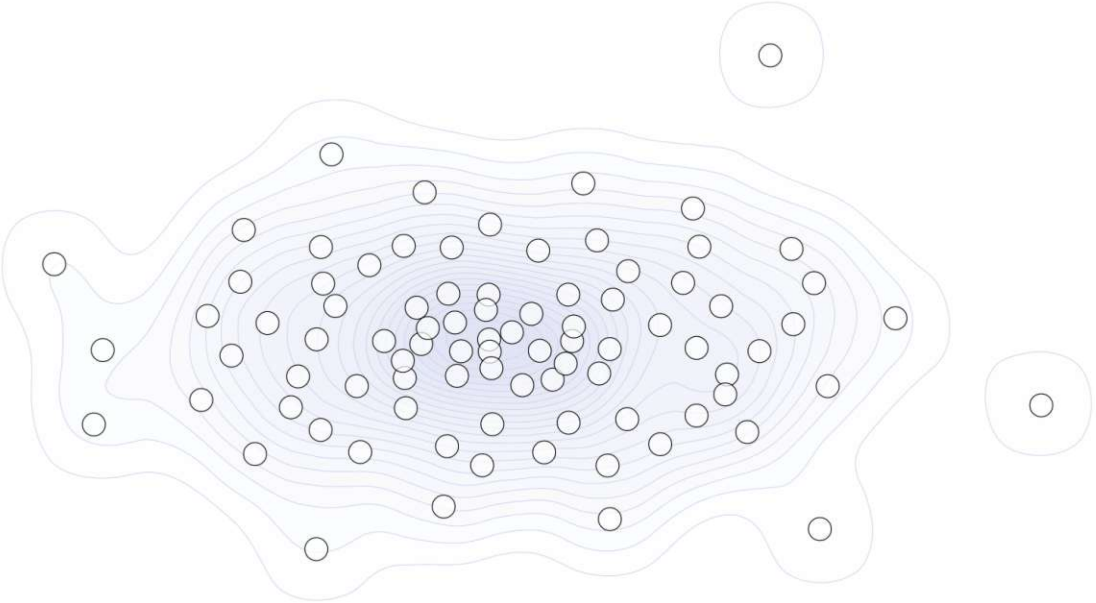


Figure 4.9: Global outlier removal is applied prior to the MDS dimensionality reduction algorithm. This opens up the data points in the available space and reveals details on the structure of the reduced space. It shows, for example, two points on the right side with a larger distance to the other data points.

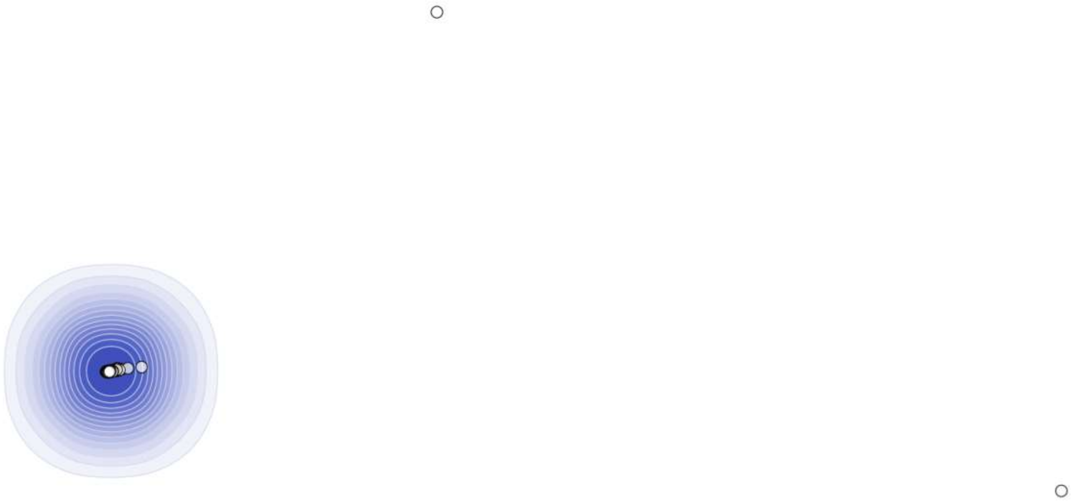


Figure 4.10: PCA dimensionality reduction is applied to the data. This reveals two outliers on the top and right sides with a larger distance to the other data points. Most of the data points are located in a very dense region visualized by the density contours and highlighted through blue color shading.

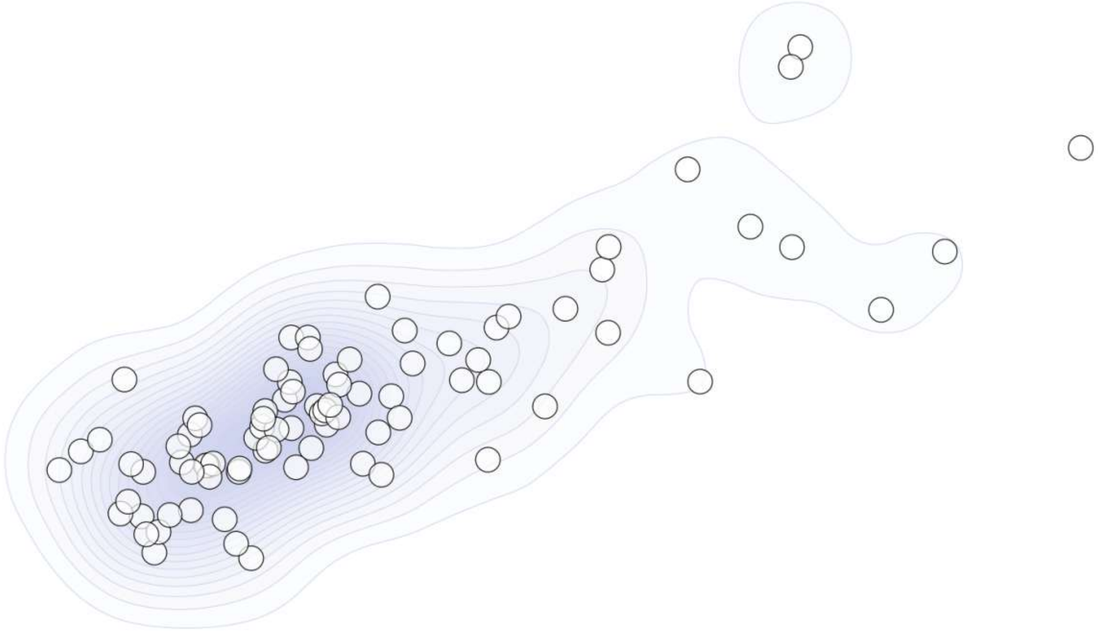


Figure 4.11: Zooming into the reduced space resulting from PCA reveals the structure of the data. The region on the left is dense and consists of overlapping points.

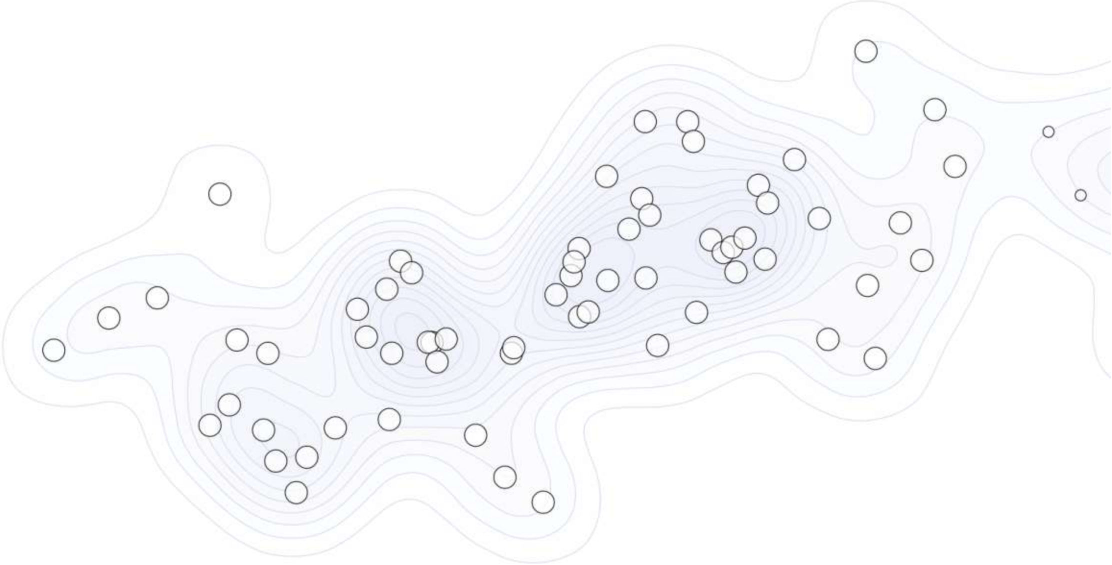


Figure 4.12: Zooming further into the data space resulting from PCA reveals more details on dense regions in the data and reduces the number of overlapping data points.

Figure 4.13: FAMD is applied on the data. It can deal with qualitative values without prior data encoding and reveals two outliers in the data. The majority of the data points are located in a very dense region on the bottom left of the image, as depicted through the density contours and the blue color shading of the region.

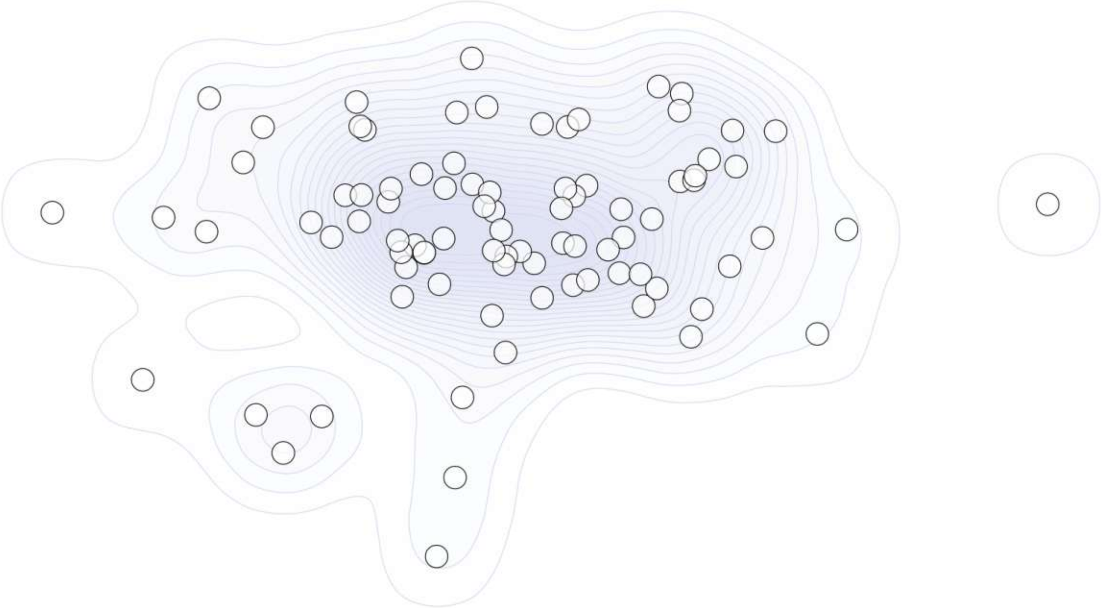


Figure 4.14: Zooming into the reduced space resulting from the FAMD algorithm reveals details on the structure of the data. For example, a small grouping of three points is shown on the bottom left side.

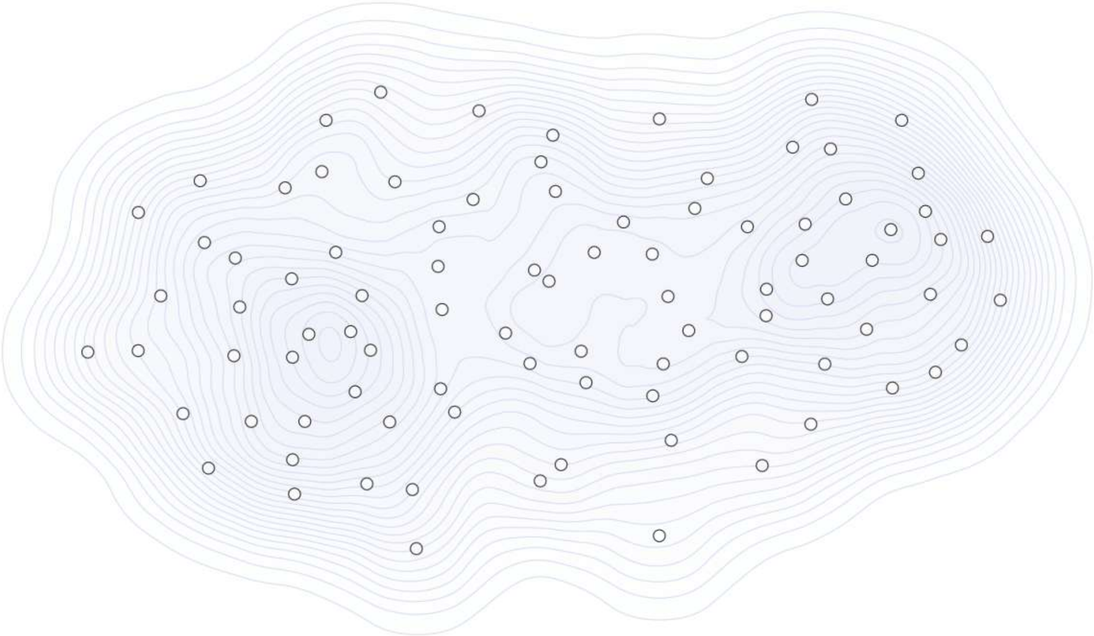


Figure 4.15: UMAP dimensionality reduction is applied to the data. It focuses on both the global and local structure of the data.

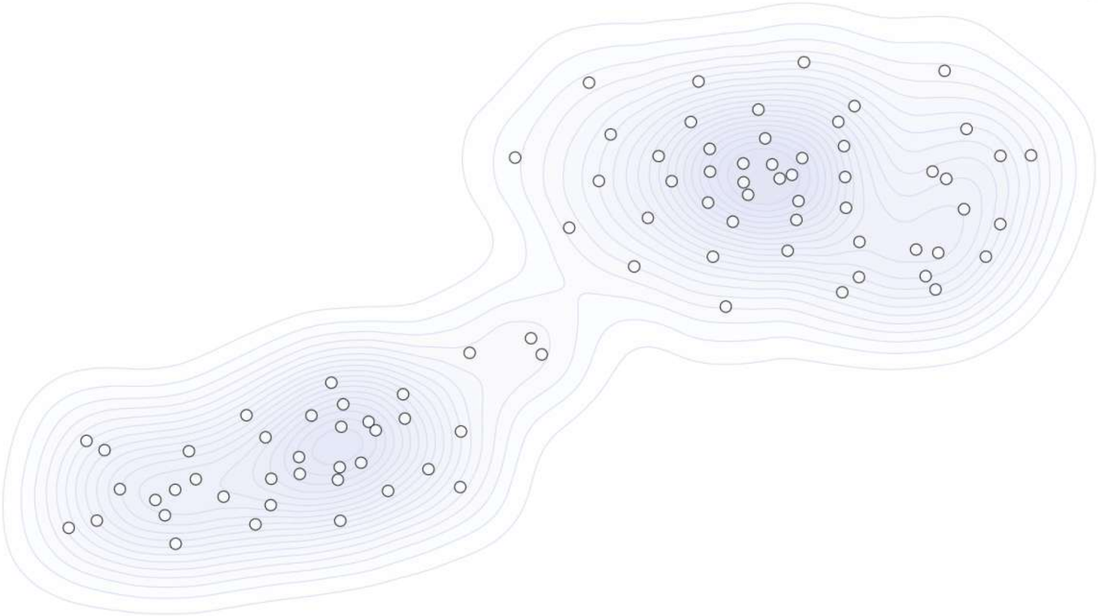


Figure 4.16: The t-SNE dimensionality reduction algorithm reveals two subclusters in the reduced space.

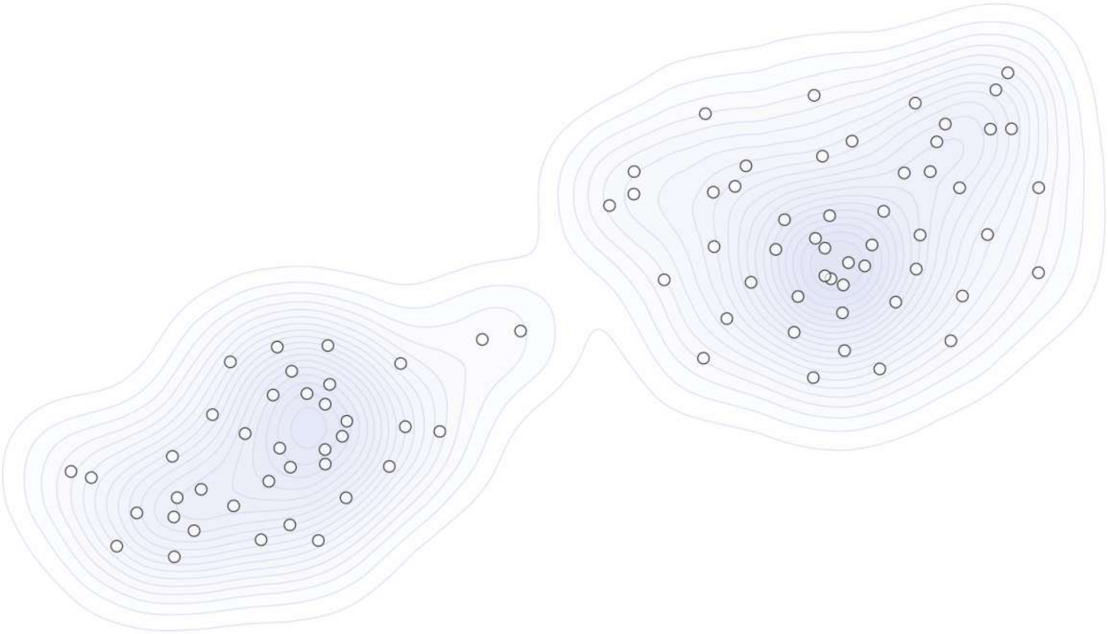


Figure 4.17: MICE imputation in combination with the t-SNE dimensionality reduction algorithm. Two subclusters are formed in the reduced space.

4.3.2 Clustering (T1b)

Why: ***produce*** clusters of patients

How: ***derive*** data divided into groups

Input: two-dimensional data from subtask **(T1a)**

Output: labels to assign each patient to a cluster

Clustering summarizes the data based on similarities and improves the data understanding [SEK03]. It supports the analysis of high-dimensional data and helps users to get insights into the data structure, such as biological processes of genomic data [KBZ+21]. To overcome the course of dimensionality, we apply clustering on the reduced two-dimensional data [SEK03, KBZ+21].

We tested and compared the following six clustering methods:

- k-Means (centroid-based)
- mean-shift (centroid-based)
- hierarchical clustering (hierarchical)
- DBSCAN: Density-based spatial clustering of applications with noise (density)
- OPTICS: Ordering points to identify the clustering structure (density-based)
- GMM: Gaussian mixture models (distribution-based)

Clustering algorithms can be categorized as hierarchical, centroid-based, distribution-based, or density-based [KBZ+21, XT15]. Centroid-based algorithms, such as k-means or mean-shift, are efficient but sensitive to initial conditions and outliers. However, they lead to the best results on our data and are therefore chosen as the default option. We determine and set the number of clusters using the elbow method [MHHWM18]. Density-based methods, such as DBSCAN or OPTICS, are suitable for arbitrary-shaped distributions. However, they cannot deal with high-dimensional data and varying densities. They also do not assign outliers to clusters. Distribution-based approaches, such as GMM, are suitable for Gaussian-distributed data. Hierarchical clustering fits datasets with a hierarchical structure the best. However, they are time-demanding and sensitive to their parameterization, such as the linkage criterion that must be defined [RG19]. The presented algorithms cluster the data based on their distribution [XT15]. An enhancement are deep learning-based approaches that consider the representation learned on the data in addition to the data distribution [KBZ+21]. However, their evaluation is hindered due to the limited amount of labeled data [KBZ+21].

In Subsection 5.2.2, we show the clustering results and evaluation of the cluster separation. Centroid-based algorithms, in our case, k-means and mean-shift, lead both to the same clustering result with the highest cluster separation scores. They detect the two visual clusters on the t-SNE dimensionality reduction result, which we determine as the default option in subtask **(T1a)**. Figure 4.18 illustrates the clustering result of both algorithms.

In k-means, the number of clusters must be specified in contrast to mean-shift. However, in mean-shift, the bandwidth parameter must be defined. We set k-means as the default option as mean-shift is computationally expensive, especially on a large number of data samples [HCL+19, ZZ21]. While mean-shift has a quadratic time complexity, the complexity of the k-means algorithm is linear [HCL+19, ZZ21]. In Figure 4.19, we demonstrate the result of the k-means algorithm on the MDS reduced space, where it assigns the data points to two clusters. The different clustering algorithms can be explored through the interface on demand. Combined with the forward and backward analysis tasks, they can be applied to the complete dataset or any subset of patients or features.

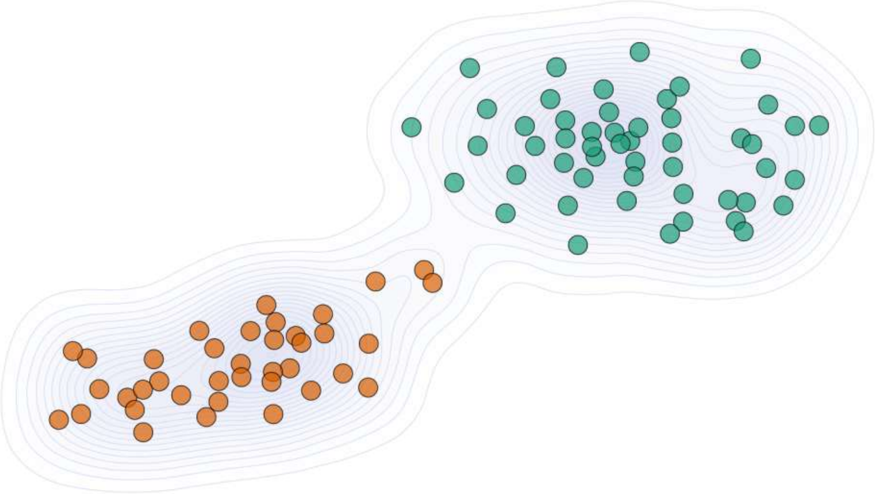


Figure 4.18: Clustering through the k-means or mean-shift algorithm on the data reduced through the t-SNE dimensionality reduction method.

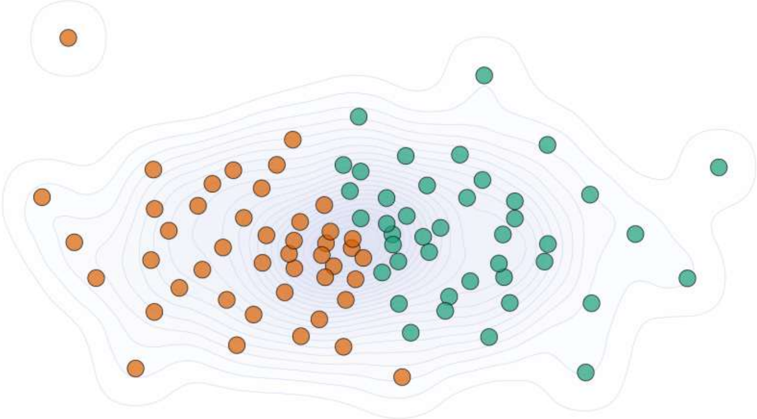


Figure 4.19: Clustering the MDS reduced space through the k-means algorithm.

4.3.3 Visualization (T1c)

Why: ***present*** the identified clusters of **(T1b)** in a visual representation

How: ***encode*** cluster similarities and differences as a visual representation

Input: two-dimensional data from subtask **(T1a)**, clusters from subtask **(T1b)**

Output: visual cluster representation

We visualize the reduced and clustered data through a scatterplot to visually present the patterns identified in the data as shown in Figure 4.18 and in Figure 4.19. Furthermore, we highlight the density of points through density contours and shading as a visual feedback on dense areas on the plot and the separation of clusters. Visualizing the high-dimensional data through a scatter matrix or a heatmap matrix of all feature correlations instead would capture only pairwise relations and does not identify complex patterns in the data. In Figure 4.20 we visualize a subset of the clincal features through a scatterplot matrix for demonstration purposes. We select one point on the scatter matrix and highlight this point in all views by applying transparency on the not selected points. However, the number of our features is larger than 10400, which hinders the data exploration in these visual encodings. They do not allow efficient pattern recognition and affect the performance of the application.

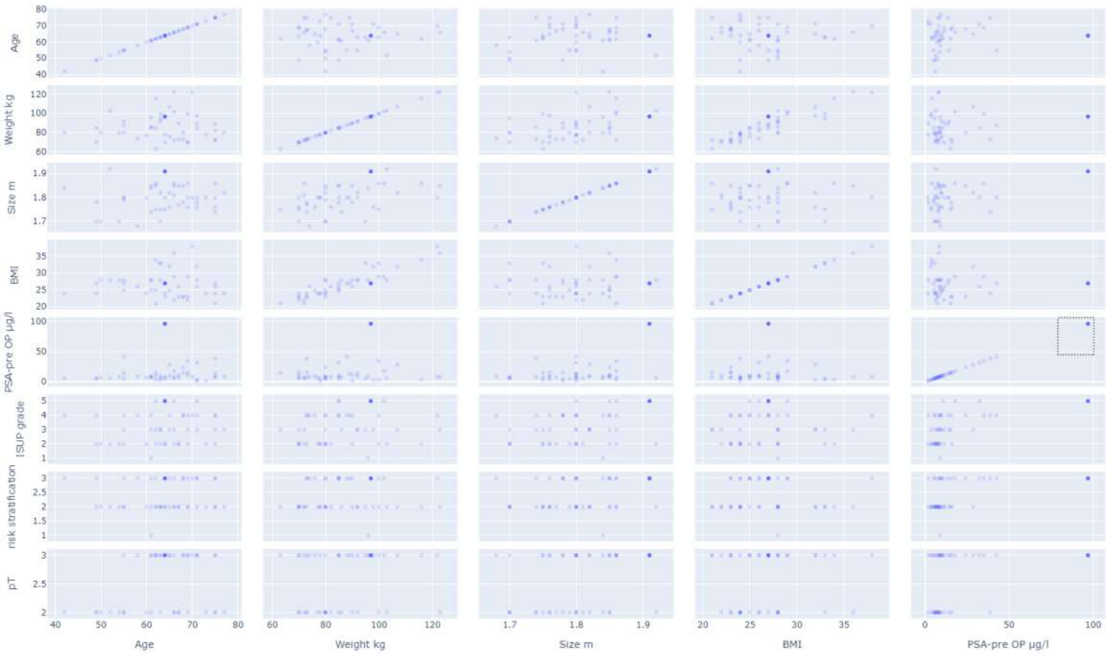


Figure 4.20: Pairwise correlations of a clinical feature subset.

4.4 Forward Analysis – Free Exploration (T2)

Why: ***explore*** the data selected in a visualization
discover patterns and correlations in the data visualization
present the selected data in a visual representation
lookup and identify data points selected in the visualization
identify data with top characteristics or differences

How: ***select*** patterns of interest in the visualization
encode the selected data in a visual representation
navigate in the visualization through zooming or by showing details on demand
arrange the data by ranking scores

Input: visualization of clusters from subtask **(T1c)**

Output: visualization of the selected data subset, its distributions, and the top features and gene mutations

After identifying and visualizing data patterns in the cohort stratification, we allow users to interact with these patterns through patient selections or subset processing and comparisons to explore and understand the data for knowledge discovery. This task further encourages hypothesis generation by identifying features that differentiate or characterize the clustered data. We highlight the density of the scatter points through a KDE plot of contours on the scatterplot. Figure 4.21 shows an enlarged view of these contours for the t-SNE reduced space. The contours reveal a subgrouping in the green points. In Figure 4.22, we demonstrate the KDE contours for FAMD. They have a detailed structure and reveal groupings in subsets of the orange data points. Moreover, we show details on the patients or feature distributions on demand by using tooltips.

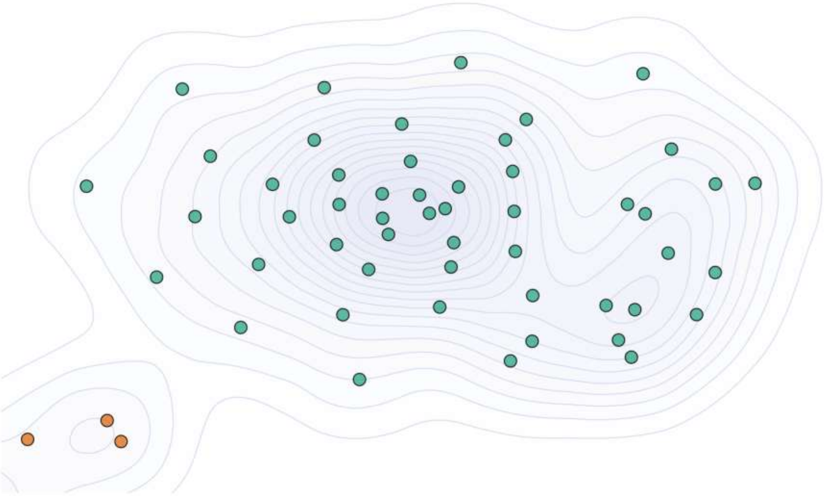


Figure 4.21: KDE contours on the t-SNE reduced space.

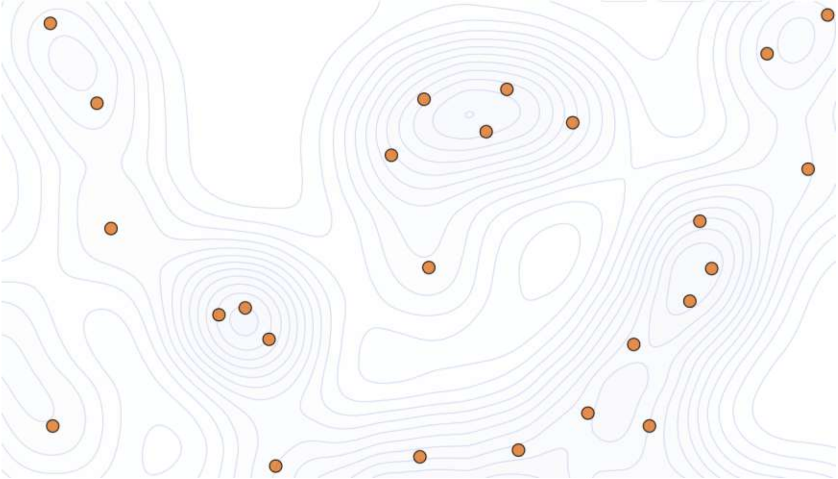


Figure 4.22: KDE contours on the FAMD reduced space.

4.4.1 Analysis of Patient Stratification (T2a)

Why: ***identify*** and explore similarities and differences of patient groups

How: ***annotate*** patient groups with similar or different features

Input: data visualization from subtask **(T1c)**

Output: data visualization with annotation of groups with similar or different features

We analyze the patient stratification to explain why the identified clusters are similar or different. By applying Shapley Additive Explanations (Shap) to the clustering result [LL17], we predict features that impact the clustering the most. We determine the top features of each cluster and combine them in the heatmap to provide users with an overview of these features as shown in Figure 4.23. On demand, we offer a detailed bar chart view that allows users to filter features by the radiomic, genomic, or clinical data, as depicted in Figure 4.24. Moreover, users can select any of these features through the heatmap or bar chart to highlight their values on the scatterplot, as illustrated in Figure 4.25. We implement Linear Discriminant Analysis (LDA) and Stochastic Gradient Descent (SGD) to pairwise determine the differentiating features between the identified clusters [OH21]. LDA fits a Gaussian density to each class, while SGD fits a linear Support Vector Machine (SVM). Both are linear methods, but SGD works with data represented as dense or sparse, which matches our sparse genomic or dense radiomic data. SGD further reveals more significant differences for our default preset of t-SNE, especially for genomic data. Therefore, we set SGD as the default method and allow users to explore LDA on demand. We sort the data by their relevance through the predicted impact on the clustering and combine the top 0.5% of the resulting feature subset in a heatmap to provide a visual and compact overview of the characteristics and differences of the identified clusters, as depicted in Figure 4.23. The higher the value is, the higher is its relevance in characterizing or differentiating clusters. The red values in the first two lines in Figure 4.23 represent the characteristics of the green or orange cluster. The red values in the third line of the heatmap represent the differences between both clusters. This allows the user to get an overview of all relevant features and compare their impact through all clusters and their pairwise differences. We use an exponential color scale and normalize all values in the range -1 to 1 to make them comparable and emphasize the diverging values at the beginning and end of the ranges. Using a linear or logarithmic color scale instead, leads to almost uniform colors on the heatmap that do not clearly depict the relevance of features for a cluster.

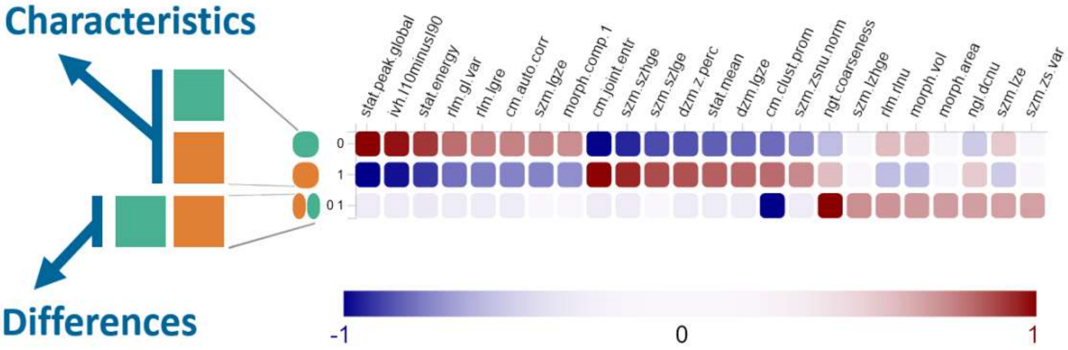


Figure 4.23: Heatmap view on characteristics and differences of a cluster.

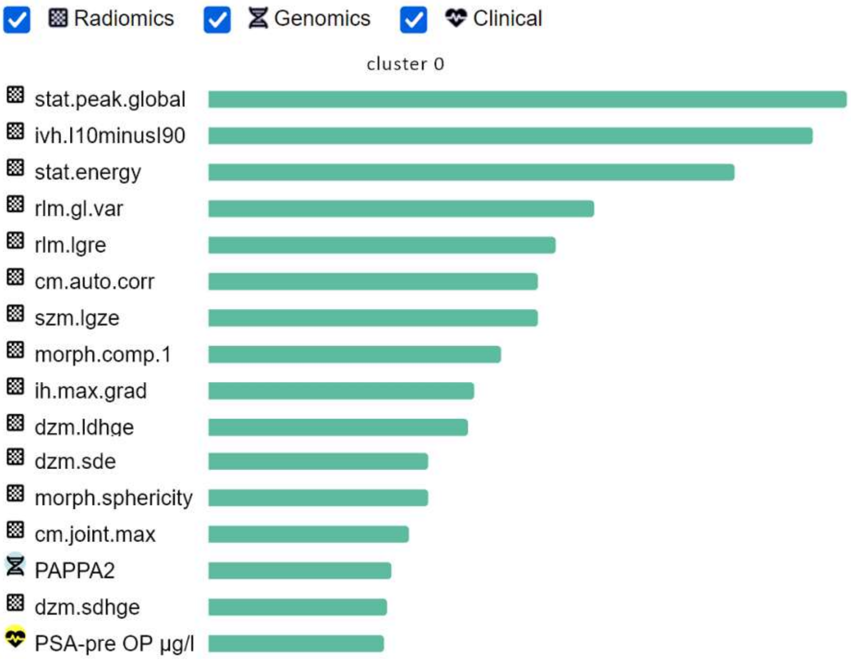


Figure 4.24: Characterizing features of the green cluster sorted by their relevance.

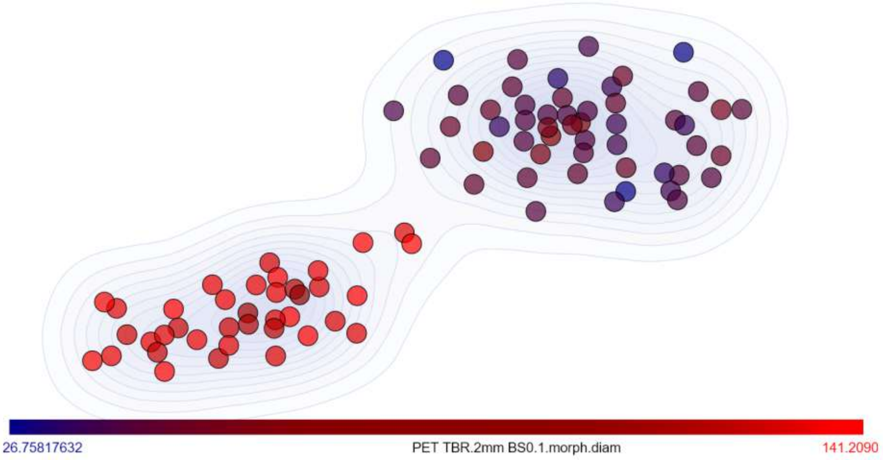


Figure 4.25: Radiomic feature values of the morphological diameter highlighted on the scatterplot. This reveals that values on the top right cluster are lower than the values on the bottom left cluster.

4.4.2 Data Selection on the Visualization (T2b)

Why: ***explore*** the data selected, discover patterns in the selection on the visualization ***present*** the selected data

How: ***select*** data in the visualization through a lasso-selection (or through a hypothesis as part of the backward analysis)

Input: data visualization from subtask **(T1c)**

Output: visualization of the selected data subset

We allow users to select any patient subset on the scatterplot through a lasso selection [BC87]. Compared to a rectangular selection, a lasso selection provides more flexible selection shapes that are not necessarily connected or neighbored on the scatterplot.

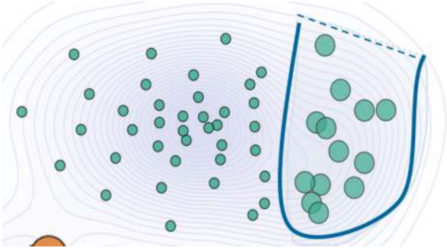


Figure 4.26: Lasso selection on the scatterplot.

4.4.3 Most Expressed Gene Mutations (T2c)

Why: ***identify*** data with top genes, and discover patterns in the visualization

How: ***arrange*** genes by their occurrence, filter the genes, and encode them visually

Input: matrix of genes and the persons that have them

Output: visual representation of the top genes in the interface

Our domain experts are interested in understanding the gene mutation data and identifying relevant gene mutations for the patients. Therefore, we determine and show the top gene mutations of any active patient subset selection on the scatterplot, as shown in Figure 4.27. The subset can be selected either manually through a lasso selection on the scatterplot or through a hypothesis-based selection as part of the backward analysis.

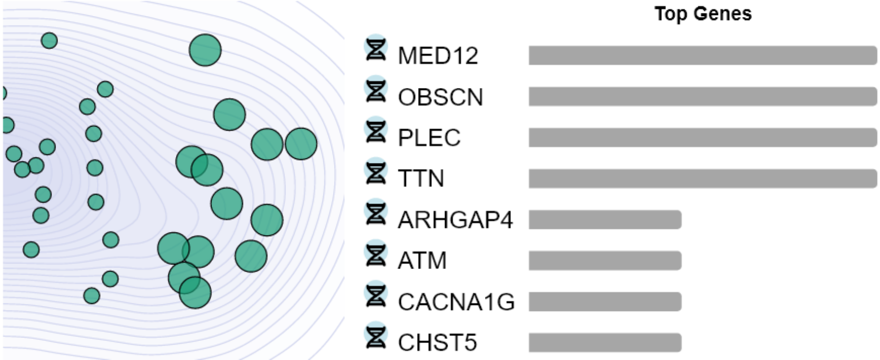


Figure 4.27: Top gene mutations of an active selection on the scatterplot.

4.4.4 Navigation in the Visualization and Details on Demand (T2d)

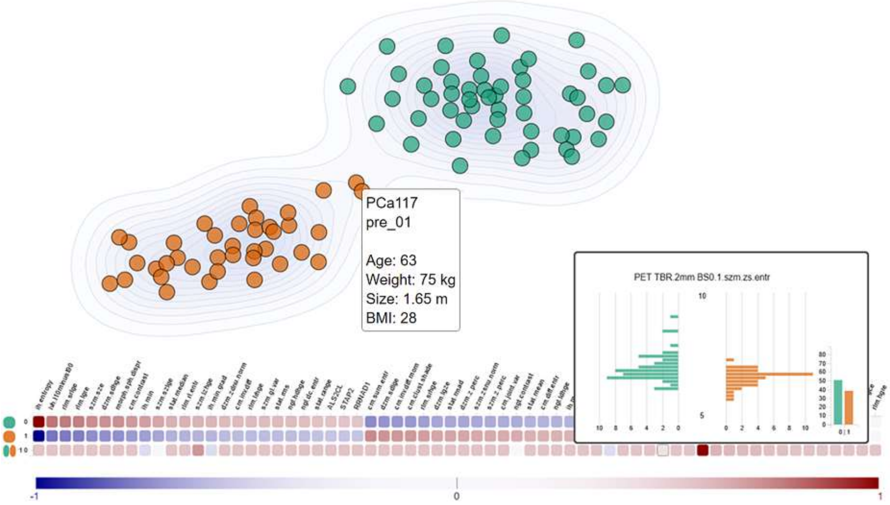
Why: ***lookup*** details on demand to discover patterns in the data (distribution, additional information on data points, number of gene mutations)

How: ***navigate*** in the visualization through zooming or by showing details on demand e.g., by hovering

Input: data visualization from subtask **(T1c)**

Output: visualization with details on demand and navigation possibilities

Users can navigate the visualization by zooming in or out on the scatterplot. We further show patient distributions on the heatmap on demand through a tooltip and display patient scores when moving the mouse over a patient in the scatterplot, as demonstrated in Figure 4.28.



PCa117
pre_01

Age: 63
Weight: 75 kg
Size: 1.65 m
BMI: 28

Figure 4.28: Tooltip on the heatmap show patient distributions of the respective heatmap feature. On the scatterplot, the tooltip reveals clinical scores of a patient.

4.4.5 Presets Selection and Parameter Change (T2e)

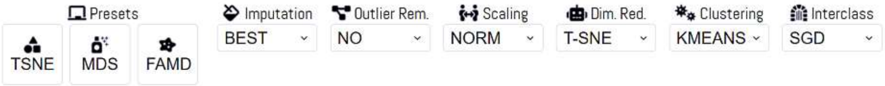
Why: ***explore*** the data of the option selected, discover patterns in the data

How: ***select*** predefined options through the interface (dropdowns, radio buttons)

Input: data visualization from subtask **(T2b)**
reduced data from subtask **(T1a)**
imputed values from subtask **(T0c)**
data without outliers from subtask **(T0d)**

Output: data of the selected options shown in a visualization

Users are provided presets for the data analysis based on evaluating the analysis and cohort stratification options and identifying the most suitable parameters for the underlying datasets. Furthermore, users can manually change any analysis option through the interface to explore how this affects the revealed patterns in the data. We show the presets and default options in Figure 4.29.



Imputation: BEST
Outlier Rem.: NO
Scaling: NORM
Dim. Red.: T-SNE
Clustering: KMEANS
Interclass: SGD

Figure 4.29: Presets and data analysis options are shown on demand for an extended analysis and exploration of the data.

4.5 Backward Analysis – Hypothesis-based Exploration (T3)

Why: ***discover*** whether a hypothesis is confirmed or rejected
lookup features of interest (e.g., genes, radiomic features, clinical data)
identify the selected subset based on a hypothesis
explore features of a selected cluster or patient
compare the selected and filtered data points and features with each other

How: ***select*** or filter the radiomic, genomic, and clinical data based on a hypothesis
encode the selected or filtered data in a visual representation
arrange features to select based on their distribution or alphabetic order

Input: data visualization from subtask **(T1c)** and a user-defined hypothesis

Output: Filtered visualization based on a hypothesis

In the backward analysis, we allow users to assess the correctness of a hypothesis by filtering the data based on a hypothesis in mind. Users can interactively filter, select, and compare data subsets [Shn94]. This enables users to identify thresholds for hypothesis assessment or to determine a new hypothesis for the underlying data, as illustrated in Figure 4.32. The resulting subset of data points is selected on the scatterplot and can be used for further forward or backward data exploration.

4.5.1 Radiomic, Genomic, and Clinical Data Filtering (T3a)

Why: ***discover*** whether a hypothesis is confirmed or rejected
lookup, ***identify***, and ***explore*** the desired data in a visualization

How: ***filter*** the data based on a hypothesis (below, above threshold, define range)
arrange features to select based on their distribution or alphabetic order

Input: data visualization from subtask **(T1c)** and a user-defined hypothesis

Output: filtered visualization based on a hypothesis

This task aims to ***discover*** whether a hypothesis is confirmed or rejected for the underlying data. We want to ***lookup*** and ***identify*** the features of interest to highlight them in the visual representation that the user can ***explore***. To fulfill this task, we ***filter*** and ***arrange*** the features matching a partial input of the feature name by alphabetic order and ***filter*** the data based on the selected features and the defined ranges [Shn94]. The backward analysis allows users to assess the correctness of any hypothesis in mind on the visual representation.

4.5.2 Feature Subset for Visual Hypothesis Assessment (T3b)

Why: ***discover*** whether a hypothesis is confirmed or rejected based on the cohort stratification of a feature subset

How: ***select*** feature subset and visualize the outcome of a hypothesis on it
arrange features to select based on their distribution or alphabetic order
Example: analyze and visualize clinical, radiomic, or genomic data separately, or use any combination of their feature subset

Input: data visualization from subtask **(T1c)** and a user-defined hypothesis

Output: filtered visualization based on hypothesis

Based on the hypothesis defined in **(T3a)**, we allow users to ***explore*** the resulting patient subset and the feature distributions of this resulting subset. Users can apply the stratification on only the identified patient subset, as illustrated in Figure 4.30.

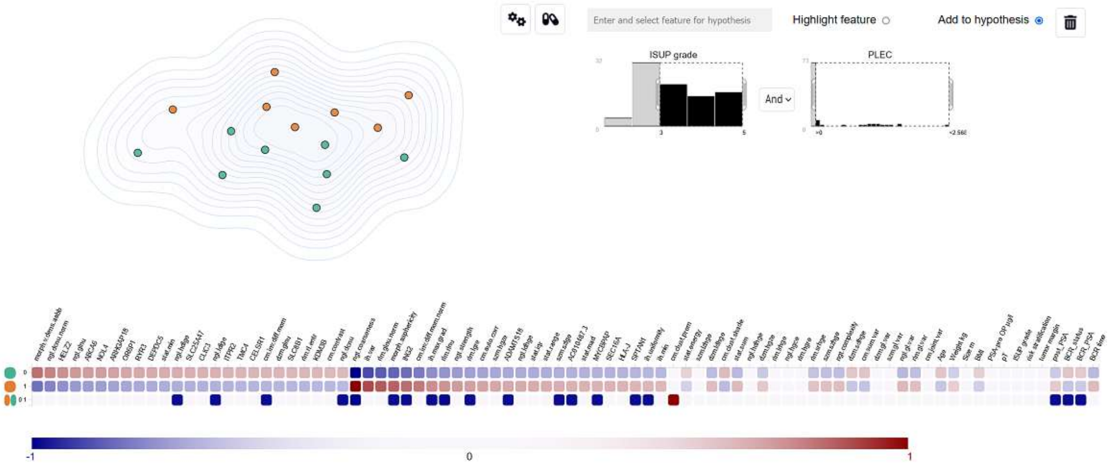


Figure 4.30: Apply cohort stratification on only the patient subset that fulfills the hypothesis. The resulting patients are shown on the scatterplot.

Furthermore, users can compare the characteristics and differences of the patient subset that fulfills the hypothesis with patients that do not fulfill the hypothesis, as shown in Figure 4.31. To accomplish this task, we automatically ***select*** patients matching the conditions specified through a hypothesis, and allow users to cluster the data based on the selected patients.

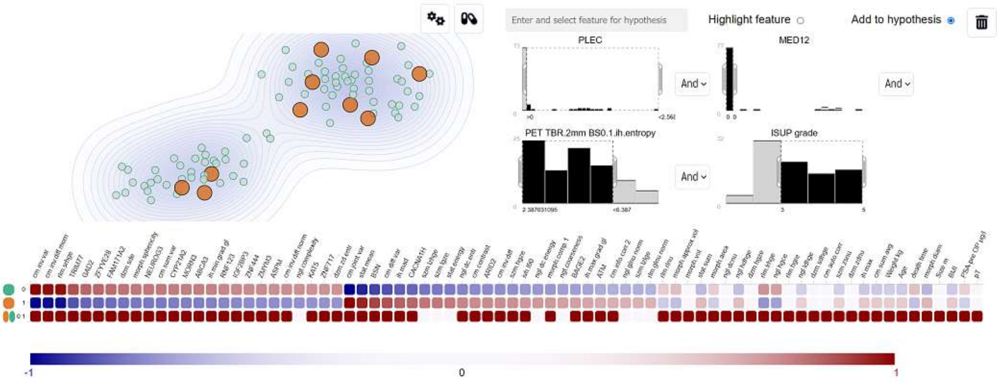

Enter and select feature for hypothesis
Highlight feature ○
Add to hypothesis ●
PLEC
MED12
And
PET TBR.2mm BS0.1.ih.entropy
ISUP grade
-1
0
1

Figure 4.31: Comparison of patients that fulfill the hypothesis (orange cluster) with patients that do not fulfill it (green cluster) by investigating the characteristics and differences of both groups on the heatmap on the bottom of the view.

4.5.3 Hypothesis-based Comparison of Patients and Features (T3c)

Why: ***compare*** patients selected through a hypothesis with not selected patients ***compare*** interactively how selected feature ranges affect the patient selection on the scatterplot

How: ***select*** data in a visualization through a hypothesis ***filter*** features, and update the differences and characteristics on the heatmap

Input: data visualization from subtask **(T1c)** and a user-defined hypothesis

Output: characteristics and differences of selected subset based on a hypothesis

To interactively ***compare*** the impact of changing feature thresholds on the visual representations and to refine a hypothesis, we allow users to interactively ***select*** thresholds for each hypothesis feature through sliders, as depicted in Figure 4.32. This enables users to ***explore*** how each change conveys to the resulting patient subset on the scatterplot.

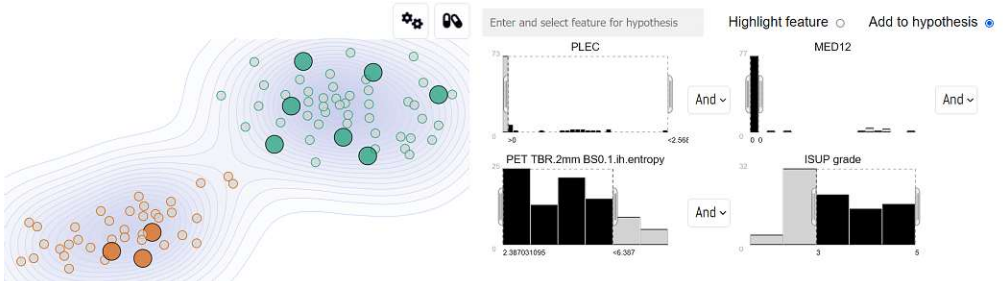


Figure 4.32: Hypothesis creation and refinement through interactively changing sliders.

4.6 Development Environment and Libraries

We implement our framework as a web application through Python and JavaScript. On the backend, we use Python due to its high machine learning and analysis capabilities through available libraries. This allows us to flexibly test, compare, and provide different analysis options on the data. For the frontend, we utilize JavaScript, which has a higher performance than Python and is therefore advantageous for interactive tasks. We use the D3 library of JavaScript to create the visualization and interaction components of our framework. It allows free customization of all visual aspects, such as the axis, legends, or the integration of additional visual elements and glyphs. The communication between the frontend and backend is performed through the Flask web framework of Python as HTTP requests. Figure 4.33 gives an overview of our workflow. We read, process, and analyze the data on the backend through Python. This includes all subtasks of the preprocessing (T0) and cohort stratification (T1) tasks. Then, we transfer the analysis results to the frontend through the Flask web framework, where we implement the HTML page structure and content, the CSS page layout and design, as well as the Javascript/D3 visualization and interaction components. Therefore, the forward analysis (T1) and the backward analysis (T2) tasks are performed on the frontend.

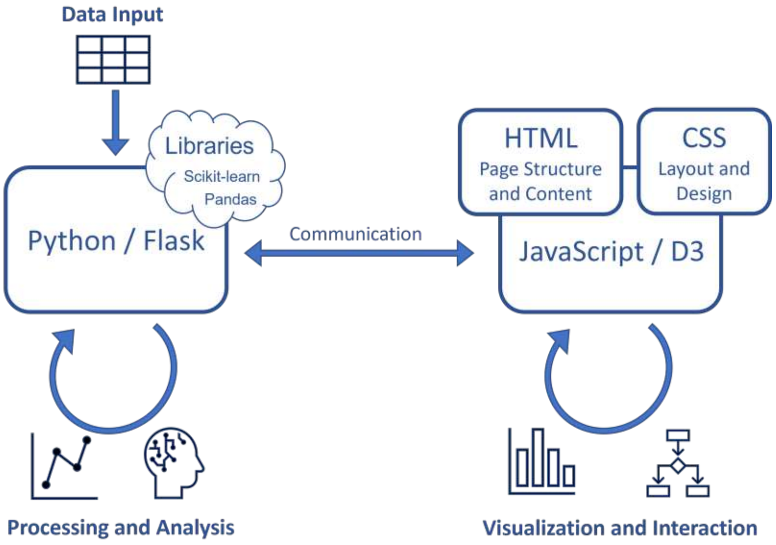


Figure 4.33: Simplified illustration of our implementation structure. We read, process, and analyze the data on the backend through Python. Then, we transfer the analysis results to the frontend through the Flask web framework, where we implement the HTML page structure and content, the CSS page layout and design, as well as the Javascript/D3 visualization and interaction components.

For the machine learning part, we mainly use the *Scikit-learn* library for the data encoding, imputation, outlier detection, dimensionality reduction, clustering, and prediction of the characterizing and differentiating features of the identified clusters. Additionally, we test the *Impyute*, *AutomImpute*, *MiceForest*, and *FancyImpute* libraries for extended imputation capabilities. We utilize *Pandas* and *Numpy* for the data management and computations on the data. Further, we use the *Chardet* library for character encoding. 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 integrating density-based contours on the scatterplot and *d3-tip* to show the heatmap tooltip of the distribution plots on demand. We perform the data filtering, sorting, and interaction components on the frontend to allow users to interactively receive feedback on the visualization. If users desire to apply the cohort stratification on a subset of patients or features, we do this on the backend, which requires a waiting time of up to 7 seconds depending on the concrete task and processing steps required. During this time, we show a waiting indicator in the framework to visually inform the user about the ongoing processing of the data. We implement predicting the characterizing and differentiating clustering features on the backend and therefore update the heatmap during interactive tasks only on demand to support the interactive exploration of the data.

4.7 Overview on our Framework Components

We divide the interface into three main views and show tooltips for the heatmap and scatterplot points on demand to reveal additional information on the data without cluttering the view and with a clear relation to the respective feature or patient. Furthermore, we integrate five tabs to visualize feature values, the top clustering features, or the most expressed gene mutations. Our tabs also consist of a processing view to specify data subsets for the analysis or a hypothesis view to visually assess or create hypotheses on the data by combining features and ranges of interest to filter the data interactively. Figure 4.34 gives an overview of our framework and its interactive views, while Figure 4.35 summarizes the functionality of the five tabs (Figure 4.34, B) of our framework.

Component (A) – (T0–T3) We show a scatterplot view with the result of the cohort stratification, where each patient represents one scatter point. The color of the points indicates the cluster the patient is assigned to and matches the color 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 in to the selection, to process and stratify only the points of this selection, or to set an active selection as its own cluster to investigate its features in relation to the not selected points through the heatmap (C) or in the **Clusters** and **Top gene** views (B).

Component (B) – (T2–T3) This view 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 4.35.

Values (T2): This default view allows users to see the distributions of patient values for an active selection on the scatterplot. The values are grouped per cluster, which serves as initial feedback on the data and the clustering scores.

Clusters (T2): By changing to this view, users are presented a ranked list of features that pairwise differentiate between two clusters. This serves as a detailed view of the heatmap features that also allows filtering the features regarding their radiomic, genomic, or clinical data. Users can also select to show the characteristics of one of the identified clusters and explore these features further.

Top genes (T2): This view shows a ranked list of the top gene mutations based on the number of patients that have them. It is displayed by default for the complete dataset but can be filtered through any selection made on the scatterplot.

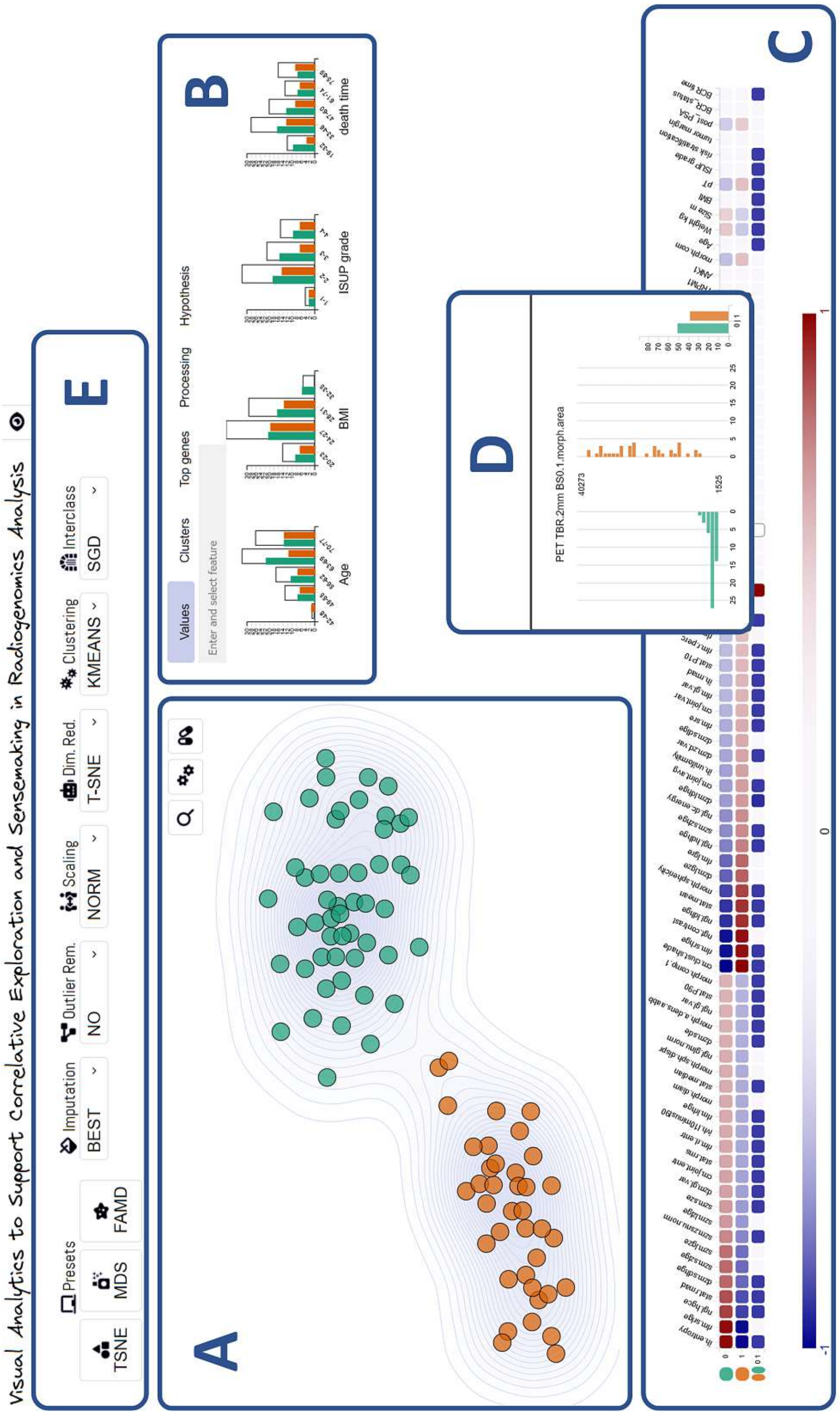
Processing (T3): The processing view allows the selection of any feature subset of the radiomic, genomic, or clinical data for the analysis process. By default, all features of all datasets are integrated that can be filtered on demand.

Hypothesis (T3): Users can highlight features of interest on the scatterplot or combine features to a hypothesis and assess its correctness for the underlying data. The feature ranges can be interactively and visually defined through sliders. We allow the combination of multiple features through a local *and* or a logical *or* operation. The resulting subset leads to a hypothesis-based selection on the scatterplot that can be explored further by investigating the top genes or features of the selection.

Component (C) – (T2–T3) The heatmap gives an overview of the features characterizing and differentiating the clusters. It consists of normalized values from -1 to 1 colored through a scale ranging from blue to red to make the ranking scores comparable and emphasize the diverging values around the beginning and end values of the ranges. The higher the value is, the higher is its predicted impact on the current clustering on the scatterplot.

Component (D) – (T2–T3) On demand, we show 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. When more than two clusters are available, the distribution of the cluster depicted by the respective heatmap line is compared with the points outside this cluster.

Component (E) – (T2–T3) Advanced options are available 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. When an active selection is made on the scatterplot, it is kept while changing the advanced options or switching the view between tabs to allow users to investigate these points further and see where they are located after a parameter change. These also consist of three presets with predefined analysis options that enable users to reset the view back to the default option through the TSNE preset or explore different dimensionality reduction methods with predefined parameters.



Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis

Presets: TSNE, MDS, FAMD

Imputation: BEST

Outlier Rem.: NO

Scaling: NORM

Dim. Red.: T-SNE

Clustering: KMEANS

Interclass: SGD

Figure 4.34: Main views (A-E) of our visual radiogenomics analysis framework.

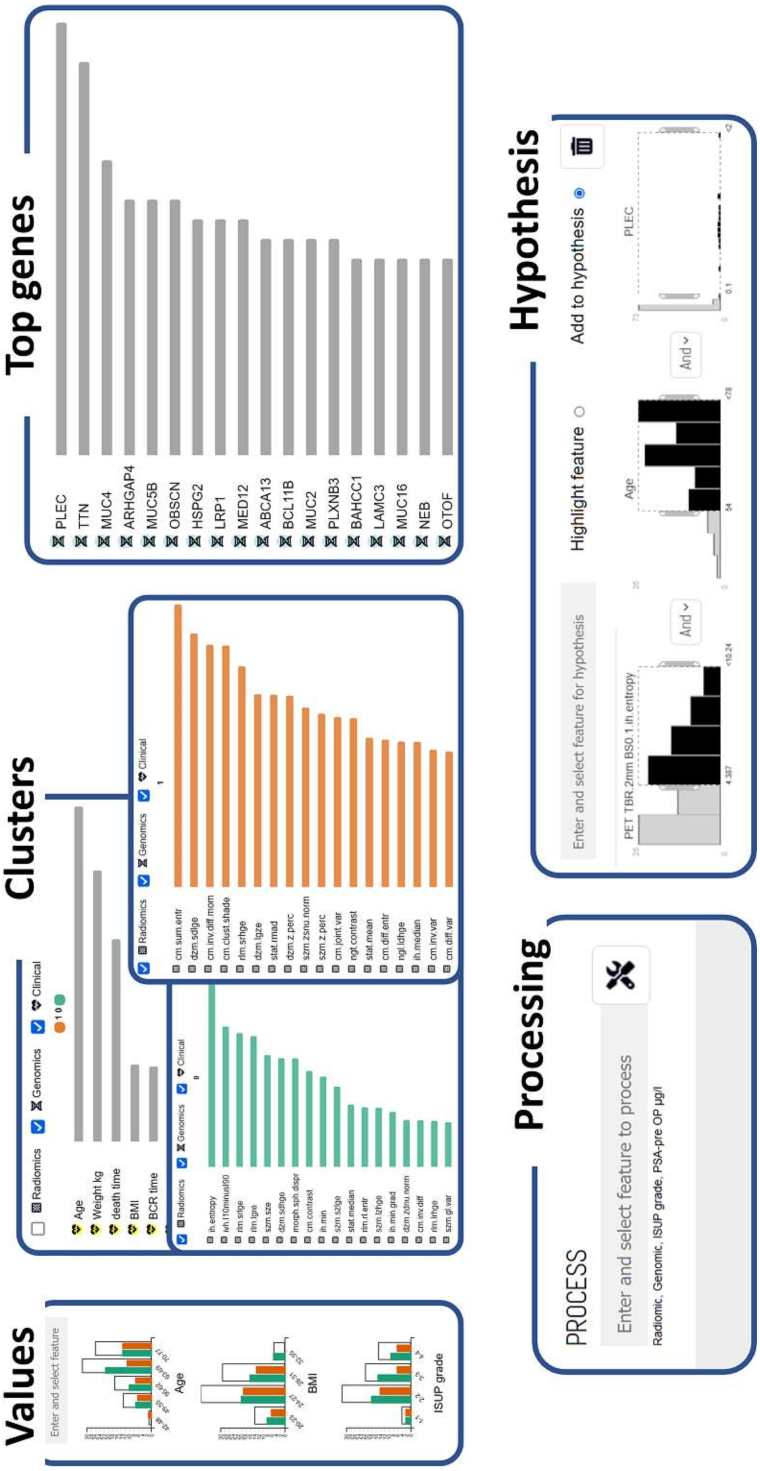
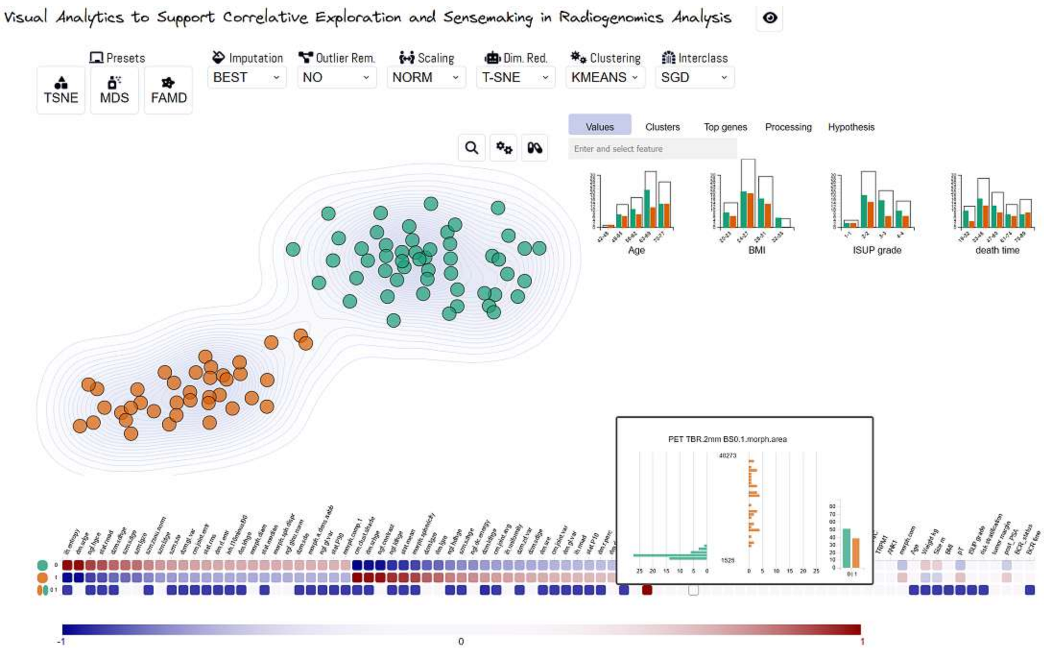


Figure 4.35: Detailed view of the five tabs of our framework. These represent component (B) of Figure 4.34.

CHAPTER 5

Results and Evaluation

This chapter presents our visual interactive and flexible framework for the combined radiogenomic and clinical data analysis. It allows domain experts to gain insights into the data for knowledge discovery and hypothesis assessment. Figure 5.1 shows our resulting framework.



Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis

Presets: TSNE, MDS, FAMD

Imputation: BEST

Outlier Rem.: NO

Scaling: NORM

Dim. Red.: T-SNE

Clustering: KMEANS

Interclass: SGD

Figure 5.1: Our visual radiogenomics framework.

We summarize evaluation techniques in Section 5.1 and present our quantitative feedback in Section 5.3 as well as our qualitative evaluation with domain experts in Section 5.3.

5.1 Evaluation Techniques

Evaluation is essential to understand whether a visualization tool achieves its goals [Mun09]. This process requires a thorough understanding of the visualization components and their complex processes [LBI+12]. Munzner et al. [Mun09] presents a model with four nested layers for the design and evaluation of visualizations, as shown in Figure 5.2. This model characterizes the domain problem and data (level 1), which are mapped into abstract operations and data types (level 2). The visual encoding and interaction techniques for the operations and data types are designed (level 3), and the algorithm for the visual encoding and interaction design is specified (level 4) [Mun09]. Figure 5.3 summarizes threats that could arise in each level of the nested model and their validation methods, from which a subset can be considered in evaluating visualization approaches [Mun09]. Examples of the proposed validations include the observation of target users, the justification of visual encodings and interaction design, the computation of time and complexity of algorithms, as well as qualitative or quantitative evaluation of the resulting analysis. An appropriate validation approach for the resulting analysis is the qualitative discussion through images [Mun09].

Lam et al. [LBI+12] present seven scenarios to evaluate information visualizations. These include, for example, the evaluation of visual data analysis, visualization algorithms, or user performance. Isenberg et al. [IIC+13] extend the evaluation scheme of Lam et al. [LBI+12] by *qualitative result inspections*. These consist of qualitative discussions and assessments of visualization results that address viewers. They comprise showing images of the end results together with a problem description and justification to clarify how the specified goal is met [IIC+13]. For frameworks that tackle data analysis, knowledge discovery, or knowledge management, Lam et al. [LBI+12] and Isenberg et al. [IIC+13] propose the evaluation through *Visual Data Analysis and Reasoning (VDAR)*. This category includes *usage scenarios* as a validation technique that describes how a hypothetical domain expert could use the visualization tool. When domain problems are handled through close collaboration between visualization researchers and domain experts, or reports are created on domain experts interacting with the visualization tool, it is considered as a *case study* instead [IIC+13].

As suggested by Lam et al. [LBI+12] and Isenberg et al. [IIC+13], we conducted the qualitative evaluation through VDAR as usage scenarios. Our usage scenario screenshots and descriptions can be further perceived as qualitative result inspection as they describe the problems and demonstrate their solutions within our tool through images and justifications [IIC+13]. We presented the tool in several iterations to our domain experts, who freely commented on it. Furthermore, our tool emerged through collaboration with domain experts. It was tested by one of our cancer experts, who interacted with the framework and described her thoughts on its functionality and usability, which we documented in parallel. Regarding the definition of Isenberg et al. [IIC+13], this is considered as a case study from domain experts and close collaborators. Our quantitative evaluation includes determining error scores for the imputation methods and clustering separation scores based on similarity measures.

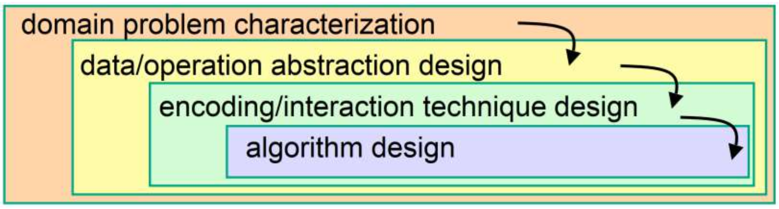

domain problem characterization
data/operation abstraction design
encoding/interaction technique design
algorithm design

Figure 5.2: Nested four-level model for visualization design and evaluation [Mun09].

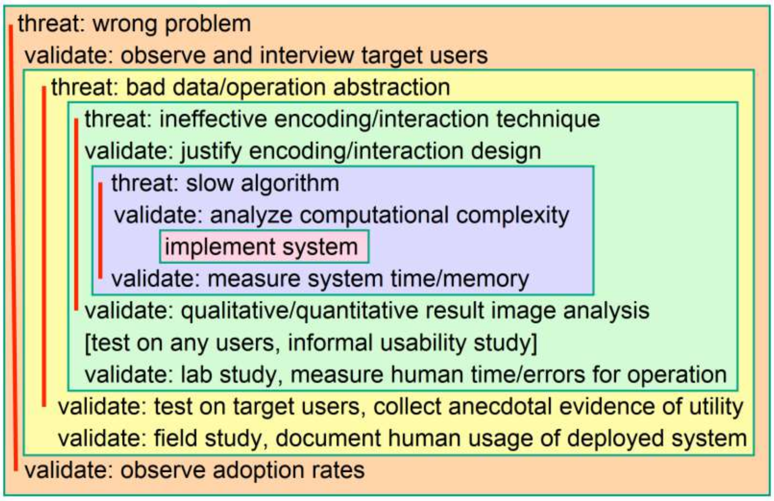

threat: wrong problem
validate: observe and interview target users
threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Figure 5.3: Threats and validation of the nested design and evaluation model [Mun09].

5.2 Quantitative Scores

We quantitatively evaluate our imputation approach through determining the MAPE and RMSE for different missingness percentages on the trainingset of our data, as presented in Subsection 5.2.1. Moreover, we evaluate the cluster separation through the Silhouette coefficient [Rou87], the Calinski-Harabasz index [CH74], and the Davies-Bouldin index [DB79]. In Sectiion 5.2.2, we show the evaluation results of the clustering.

5.2.1 Imputation Scores

For the data imputation, we divide the subset of patients with complete features into a training and test set. We randomly simulate missingness in the datasets and calculate the RMSE and MAPE of the result. We repeat this process five times with randomly simulated missingness percentages and data splitting to ensure its robustness and plot the mean error of these five iterations. Figure 5.4 – Figure 5.9 show the distribution of the features before imputation and with the three main options we provide through the interface. These options are BEST, MICE, and KNN. The BEST imputation applies the best suitable method on each feature considering the data types and RMSE evaluation. MICE and KNN are applied on all features with missingness when these options are selected. The RMSE behaves similarly for features of the same data type. Therefore, we demonstrate in this thesis only the RMSE for one representative feature of each data type on the training and test set. Figure 5.10 and Figure 5.11 show the RMSE for the binary *BCR status* feature on the training and test set. The y-axis depicts the RMSE, while the x-axis shows the missingness percentages simulated in the data. Figure 5.12 and Figure 5.13 depict the RMSE for the discrete *BCR time* feature. Figure 5.14 and Figure 5.15 illustrate the RMSE for the continuous *post PSA* feature. In general, these three methods are stable for our default t-SNE preset as they do not affect the clustering of the patients. Even imputing the missingness through a constant value leads to a deviation of only one patient that is assigned to a different cluster compared to using other imputation methods.

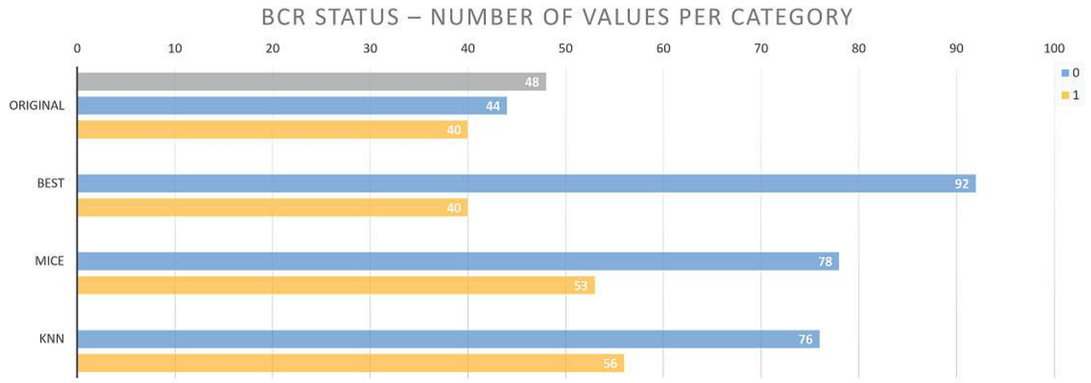


Figure 5.4: BCR status feature distribution before and after imputation.

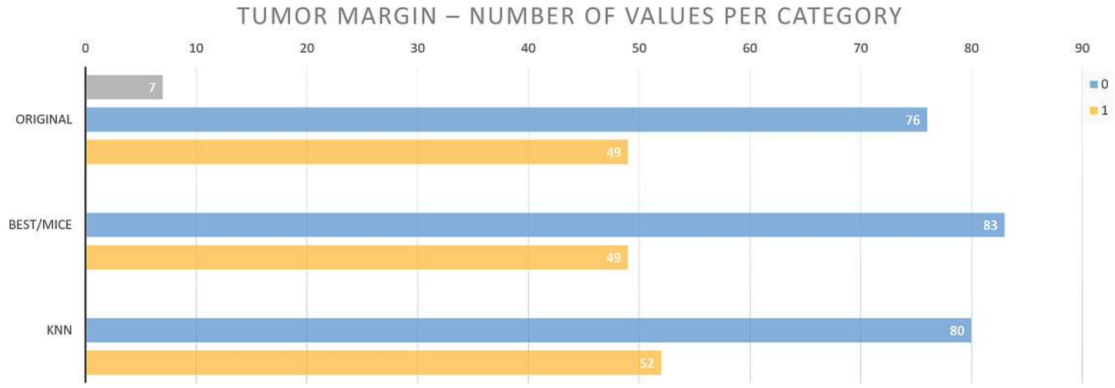


Figure 5.5: Tumor margin feature distribution before and after imputation.

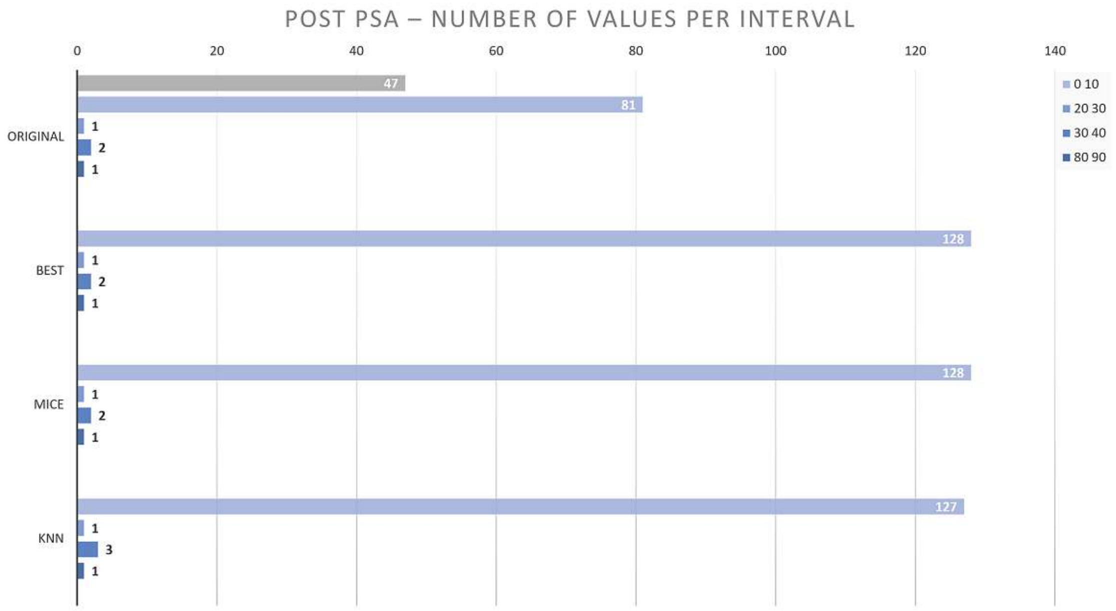


Figure 5.6: Post PSA feature distribution before and after imputation.

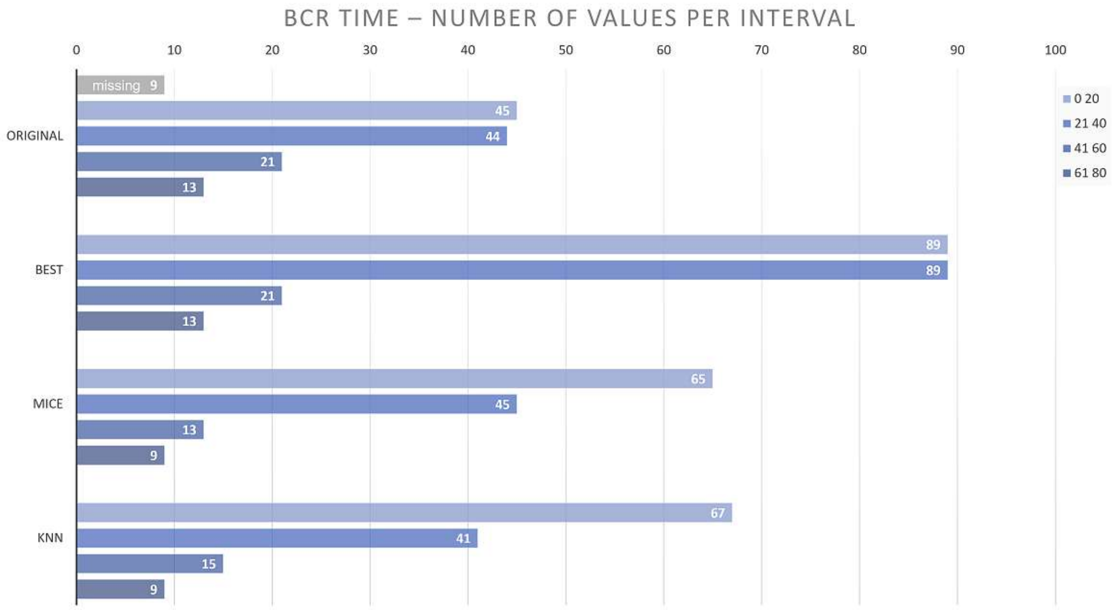


Figure 5.7: BCR time feature distribution before and after imputation.

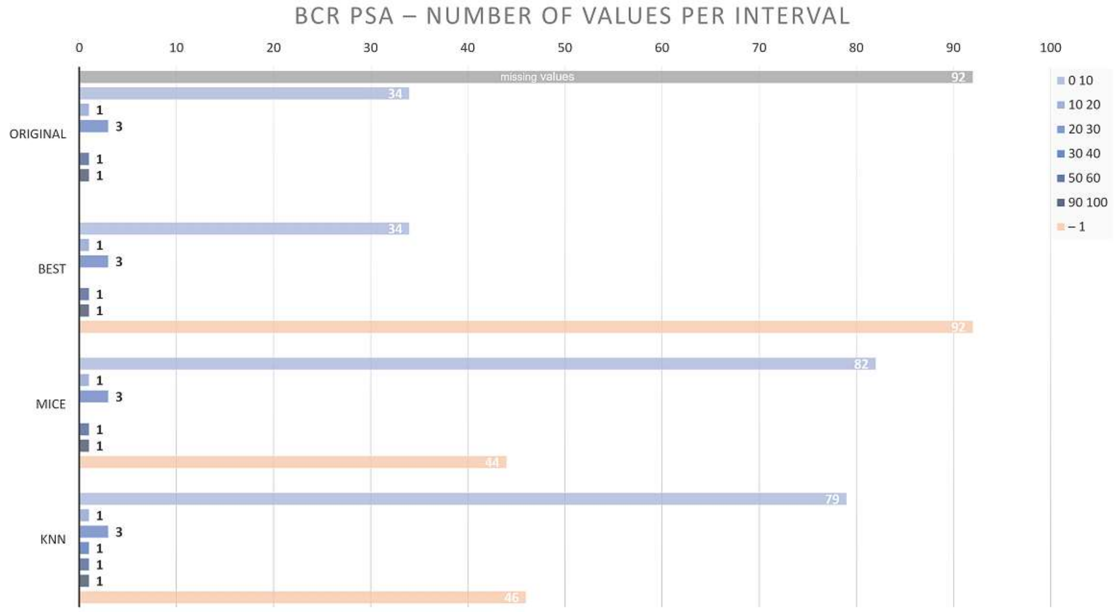


Figure 5.8: BCR PSA feature distribution before and after imputation.

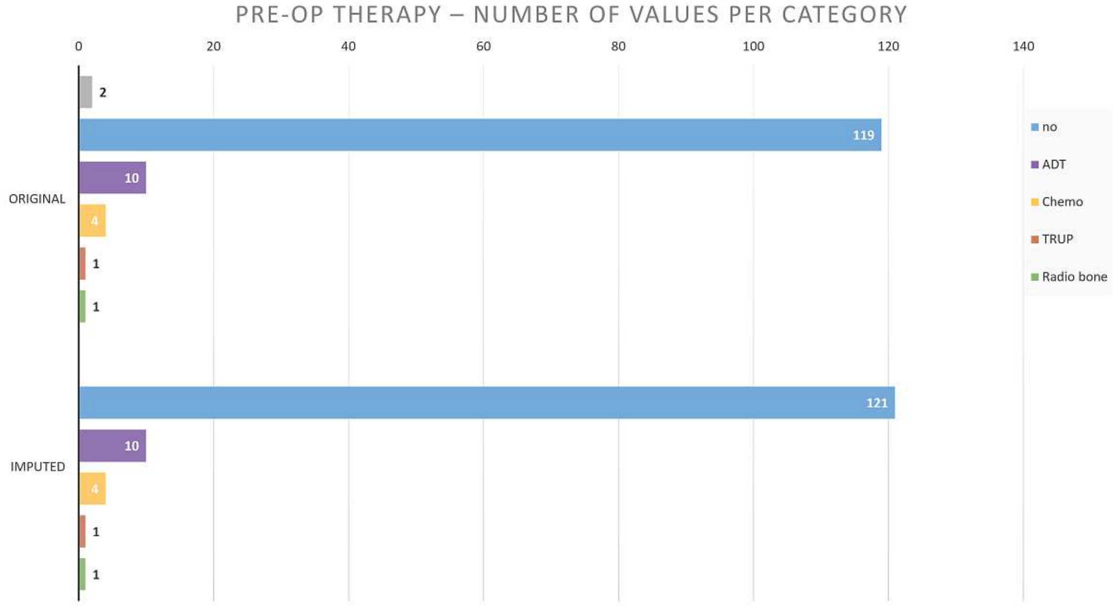


Figure 5.9: Pre-OP therapy feature distribution before and after imputation.

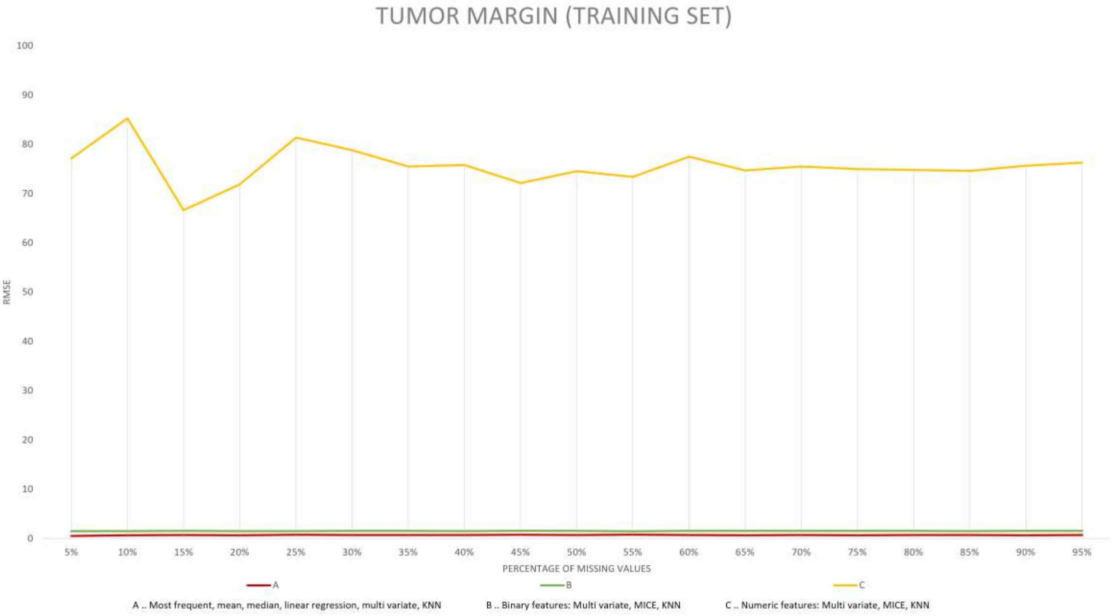


Figure 5.10: RMSE error for imputing a binary feature on the training set.

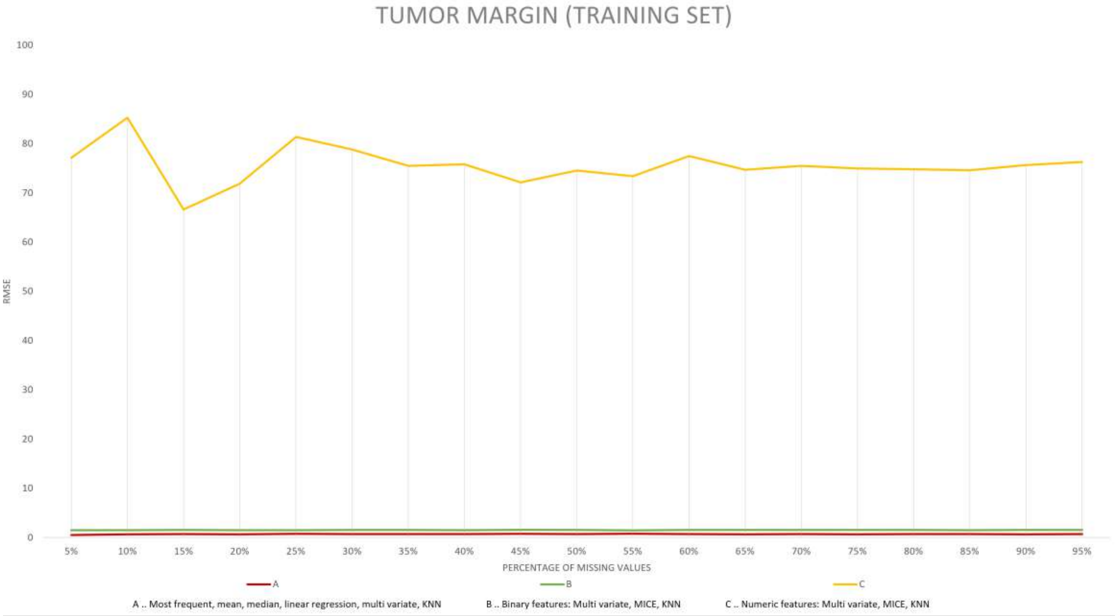


Figure 5.11: RMSE error for imputing a binary feature on the test set.

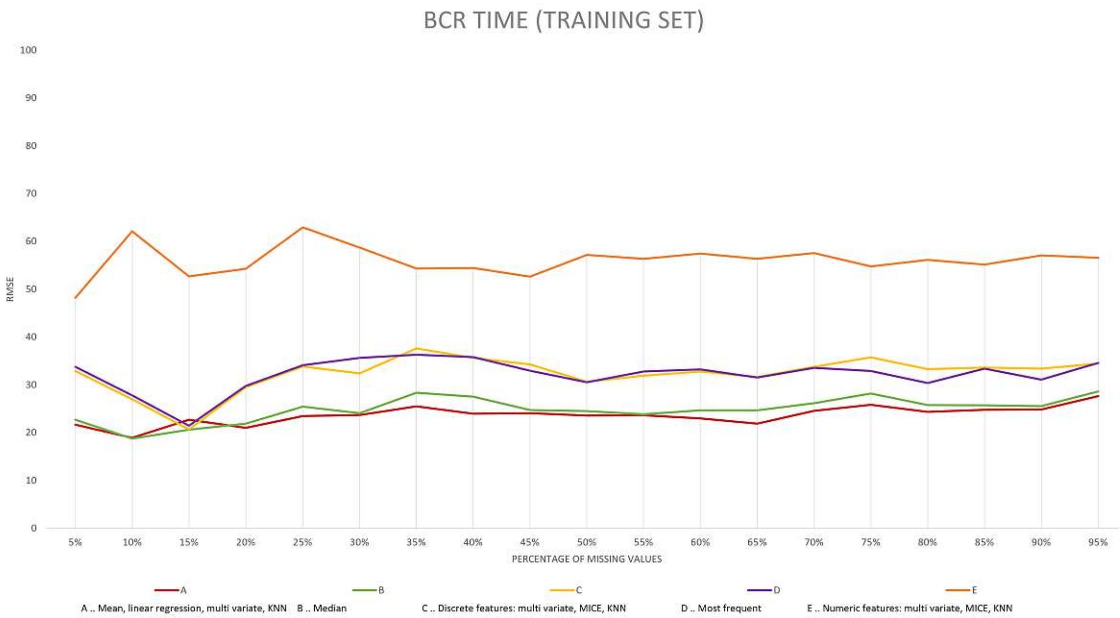


Figure 5.12: RMSE error for imputing a discrete feature on the training set.

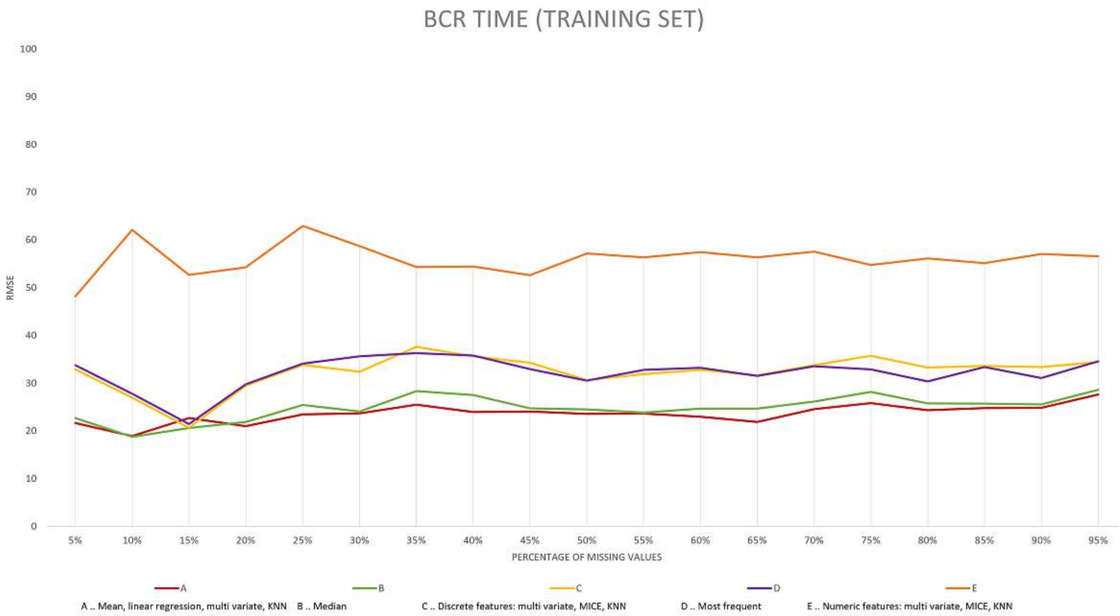


Figure 5.13: RMSE error for imputing a discrete feature on the test set.

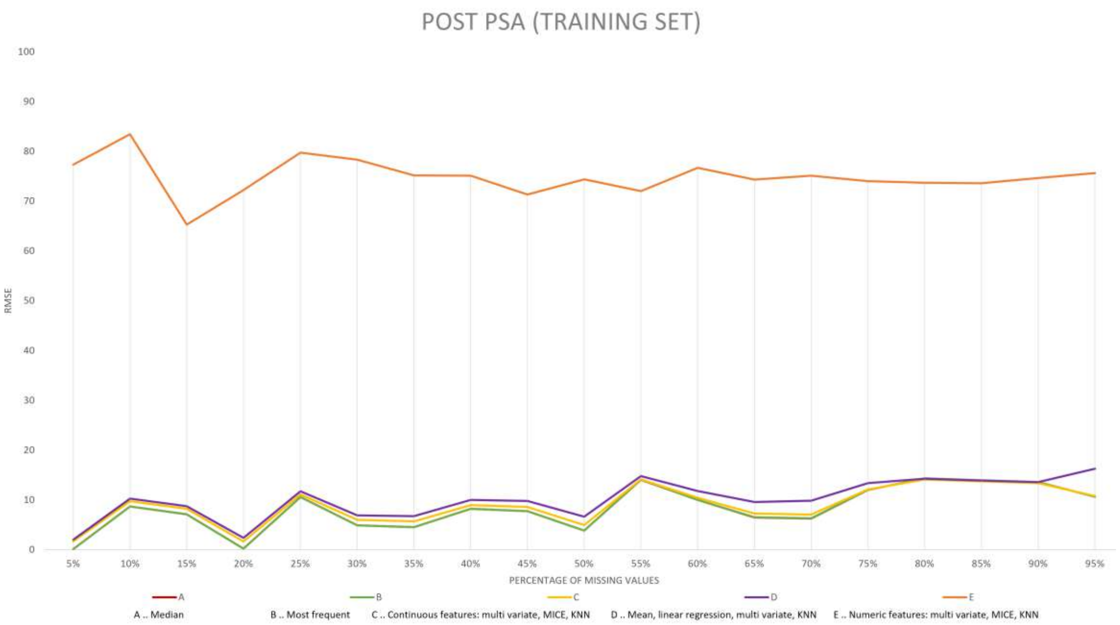


Figure 5.14: RMSE error for imputing a continuous feature on the training set.

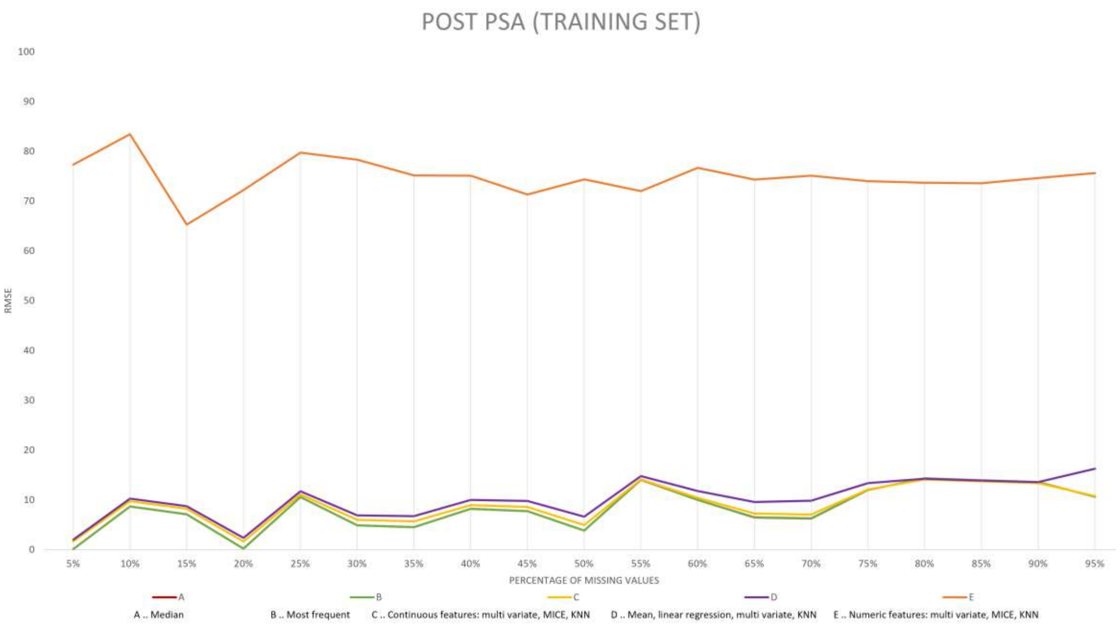


Figure 5.15: RMSE error for imputing a continuous feature on the test set.

5.2.2 Clustering Results and Evaluation

Table 5.1 shows the evaluation of the 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 is, the better the clusters are defined. In contrary, a lower Davies-Bouldin index is related to a better separation of clusters.

Table 5.1: 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.

	t-SNE	**MDS**	**FAMD**	**UMAP**	**PCA**	
k-means (2 clusters)	0.89	0.42	0.97	0.65	0.89	Silhouette coefficient
	414.15	37.70	104.65	111.35	118.47	Calinski-Harabasz index
	0.40	1.30	0.03	0.81	0.02	Davies-Bouldin index
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	0.0039	
Hierarchical (4 clusters)	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	
Hierarchical (6 clusters)	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	

For t-SNE, k-means and mean-shift lead to the best cluster separation. Both methods detect the two visual clusters in the data. For MDS, mean-shift leads to the highest separation score. However, mean-shift defines one cluster with only one point assigned, which is not the desired space division. The next methods with the highest separation values are k-means and hierarchical clustering. While k-means identify two clusters in the data, the hierarchical clustering algorithm determines the number of clusters specified. For FAMD, k-means leads to the highest separation. However, it determines two clusters, where one of them consists of only one point. Mean-shift and hierarchical clustering lead to the next highest scores. Mean-shift identifies five clusters that also match the visual grouping of the data through the FAMD algorithm. For UMAP, the k-means and mean-shift algorithms identify both two clusters. However, they differ in the structure and are not well separated. PCA has the highest separation scores for k-means and mean-shift, but both of them consist of clusters that contain only outliers, which is not the desired space division. The next highest score for PCA is reached through hierarchical clustering by using four clusters. The results of these clustering algorithms applied to the dimensionality-reducted data are shown in Figure 5.16 – Figure 5.40. We set three dimensionality reduction and clustering combinations as presets in the interface to facilitate applying them to the data. Furthermore, we allow users to freely investigate different methods by applying them to the data.

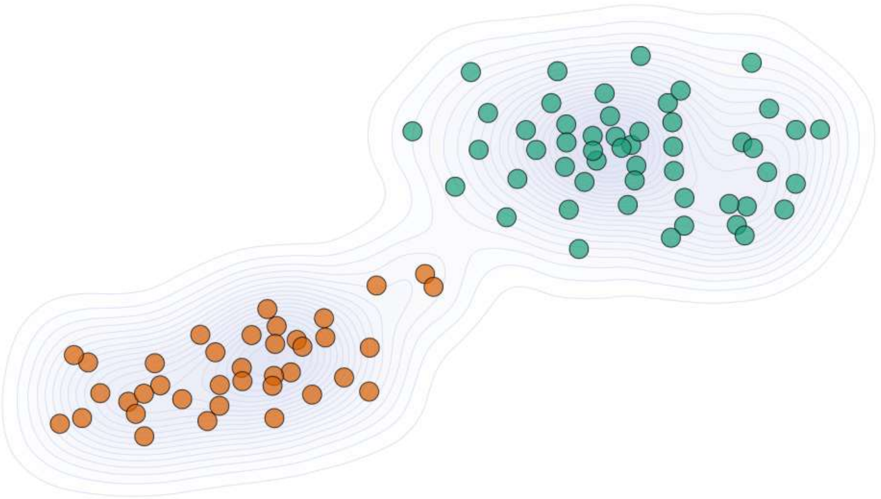


Figure 5.16: K-means and mean shift clustering lead both to the same clustering on the t-SNE dimensionality reduction result.

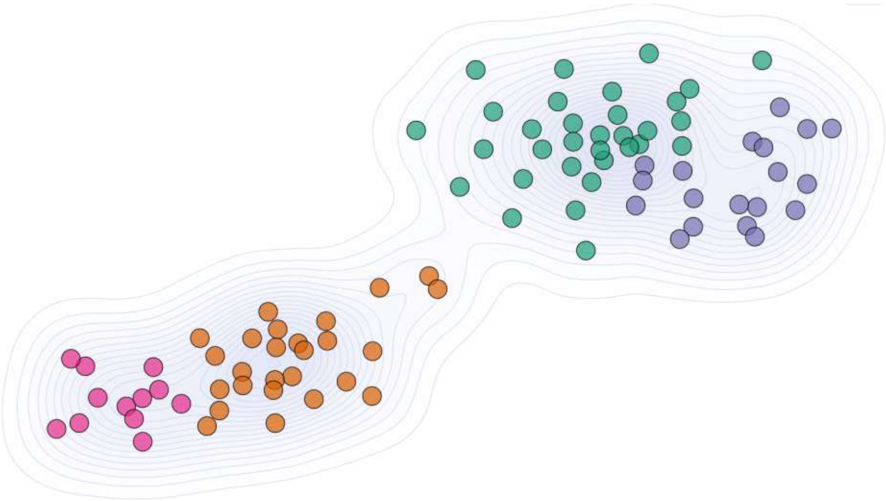


Figure 5.17: Hierarchical clustering with four clusters applied to the t-SNE dimensionality reduction result.

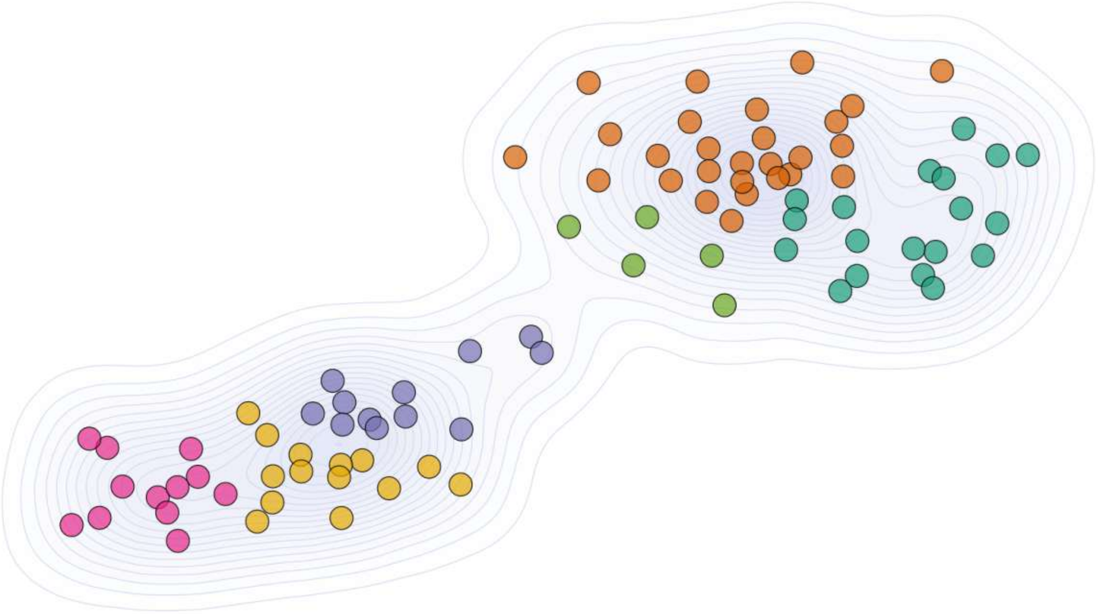


Figure 5.18: Hierarchical clustering with six clusters applied to the t-SNE dimensionality reduction result.

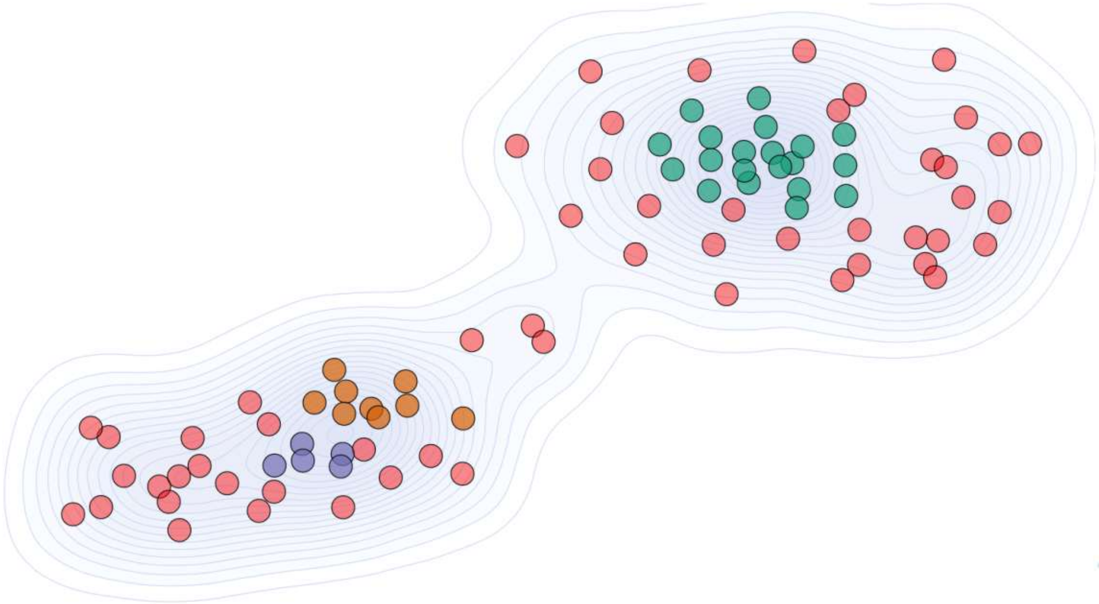


Figure 5.19: OPTICS clustering applied to the t-SNE dimensionality reduction result.

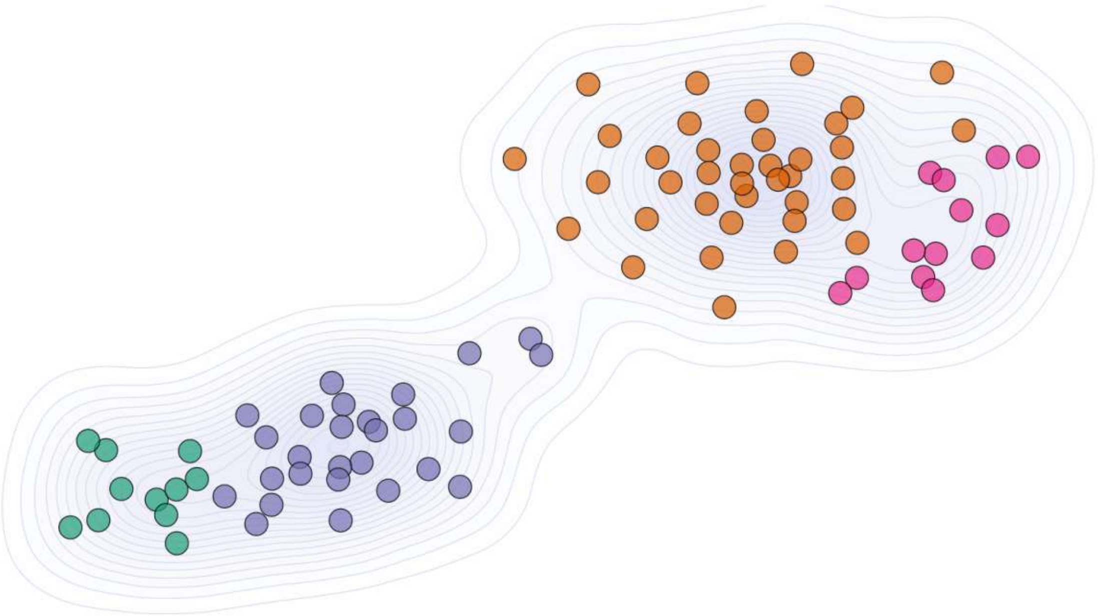


Figure 5.20: GMM clustering applied to the t-SNE dimensionality reduction result.

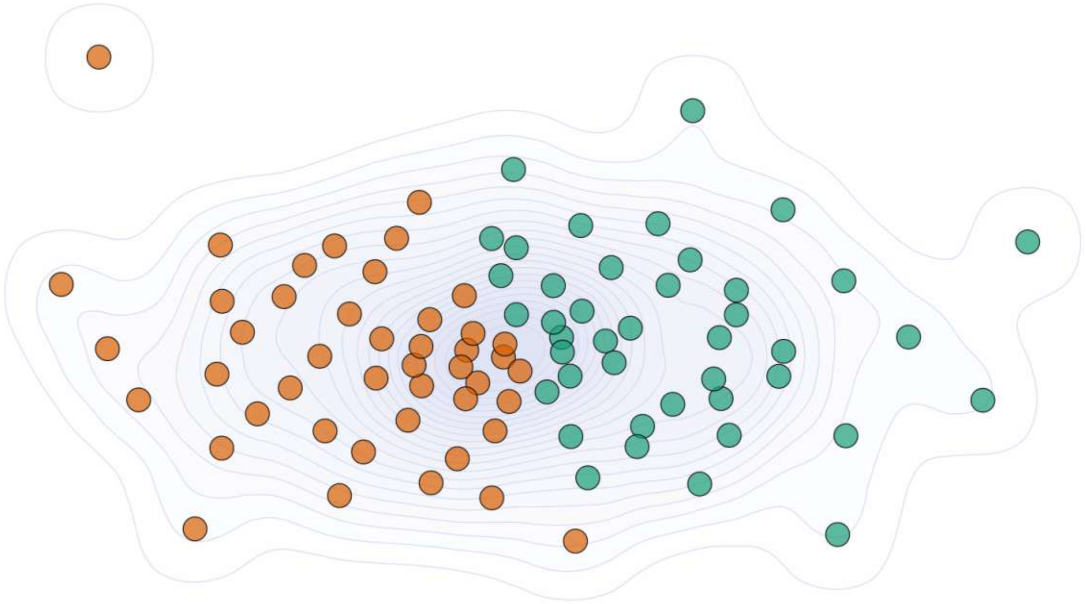


Figure 5.21: K-means clustering applied to the MDS dimensionality reduction result.

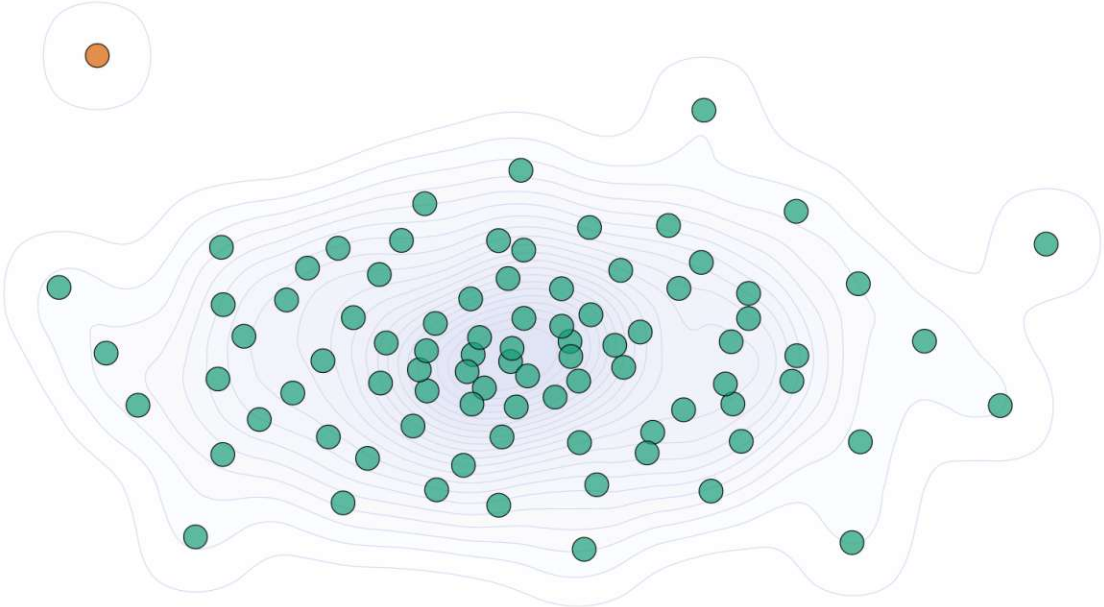


Figure 5.22: Mean-shift clustering applied to the same clustering on the MDS dimensionality reduction result.

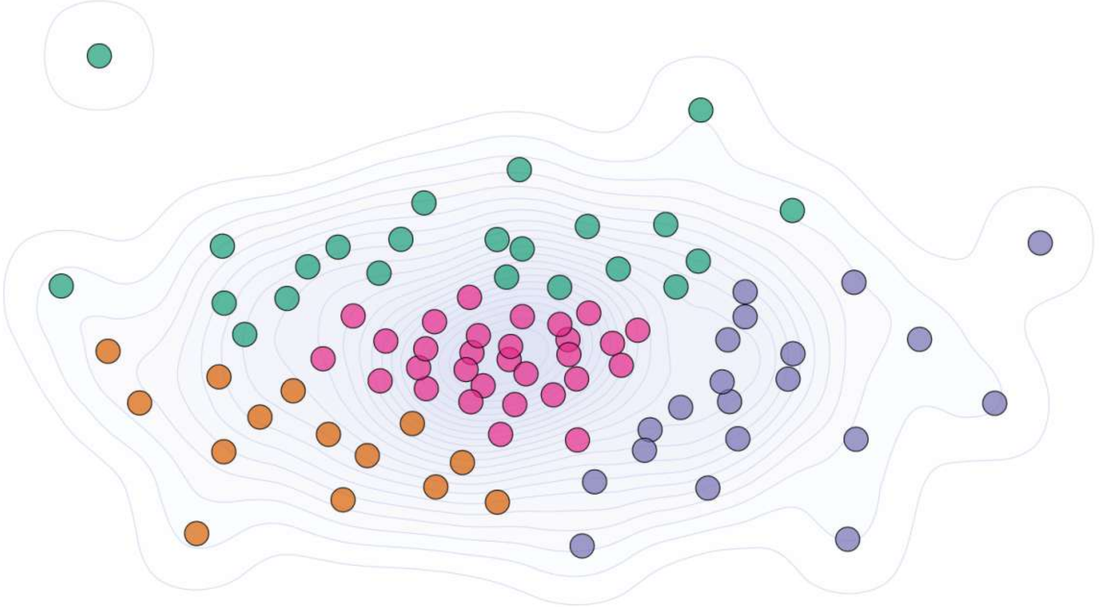


Figure 5.23: Hierarchical clustering with four clusters applied on the MDS dimensionality reduction result.

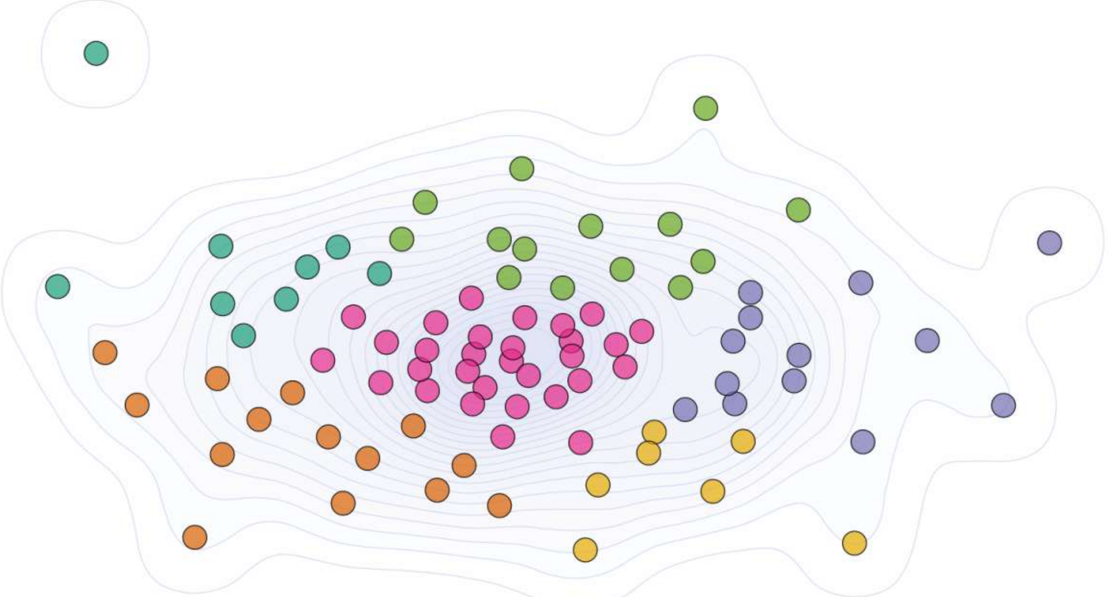


Figure 5.24: Hierarchical clustering with six clusters applied on the MDS dimensionality reduction result.

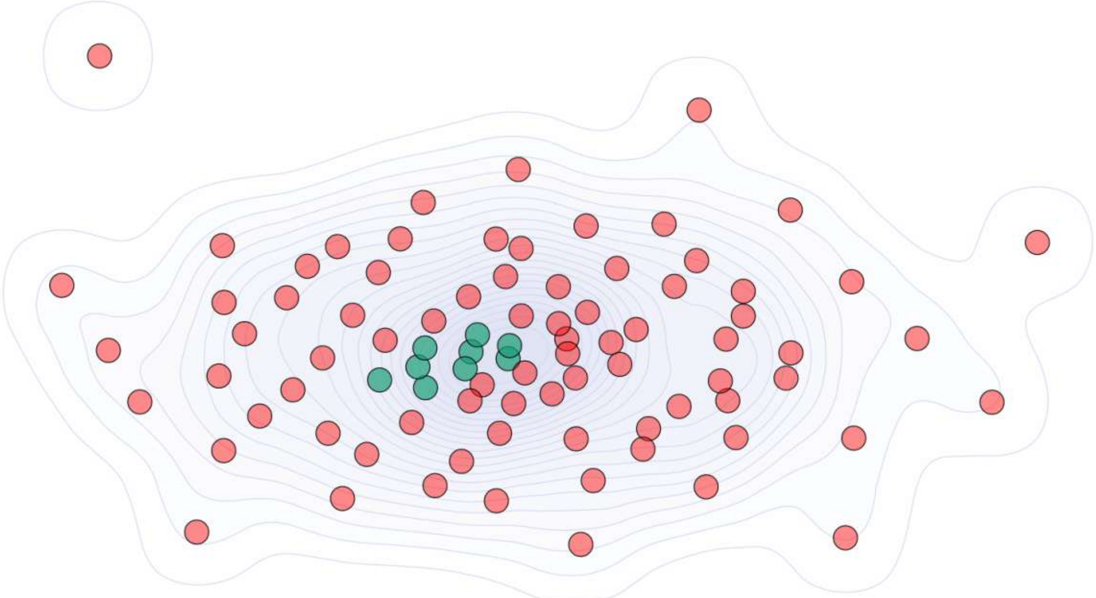


Figure 5.25: OPTICS clustering applied on the MDS dimensionality reduction result.

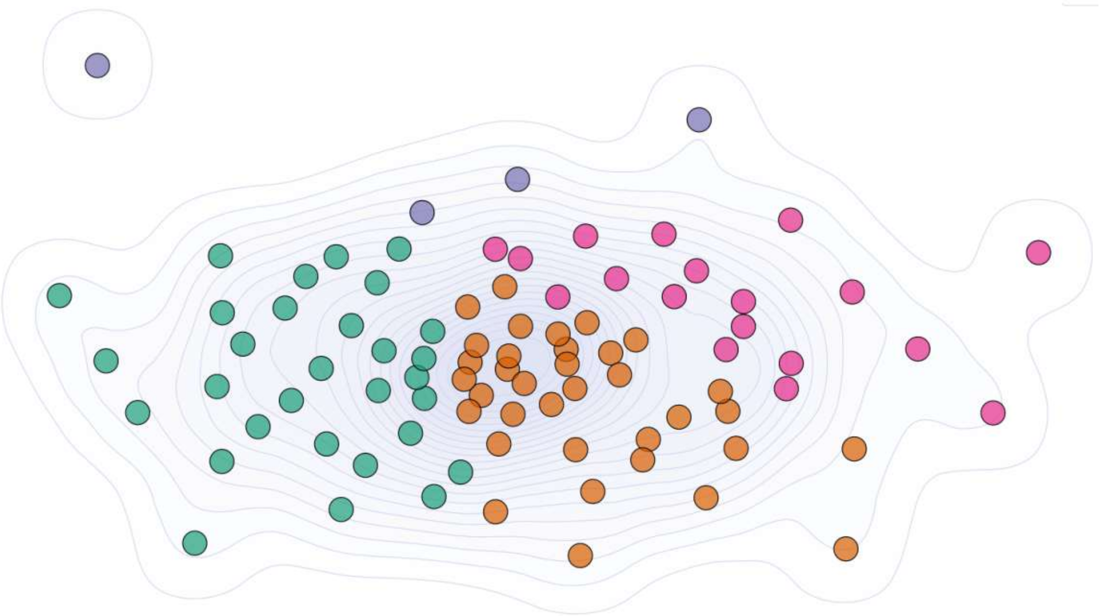


Figure 5.26: GMM clustering applied on the MDS dimensionality reduction result.

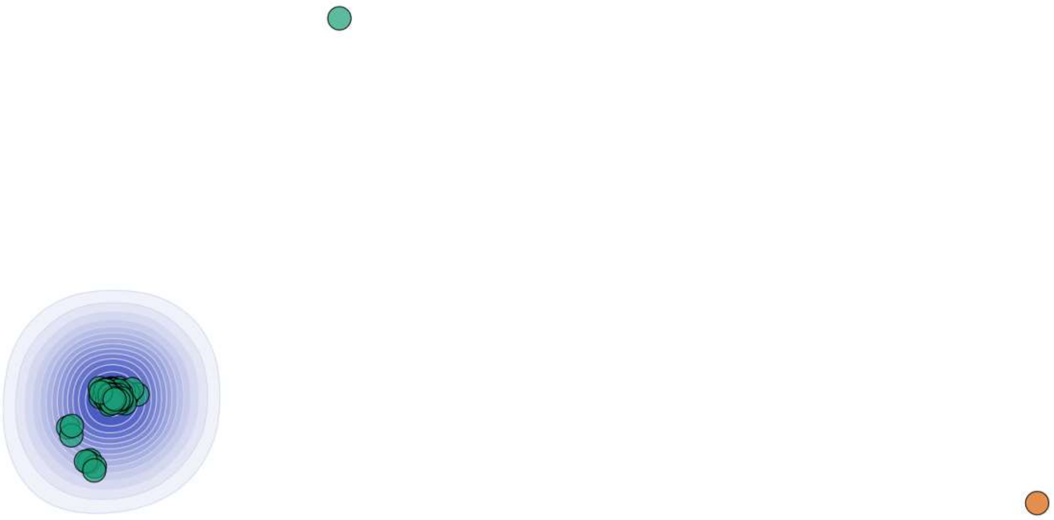


Figure 5.27: K-means clustering applied to the FAMD dimensionality reduction result.

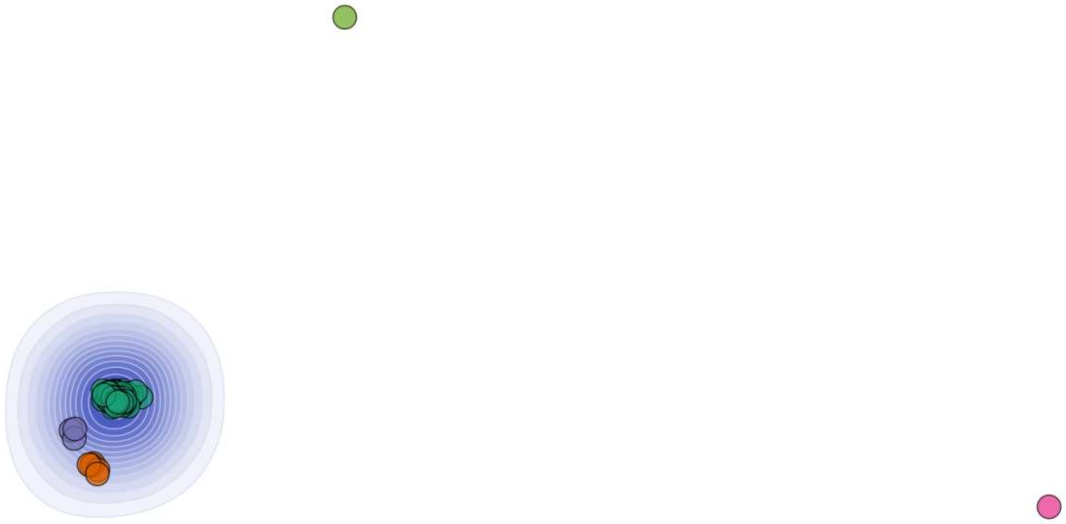


Figure 5.28: Mean-shift clustering and hierarchical clusters with four clusters lead both to the same clustering result when applied to the FAMD dimensionality reduction result.

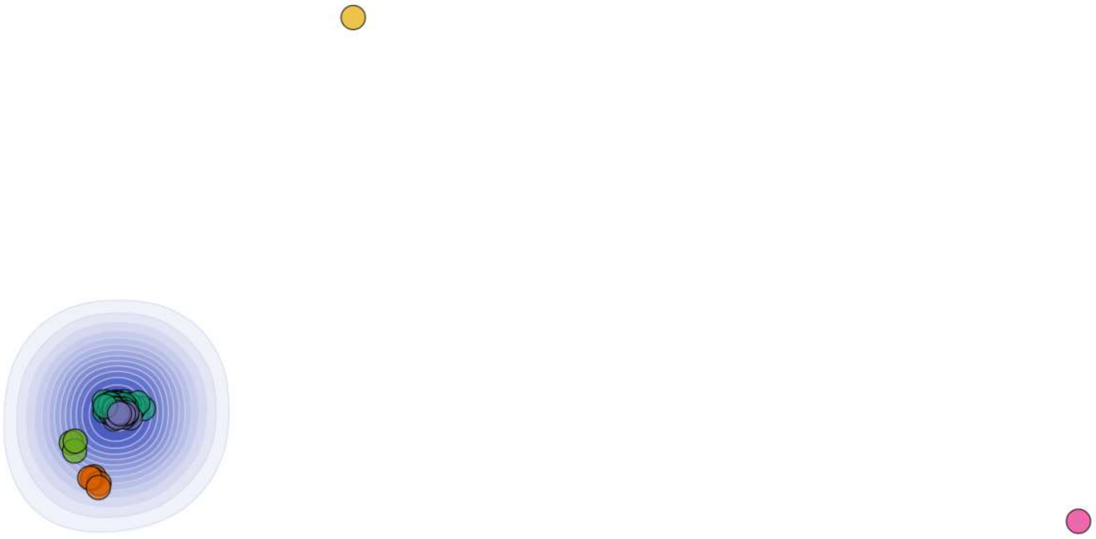


Figure 5.29: Hierarchical clustering with six clusters applied on the FAMD dimensionality reduction result.

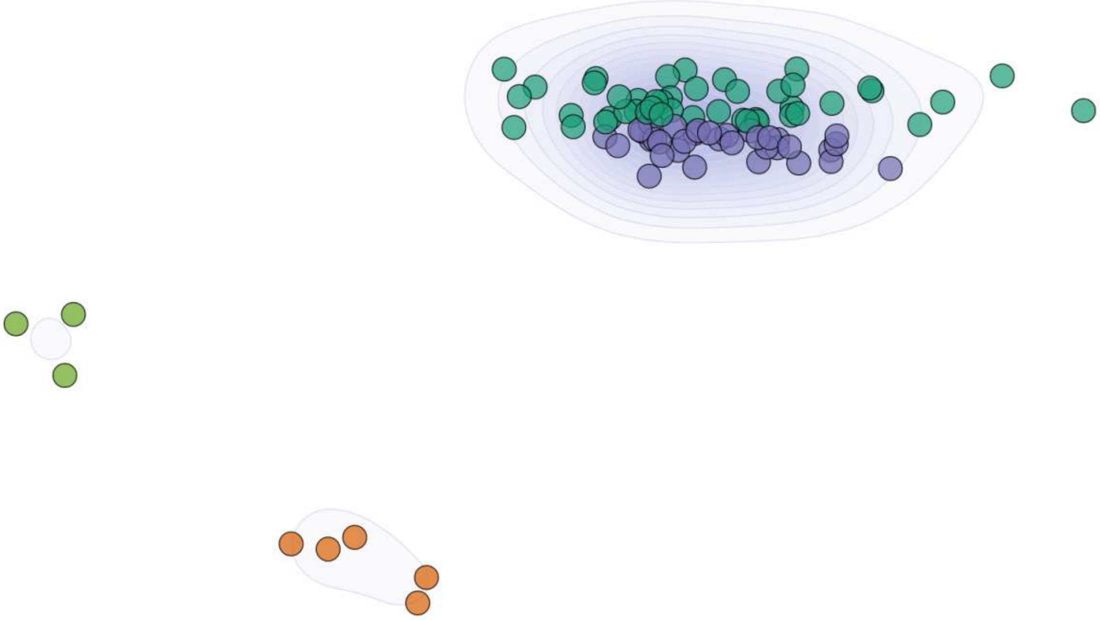


Figure 5.30: Zoomed result of the hierarchical clustering with six clusters applied on the FAMD dimensionality reduction result.

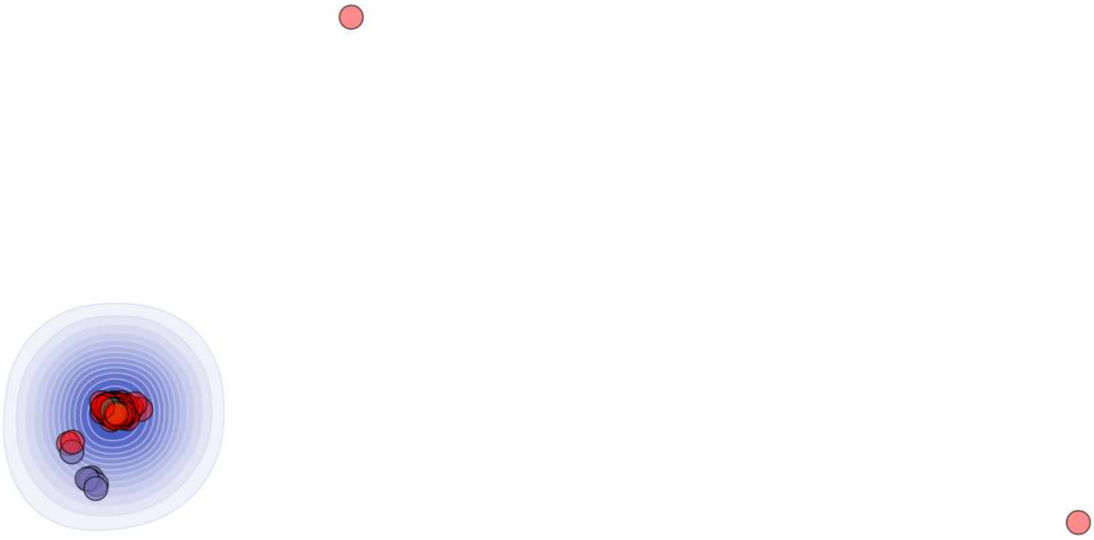


Figure 5.31: OPTICS clustering applied on the FAMD dimensionality reduction result.

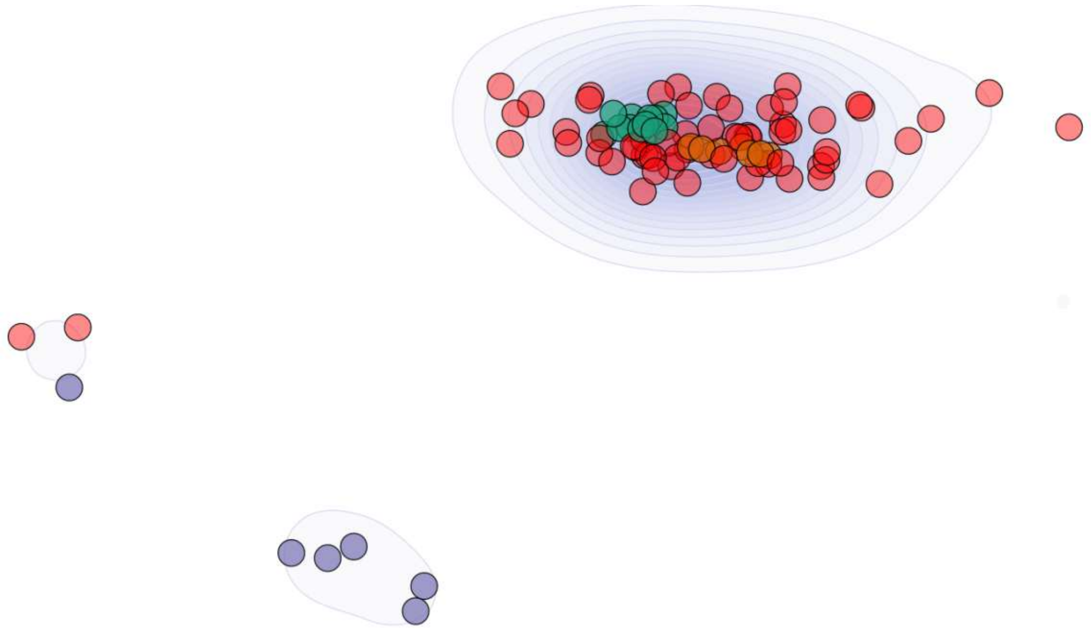


Figure 5.32: Zoomed result of the OPTICS clustering applied on the FAMD dimensionality reduction result.

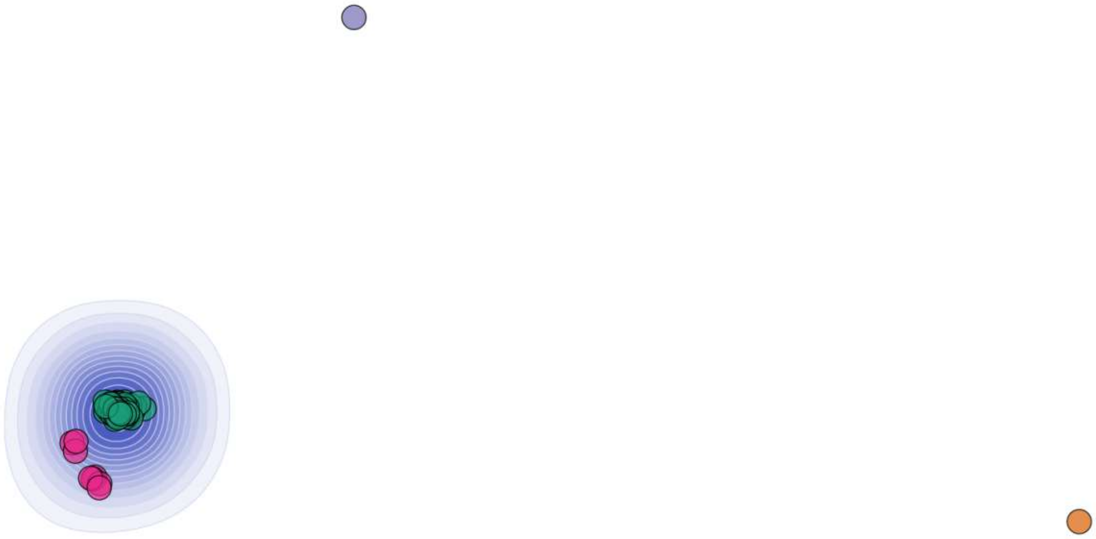


Figure 5.33: GMM clustering applied on the FAMD dimensionality reduction result.

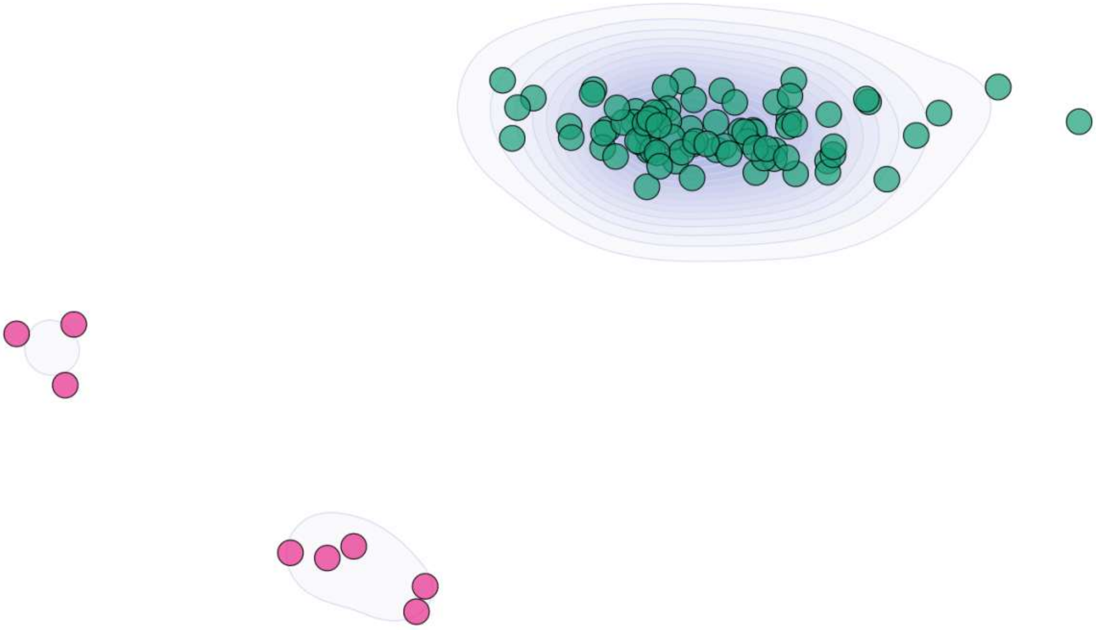


Figure 5.34: Zoomed result of the GMM clustering applied on the FAMD dimensionality reduction result.

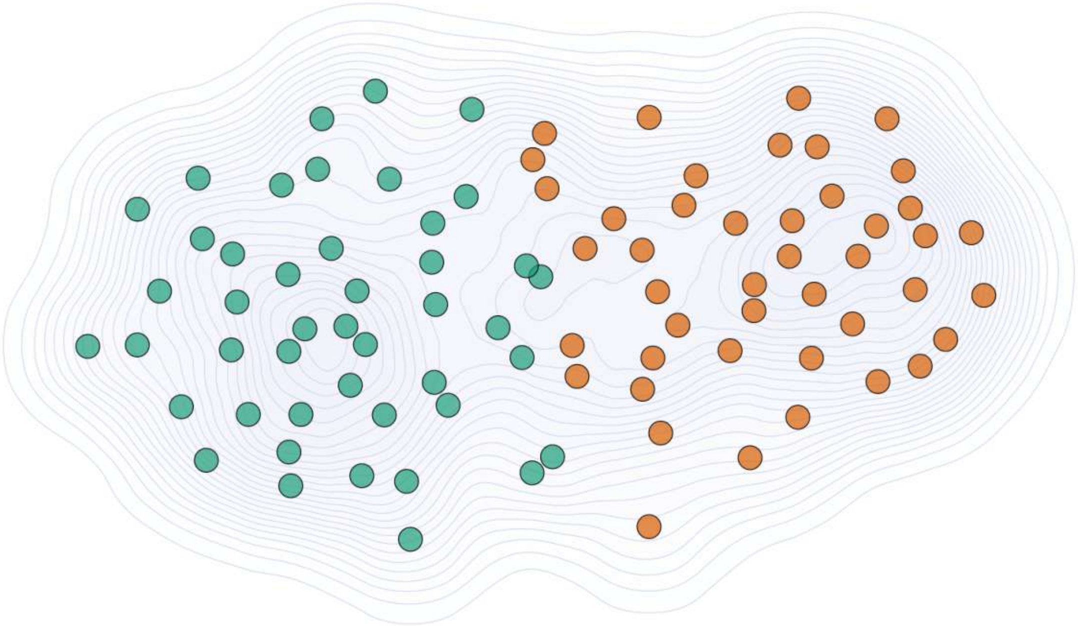


Figure 5.35: K-means clustering applied to the UMAP dimensionality reduction result.

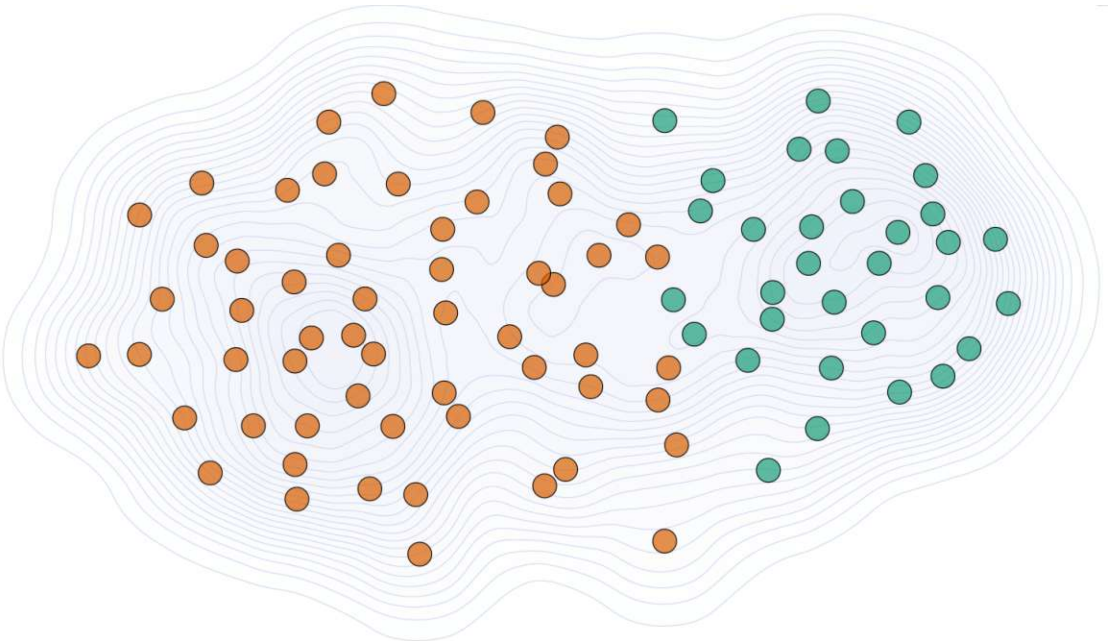


Figure 5.36: Mean-shift clustering applied to the same clustering on the UMAP dimensionality reduction result.

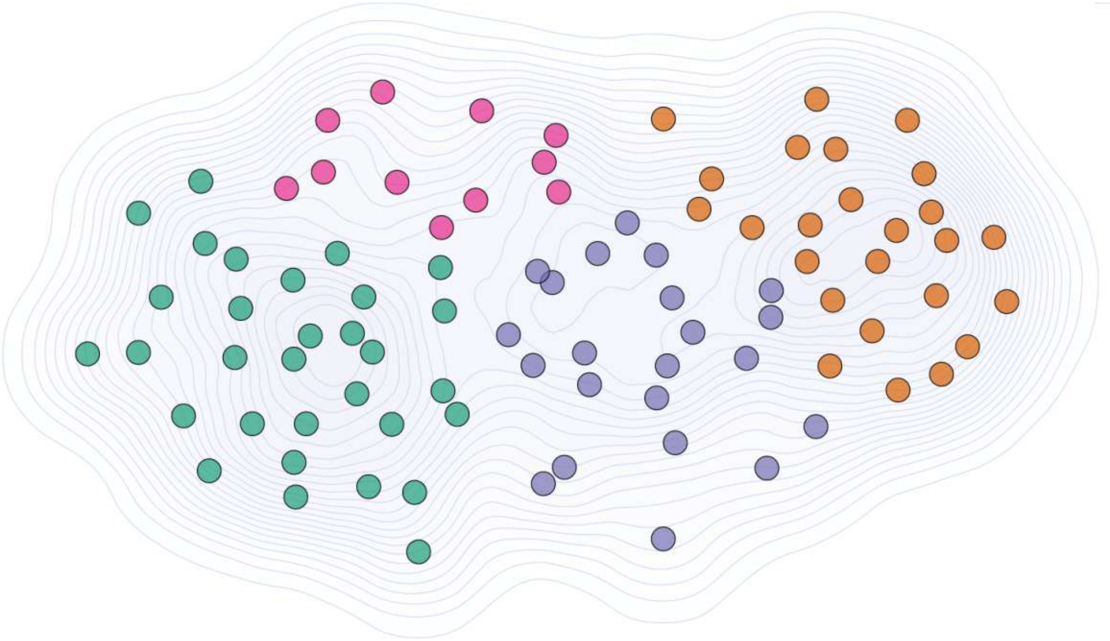


Figure 5.37: Hierarchical clustering with four clusters applied on the UMAP dimensionality reduction result.

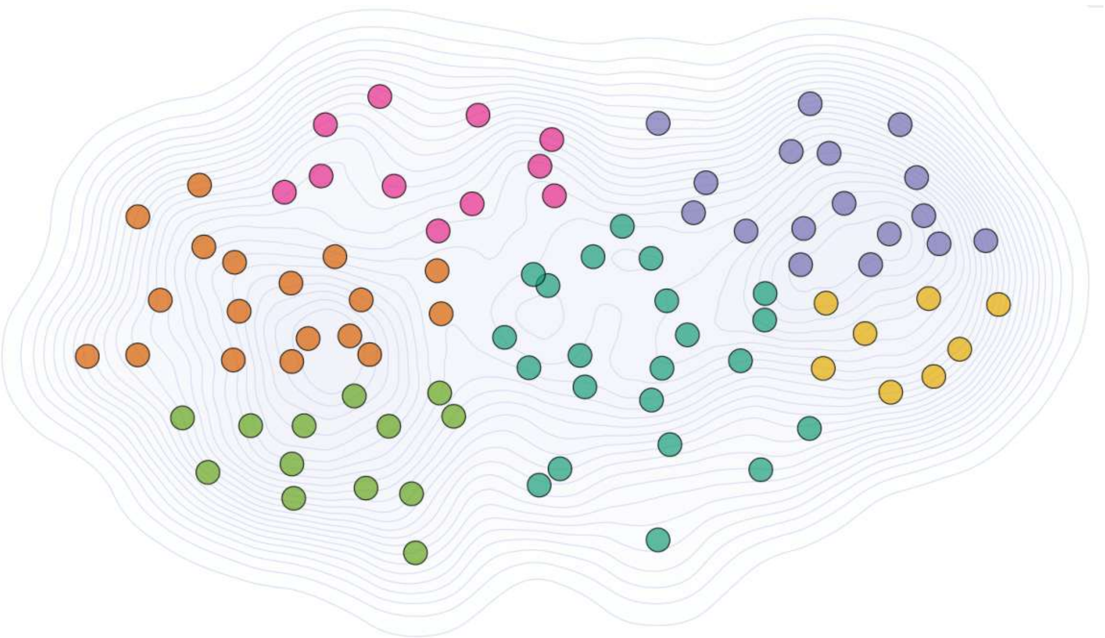


Figure 5.38: Hierarchical clustering with six clusters applied on the UMAP dimensionality reduction result.

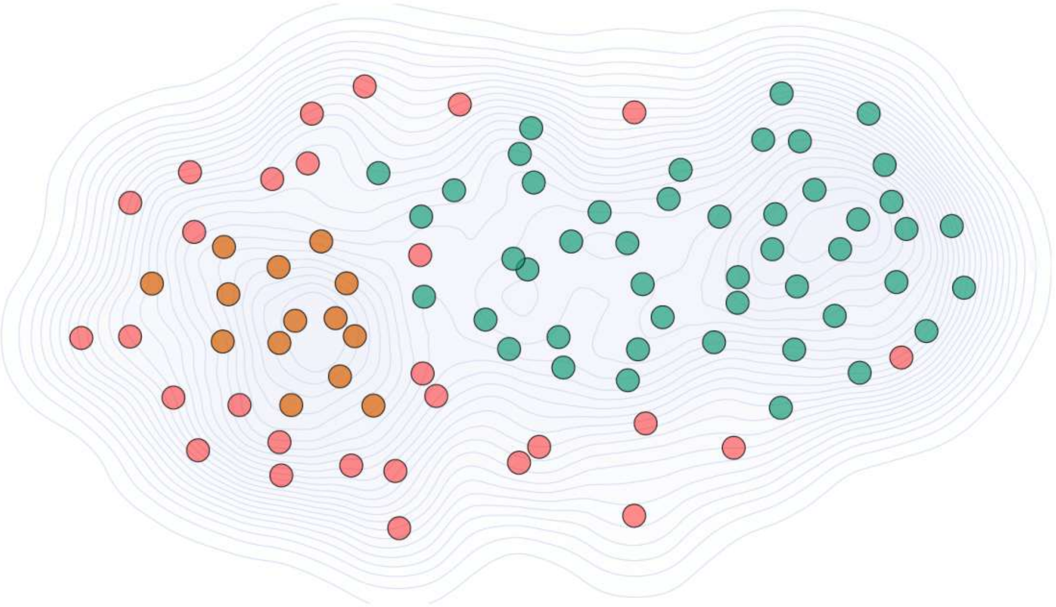


Figure 5.39: OPTICS clustering applied on the UMAP dimensionality reduction result.

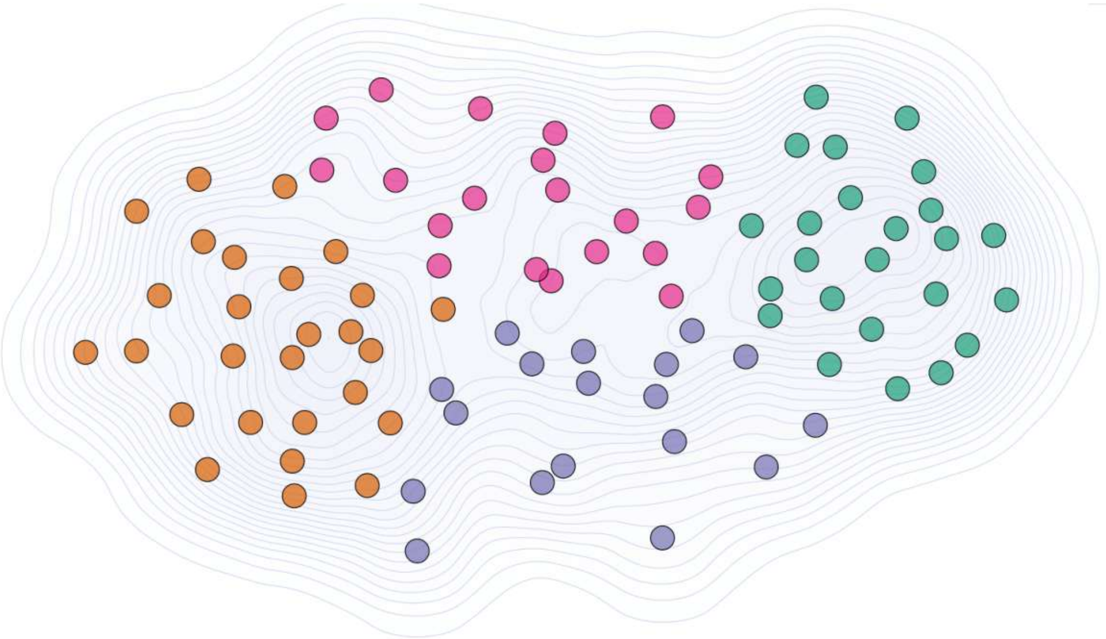


Figure 5.40: GMM clustering applied on the UMAP dimensionality reduction result.

5.3 Qualitative Evaluation

We qualitatively evaluate our framework through usage scenarios, through a cancer expert testing and commenting on our application, and through regularly presenting the progress of the application to domain experts and receiving comments on its functionality and clinical applicability. We summarize the feedback of domain experts in Subsection 5.3.2 and present a selection of ten usage scenarios out of sixty in Subsection 5.3.1.

5.3.1 Usage Scenarios

To evaluate the data analysis, knowledge discovery, or knowledge management capabilities of our framework, we create usage scenarios based on the hypothetical view of domain experts as described by Isenberg et al. [IIC+13] and Lam et al. [LBI+12]. We mainly include the cases that our domain experts communicate as being of interest in the scenarios. Furthermore, we validate the importance of the scenarios with them and receive extensive comments and extended cases for further scenarios that we included. This resulted in 60 usage scenarios, from which we select ten examples together with our cancer experts to demonstrate the usage of our application for knowledge discovery and hypothesis assessment.

Free selection of patients on the scatterplot By clustering the data, the patients are grouped into two groups with similar radiogenomic and clinical profiles. To understand why the two clusters differ, we show the top features characterizing or differentiating them. To further reveal why these features characterize or differentiate the clusters, we visualize their distributions for both clusters on demand. We additionally allow users to see feature distributions of patients grouped by the identified clusters as initial feedback on the data. By default, we show the values of the complete data to highlight differences between the clusters. Users can then make any free selection of patients on the scatterplot to filter the distribution plots for the selected patients. Figure 5.41 shows a selection of three orange points in the center and a subset of green points that visually form a subcluster on the top right. These are indicated through the inflated points in the scatterplot. This reveals that in the orange cluster, the selected patients all have low *PSA-pre OP* values, while the *PSA-pre OP* values of the selected patients in the green cluster also include higher scores as shown in the pyramid plot that is brought into the view on demand. Furthermore, this highlights that the distributions of the *age* or *ISUP* grade of these patients differ between the orange and green cluster and that none of the selected patients has an *ISUP* grade of 1.

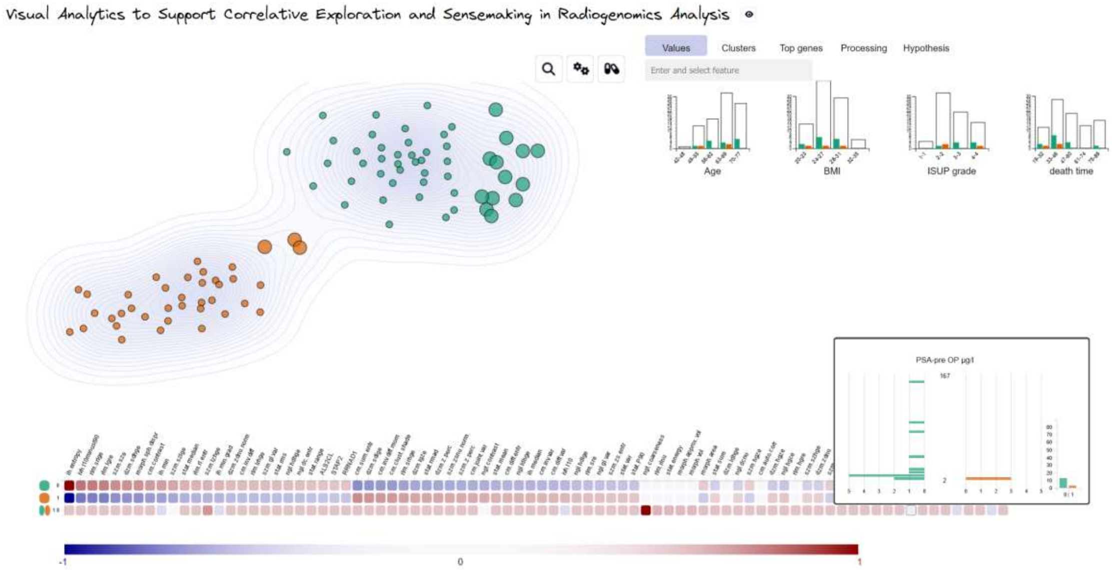


Figure 5.41: Free patient selection on the scatterplot to see the feature distributions for the selected patients.

Characteristics of a patient cluster Under the *Clusters* tab, we show an enlarged view of the characteristics or differences of clusters. By default, we show the differences between the identified clusters. In Figure 5.42, the green cluster (the grouping on the top right of the scatterplot) is selected to display its characterizing features. In this case, all top features are from the radiomic dataset. We allow users to filter the datasets to see also the characterizing genomic or clinical features through checkboxes. In Figure 5.42, the first radiomic feature is selected, which represents the entropy of the intensity histogram. Entropy is a measure of uncertainty or randomness in the image that characterizes the image texture [TLM08]. It can characterize heterogeneity or homogeneity of the tumor and might indicate tumor regression [DAB+17, MAM+23]. In this example, the patients in the green cluster (on the top right of the scatterplot) have higher entropy values compared to patients in the orange cluster (on the bottom left of the scatterplot).

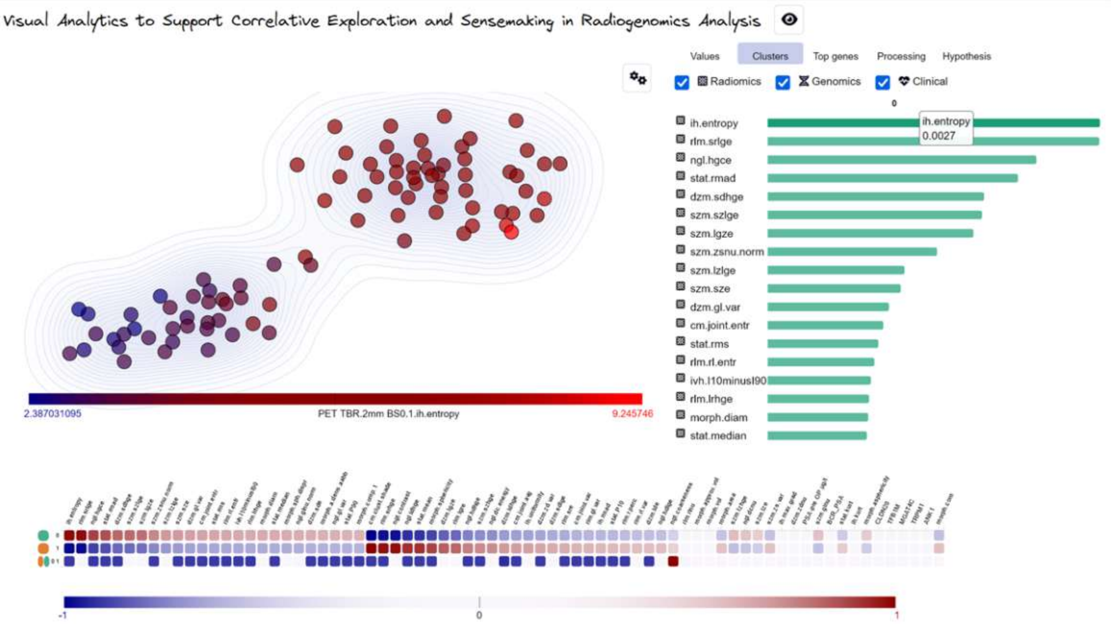


Figure 5.42: Investigation of the radiomic feature *ih.entropy* as one of the characteristics of the green cluster. Its values are higher in the green cluster (top right of the scatterplot) compared to the data points on the bottom left.

Top genes based on a specified condition The *Top genes* tab highlights the top gene mutations for all patients or an active patient selection on the scatterplot. These represent gene mutations that occur for most patients with a value higher than zero. In Figure 5.43, all patients with an *ISUP* grade of four or five are selected to investigate their top gene mutations. Furthermore, the first gene mutation, in this case *PLEC*, is highlighted on the scatterplot to show its values for the selected patients (large scatter points that have an *ISUP* grade of four or five) and the not selected patients (small scatter points that have an *ISUP* grade of one to three) in comparison. The points with no filling do not have this gene mutation, while all colored points have it. Larger points are the ones that also match the current hypothesos. *PLEC* ist of relevance as it maintains tissue integrity that regulates cell survival [RTS+19, WBKA21].

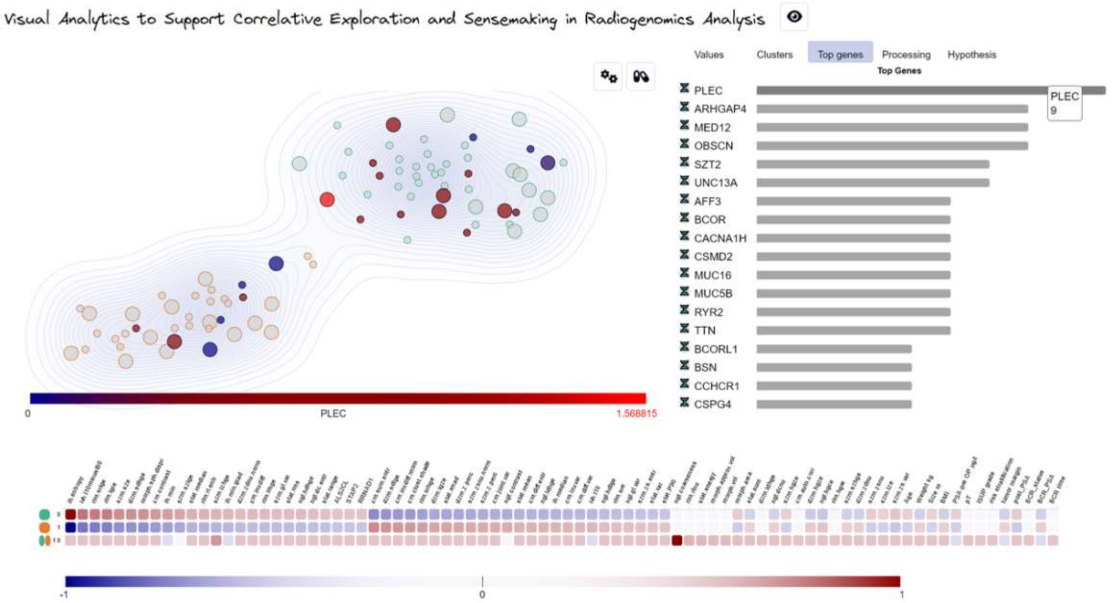


Figure 5.43: Investigation of the top genes of all patients with an *ISUP grade* of four or five (large points on the scatterplot) and selection of the *PLEC* gene mutation. Filled scatter points represent patients with the *PLEC* gene mutation.

Comparison of clusters formed through a specified condition In the scenario in Figure 5.44, patients with an *ISUP* grade of four or five are selected (black selection in the histogram). The heatmap features are updated to compare patients that match this condition of having an *ISUP* of at least four with patients that do not match it and have an *ISUP* of three or smaller. This allows users to identify features that differentiate or characterize the specified groups. For example, the *BCOR* gene mutation occurs only for patients that match the specified condition. *BCOR* indicates aggressive cancer diseases [AFM+19].

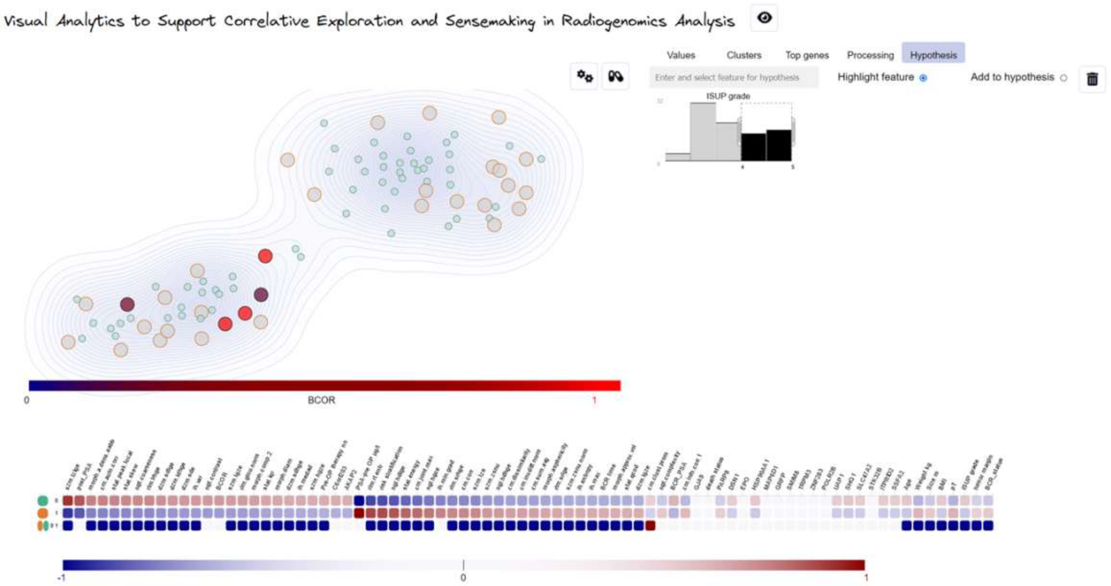

Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis
Values
Clusters
Top genes
Processing
Hypothesis
Enter and select feature for hypothesis
Highlight feature
Add to hypothesis
ISUP grade
4
5
0
BCOR
1
-1
0
1

Figure 5.44: Comparison of patients with an *ISUP* grade of four or five (large points on the scatterplot with an orange border) with patients that have lower *ISUP* grades. The heatmap shows which features characterize or differentiate these two *ISUP* groupings the most. For example, the *BCOR* gene mutation is a characteristic of patients with an *ISUP* of four or five located in the bottom left cluster. Filled scatter points represent patients with the *BCOR* gene mutation.

Assessment of a hypothesis involving a gene and clinical score A scenario for the assessment of the following hypothesis is shown in Figure 5.45:

All patients that have the MED12 gene mutation, have an ISUP grade of at least 2.

This hypothesis is confirmed for the underlying data, which is revealed through interacting with the feature sliders. After selecting all patients that have the *MED12* gene mutation (black selection on the histogram), and adding the *ISUP* grade feature to the hypothesis (textbox with suggestions on top of the histograms), moving the minimum limit of the *ISUP* grade slider to 1 keeps the selected patients on the scatterplot unchanged. In contrast, a higher grade leads to a filtering of the selected patients on the scatterplot. Users can freely choose any thresholds for the features of interest (through the histograms) to interactively explore how this selection influences the result.

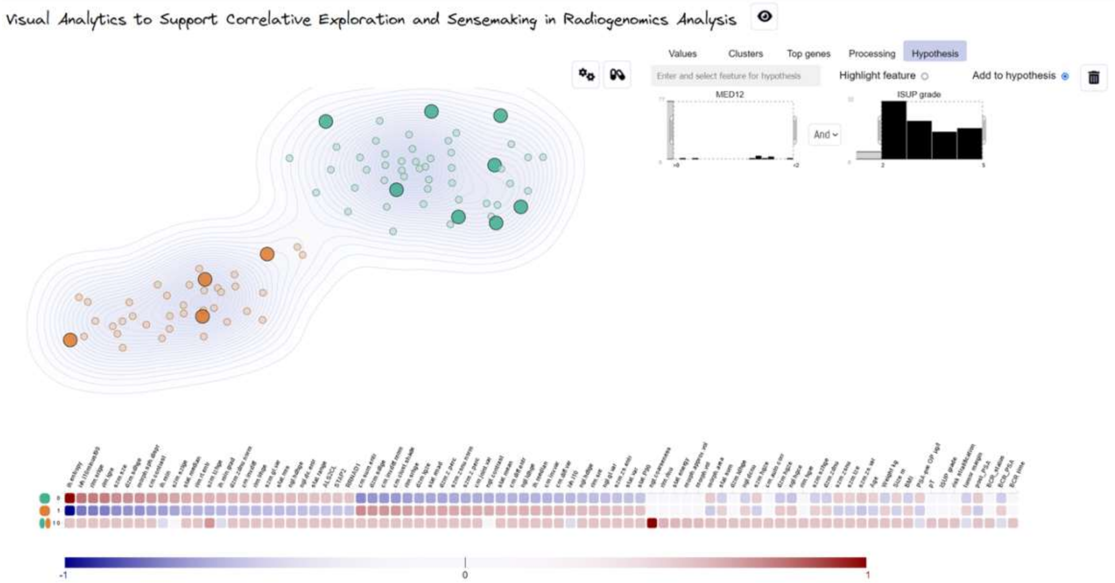

Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis
Values
Clusters
Top genes
Processing
Hypothesis
Enter and select feature for hypothesis
Highlight feature ○
Add to hypothesis ●
MED12
And
ISUP grade
2
5
-1
0
1

Figure 5.45: Assessment of a hypothesis involving the *MED12* gene mutation and the *ISUP* grade. All patients that have the *MED12* gene mutation have an *ISUP* grade of at least two, which is revealed by interactively changing the *ISUP* grade sliders.

Assessment of a hypothesis involving two genes This scenario assesses the following hypothesis as shown in Figure 5.46:

All patients that have the PLEC gene mutation do not have the MED12 gene mutation.

This hypothesis is rejected for the patient depicted by the red arrow (Figure 5.46) as this patient has both gene mutations. However, it is confirmed for all other patients. Users can further adjust the limits to identify, for example, patients that do not have any of the two gene mutations or have at least one of them and freely combine the resulting subset with other features of interest. The heatmap on the bottom supports users in identifying features of relevance for the current clustering on the scatterplot.

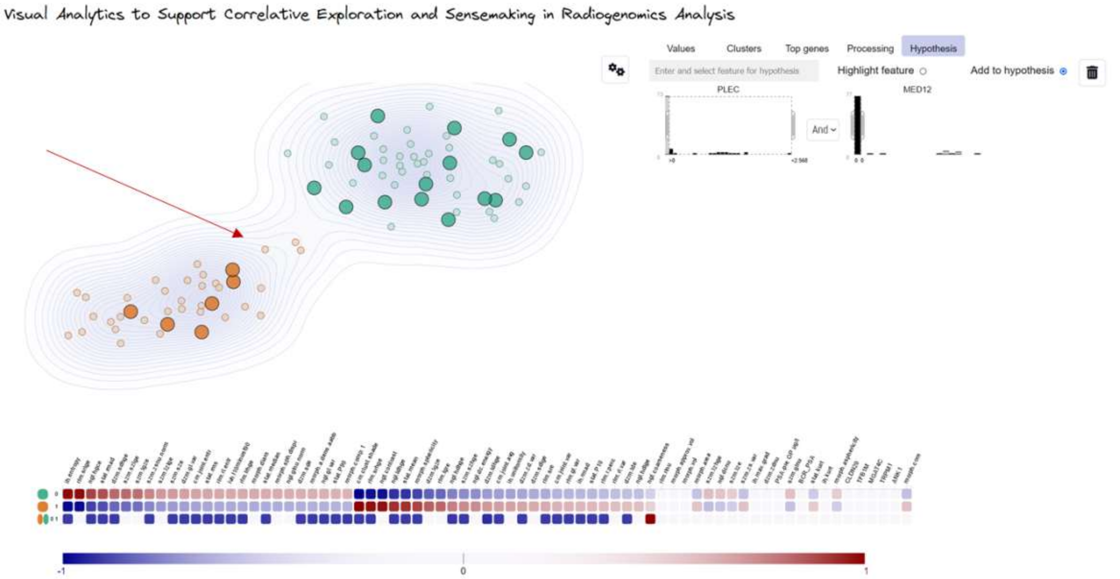

Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis
Values
Clusters
Top genes
Processing
Hypothesis
Enter and select feature for hypothesis
Highlight feature ○
Add to hypothesis ●
PLEC
And
MED12
>0
<2.568
0 0
-1
0
1

Figure 5.46: Assessment of a hypothesis involving the *PLEC* gene mutation and the *MED12* gene mutation. Only one patient (depicted by the red arrow) has both gene mutations, while all other patients that have the *PLEC* gene mutation do not have the *MED12* gene mutation. This is revealed through interacting with the feature sliders.

Hypothesis generation and refinement Figure 5.47 shows a scenario to create and refine a hypothesis based on features of interest. The users are interested in high *PSA-pre OP* scores with no concrete hypothesis in mind. By adding all features of interest, users can interactively identify and refine ranges that are true for high *PSA-pre OP* scores. In this case, we identify the following true statement for the underlying data:

Patients that have the highest PSA-pre OP values starting from 665, have post PSA values starting from 36 and BCR PSA values starting from 8 and do not have the MED12 gene mutation and they have an ISUP grade of 5.

This statement is generated through adding histograms for the features, and interactively selecting the threshold ranges (marked in black on the histograms). The resulting patient subset is then highlighted on the scatterplot (three large orange points in this case).

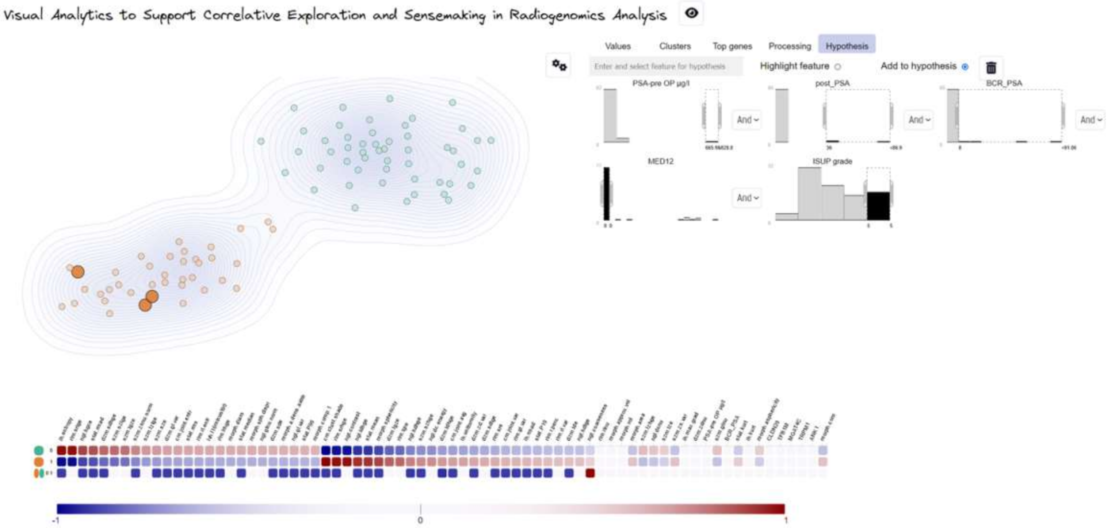


Figure 5.47: Identification of feature ranges for patients with a high *PSA-pre OP* value. Users can add features of interest and interactively identify ranges for a matching condition.

Filtering based on features of different datasets The scenario in Figure 5.48 combines 19 features as a hypothesis example with a larger number of features. In general, we allow users to combine, explore, and interact with any number of features of the radiomic, genomic, or clinical dataset for hypothesis filtering. These features can be connected through a logical *and* or a logical *or* operation. We evaluate the statement always from the left side to the right side, which is relevant when mixing different operators in one statement. As an example, the statement *A or B and C* for three features *A, B, C* is evaluated as *((A or B) and C)*, which does not necessarily lead to the same result as *(A or (B and C))*.

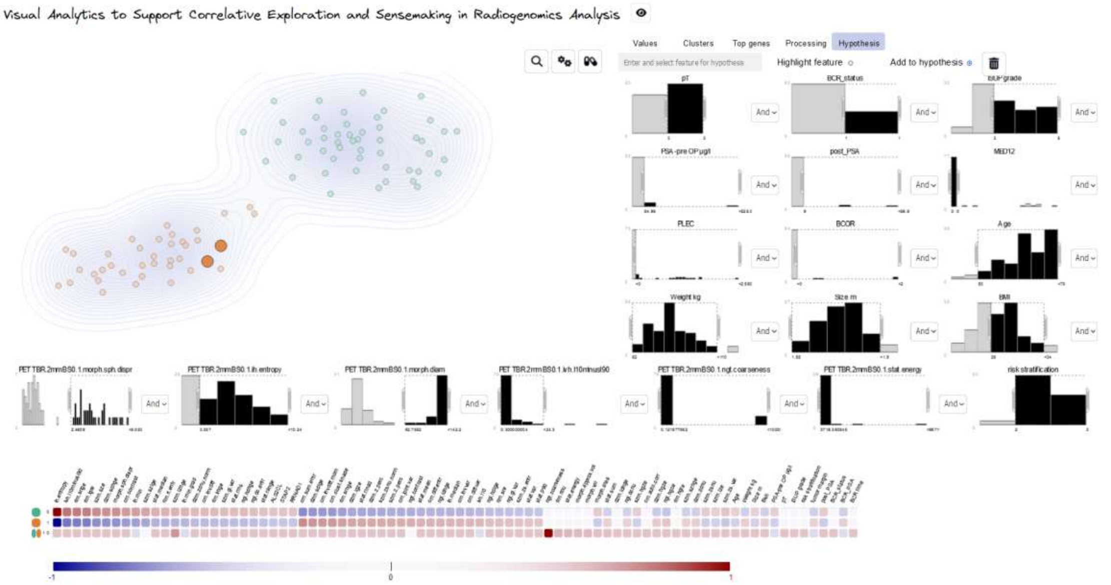

Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis
Values
Clusters
Top genes
Processing
Hypothesis
Enter and select feature for hypothesis
Highlight feature
Add to hypothesis
pT
BCR_status
ISUP grade
PSA-pre OP µg/l
post_PSA
MED12
PLEC
BCOR
Age
Weight kg
Size m
BMI
PET TBR.2mmBS0.1.morph.sph.dispr
PET TBR.2mmBS0.1.ih.entropy
PET TBR.2mmBS0.1.morph.diam
PET TBR.2mmBS0.1.ivh.I10minus I90
PET TBR.2mmBS0.1.ngt.coarseness
PET TBR.2mmBS0.1.stat.energy
risk stratification
And
-1
0
1

Figure 5.48: Our framework supports adding any number of features of the radiomic, genomic, or clinical dataset to the hypothesis. This example combines 19 features for demonstration purposes.

Highlighting feature values on a hypothesis result We allow users to select any subset of radiomic, genomic, and clinical data features to apply the preprocessing steps and cohort stratification on it. As the radiomic features dominate when combining all features of the three datasets, the scenario in Figure 5.49 excludes the radiomics data from the analysis to see how this affects the grouping of the features. In this example, all genomic and clinical features are included in the analysis through the *Processing* tab, but none of the radiomics features. Filtering is applied to show patients with a pT and *BCR status* of 1 and an *ISUP* grade starting from 1. In this example, the identified feature combination and ranges correlate with the green cluster (on the top part of the scatterplot). We highlight the *ISUP* grade feature on the data to get an overview of the *ISUP* grade values on the scatterplot. This also shows that the *ISUP* grade has higher values for patients in the green cluster than those in the orange cluster. The user can explore the data further by selecting different datasets to process to explore how this affects the resulting patterns.

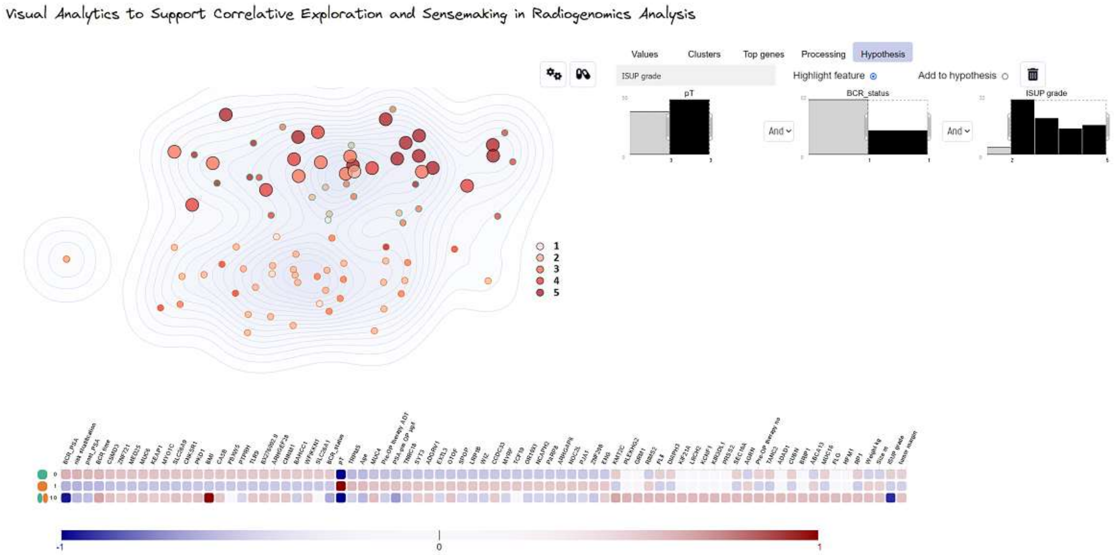
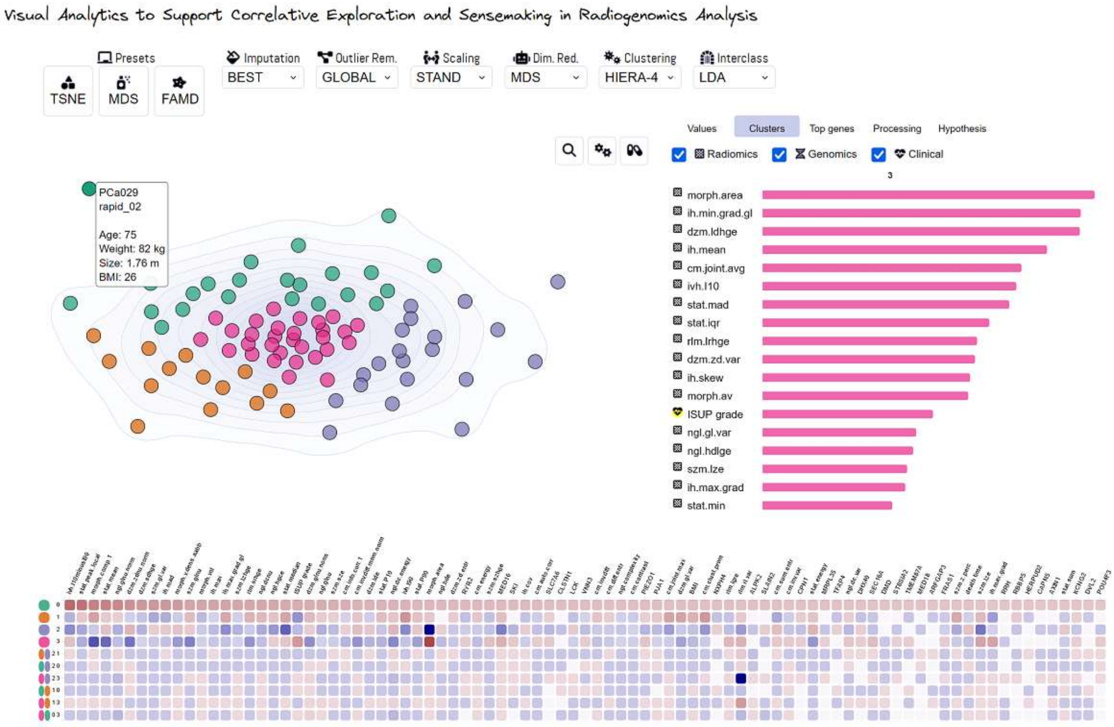


Figure 5.49: Hypothesis selection (large points on scatterplot) of patients with a pT and *BCR status* of 1, and *ISUP* grades starting from 2 to identify whether this subset correlates with a cluster using only the genomic and clinical data. The values of the *ISUP* grade are highlighted on the scatterplot.

Investigating presets and advanced options on demand Figure 5.50 shows advanced options displayed on demand. Contrary to the nine previous scenarios, this scenario addresses biomedical data scientists instead of cancer experts. Biomedical data scientists working with these data are interested in comparing and understanding how different parameters influence the resulting patterns and clusters and which analysis steps are applied to the data. In this example, the *MDS* preset is used that involves a global outlier removal and hierarchical clustering with four clusters on the complete radiogenomic and clinical data by default. This results in four different clusters highlighted on the scatterplot. Their characterizing and differentiating featurs are further shown on the heatmap. For example, the radiomic feature for the morphological area has a dark red color and therefore high relevance on the heatmap for the pink cluster. In contrary, it has a dark blue and therefore lower impact on the clustering outcome in the violett cluster. The user can investigate any features further by moving the mouse over them to show their distributions or by clicking on them to visualize their values on the scatterplot. We allow exploring different patient or feature subsets, imputation methods, outlier removal options, data scaling, dimensionality reduction, clustering, and the interclass methods used for pairwise cluster comparison on demand.



Visual Analytics to Support Correlative Exploration and Sensemaking in Radiogenomics Analysis

Imputation: BEST
Outlier Rem.: GLOBAL
Scaling: STAND
Dim. Red.: MDS
Clustering: HIERA-4
Interclass: LDA

☑ Radiomics ☑ Genomics ☑ Clinical

PCa029
rapid_02

Age: 75
Weight: 82 kg
Size: 1.76 m
BMI: 26

Figure 5.50: Advanced options can be investigated on demand. These include selecting predefined presets or changing the imputation, outlier removal method, scaling, dimensionality reduction, clustering, or interclass method. In this example, the MDS preset is used, which applies hierarchical clustering to the data.

5.3.2 Feedback from Domain Experts

One cancer expert interacted with the tool to explore the data. She was sent usage scenarios before that showed the basic functionality of the tool and knew it from demonstration videos and meetings, but she has not tested it herself before. She found the clear cluster separation on the scatterplot interesting and would desire to see all patient IDs on the scatterplot at a time to get an overview of the patients that are grouped together. We currently show the patient IDs and scores only on demand, as these would otherwise clutter the view, especially in dense regions of the scatterplot. Therefore, we identified an export function of the patients per cluster as a better suitable solution. Furthermore, she was interested in the differentiating and characterizing features shown in the heatmap and would like to understand why the radiomic features are unevenly distributed between the two clusters. She tested the feature highlight function for the *ih.entropy* and *post PSA* values. She commented on it as being helpful in revealing where the points with high or low values are located in the cluster. Highlighting the *post PSA* on the scatterplot identified three red scatter points with high values that she was interested in comparing with the other patients to determine features that differentiate between them. She commented on the function of comparing a free selection on the scatterplot with other not selected patients as being practical. She would like to export a selection of features shown there to analyze them further and confirm their relevance. She tested this functionality also through a hypothesis to determine the characterizing and differentiating features between patients that have a low *ISUP* grade and patients that have a high *ISUP* grade in combination with patients that have the *PLEC* gene. She showed the detailed bar chart view of the differentiating features and filtered the radiomic data to see also the relevant genomic and clinical scores. Finally, she investigated the top gene mutations of the active hypothesis and would also desire to export these for further analysis. In general, she could deal with the interface and liked its interactivity. She expects it to help her effectively generate new hypotheses and insights on the data and comment on its interactivity as being of great value that cannot be perceived just through screenshots or demonstration videos.

During the progressing development of this framework, we further received feedback from domain experts after demonstrating the functionality of our framework at meetings or presentations. Our domain experts commented that the resulting interface was very clear and easy to understand. However, they do not interactively test or use the framework themself. They commented on it as being helpful for them in getting an understanding or feeling for the data and in checking a hypothesis or the role of any feature combination on an interactive visual basis. Regarding the data imputation and outlier removal steps, they were afraid that these cause bias in the data or remove outliers they are interested in. However, the outlier removal is only performed on demand and can also be applied to detect outliers and explore them further. The imputation is evaluated on the data and can be changed to a different method to see how it influences the resulting clusters or patterns. For our default preset of *t-SNE* the result is very stable on all imputation methods except the imputation through a constant, which assigns one of the patients to a different cluster. We allow users to select the complete subset of patients with no imputed values to identify and explore these patients further. In general, they commented that the framework indicates that the data consist of more aggressive tumors that show specific gene mutations and radiomic features.

CHAPTER 6

Conclusion and Future Work

In this work, we address the problem of getting insights into large, high-dimensional, and complex radiogenomic data with respect to clinical data. We investigate, design and develop a flexible interactive visual interface that combines radiogenomic data and clinical data into one framework for cancer experts and biomedical data scientists. We offer a preprocessing step to prepare the data for automated analysis and visualization. This step includes resolving data inconsistencies, handling mixed data types, imputing missingness in data values, detecting outliers in the data, and scaling feature ranges. Our cohort stratification divides patients into groups with similar radiomic, genomic, or clinical profiles to identify patterns and correlations in the data. It includes a dimensionality reduction and clustering step that we evaluate based on clustering separation scores. Based on our evaluation, we suggest the best suitable options for the data through presets. Furthermore, we allow domain experts to investigate and compare different imputations, outlier detection techniques, scaling methods, dimensionality reduction algorithms, and clustering approaches on demand. To explain the identified patterns and help our domain experts to understand the data, we determine and visualize the most characterizing and differentiating features per cluster and show feature distribution plots on demand as part of the forward analysis. In the forward analysis, we allow our domain experts to gain knowledge from the data by freely interacting with the data through exploring and comparing the characterizing and differentiating features of patient subsets or applying the processing steps and cohort stratification on only a selected patient subset for further exploration. Finally, we allow domain experts to interactively assess a hypothesis in mind on the data or to create and refine hypotheses by exploring characterizing and differentiating features and top gene mutations of the data. As part of the backward analysis, users can interactively test and identify feature thresholds for their hypotheses. They can further define any feature subset of the radiomic, genomic, and clinical data to base the analysis on and interactively investigate and compare the resulting patterns and correlations in the data. We qualitatively evaluate our framework with one of our cancer experts and by regular feedback on its progress from our domain experts. The feedback from our domain experts shows that our framework is a suitable technique to get insights into the data and supports the data sensemaking process.

In future work, we plan to apply our approach to datasets of different cancer types to assess its clinical applicability. Furthermore, using datasets with a larger sample size would confirm the scalability of our technique and might lead to more insights. To allow cancer experts to test the causality of the identified correlations or patterns through extended data examinations and analysis outside our framework, an export function for selected features, genes, patients, and clusters is helpful for them. Moreover, the visualization of gene pathways or integration of pathway data in the analysis is of interest to our cancer experts but is beyond the scope of this thesis.

List of Figures

List of Tables

Index

Glossary

Correlation Statistic relation between random variables, independent of whether this relationship is causal or not. 41

Correlative Exploration Exploration of correlations in the data independent of whether these are causal or not. 41

D'Amico Risk Statification Based on the PSA, GS, and Clinical staging values, patients are strarified into three risk groups that represent the estimated failure in five years of post treatment. 42

Genomics Gene sequencing data retrieved by analyzing DNA or RNA sequences of a genome. 41

Machine Learning Algorithms that automatically learn patterns from the data and make predictions or decisions that are not explicitly programmed. 41

Omics Biochemical assays that measure molecules from a biological sample. 41

Radiogenomics Bridging radiomic imaging features and gene sequencing data. 42

Radiomics Quantitative features extracted from medical imaging by data characterisation algorithms. 41

Sensemaking Process in which a meaning is given to the data or the analysis. 41

Unsupervised ML Algorithms that automatically identify patterns in unlabeled data through similarities and differences in the data. 42

Visual Analytics Combination of analytical reasoning with interactive visual interfaces to get insight into complex data and make effective decisions. 41

Acronyms

BCR Biochemical Recurrence. 41

CT Computed Tomography. 41

DNA Deoxyribonucleic Acid. 41

GS Gleason Score. 41

ISUP International Society of Urological Pathology. 41

LoG Laplacian of Gaussian. 42

ML Machine Learning. 41, *Glossary:* Machine Learning

MRI Magnetic Resonance Imaging. 41

PET Positron Emission Tomography. 41

PS Pathological Staging. 41

PSA Prostate-Specific Antigen. 41

PSMA Prostate-Specific Membrane Antigen. 41

RNA Ribonucleic Acid. 41

SUV Standard Uptake Value. 41

TBR Target to Blood Pool Ratio. 41

TM Tumor Margin. 41

TNM Staging Tumor Node Metastasis Staging. 41

VA Visual Analytics. 41, *Glossary:* Visual Analytics

VOI Volume Of Interest. 41

WHO World Health Organization. 41

Bibliography

[ABCA22] Ali Abbasian Ardakani, Nathalie J Bureau, Edward J. Ciaccio, and U Rajendra Acharya. Interpretation of radiomics features–A pictorial review. *Computer Methods and Programs in Biomedicine*, 215:2–19, 2022. doi:10.1016/j.cmpb.2021.106609.

[Ada23] David Adams. DNA Sequencing. `https://www.genome.gov/genetics-glossary/DNA-Sequencing`, 2023. Last access on the 2nd of February 2023.

[AFM+19] Annalisa Astolfi, Michele Fiore, Fraia Melchionda, Valentina Indio, Salvatore N Bertuccio, and Andrea Pession. BCOR involvement in cancer. *Epigenomics*, 11(7):835–855, 2019.

[AMAK19] Monerah Al-Mekhlal and Amir Ali Khwaja. A Synthesis of Big Data Definition and Characteristics. In *International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 314–322. IEEE, 2019. doi:10.1109/CSE/EUC.2019.00067.

[ANI+20] Shiva Alemzadeh, Uli Niemann, Till Ittermann, Henry Völzke, Daniel Schneider, Myra Spiliopoulou, Katja Bühler, and Bernhard Preim. Visual Analysis of Missing Values in Longitudinal Cohort Study Data. *Computer Graphics Forum (CGF)*, 39(1):63–75, 2020. doi:10.1111/CGF.13662.

[AOH+14] Paolo Angelelli, Steffen Oeltze, Judit Haasz, Cagatay Turkay, Erlend Hodneland, Arvid Lundervold, Astri J. Lundervold, Bernhard Preim, and Helwig Hauser. Interactive Visual Analysis of Heterogeneous Cohort-Study Data. *IEEE Computer Graphics and Applications*, 34(5):70–82, 2014. doi:10.1109/MCG.2014.40.

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

[AVL+14] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René

Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1):4006, 2014. doi:10.1038/ncomms5006.

[BBJ+17] Andreas Bannach, Jürgen Bernard, Florian Jung, Jörn Kohlhammer, Thorsten May, Kathrin Scheckenbach, and Stefan Wesarg. Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. In *Workshop on Visual Analytics in Healthcare (VAHC)*, pages 84–91. IEEE, 2017. doi:10.1109/VAHC.2017.8387545.

[BC87] Richard A. Becker and William S. Cleveland. Brushing Scatterplots. *Technometrics*, 29(2):127–142, 1987. doi:10.1080/00401706.1987.10488204.

[BM13] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi:10.1109/TVCG.2013.124.

[brg22] brgfx. Normal and Cancer Cell Image. https://www.freepik.com/free-vector/process-cancer-cell-development_25592548.htm, 2022. Last access on the 16th of December 2022.

[Bro08] Stuart M. Brown. Cancer Genomics. In *Essentials of Medical Genomics*, pages 328–338. John Wiley & Sons, Inc., 2008. doi:10.1002/9780470336168.

[BSM+15] Jürgen Bernard, David Sessler, Thorsten May, Thorsten Schlomm, Dirk Pehrke, and Jörn Kohlhammer. A Visual-Interactive System for Prostate Cancer Cohort Analysis. *IEEE Computer Graphics and Applications*, 35(3):44–55, 2015. doi:10.1109/MCG.2015.49.

[BTNK+19] Zuhir Bodalal, Stefano Trebeschi, Thi Dan Linh Nguyen-Kim, Winnie Schats, and Regina Beets-Tan. Radiogenomics: bridging imaging and genomics. *Abdominal Radiology*, 44(6):1960–1984, 2019. doi:10.1007/s00261-019-02028-w.

[BZA20] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, 53(3):55, 2020. doi:10.1145/3381028.

[CCW+21] Alberto Corvò, H. S. Garcia Caballero, Michel A. Westenberg, Marc A. van Driel, and Jarke J. van Wijk. Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(10):3851–3866, 2021. doi:10.1109/TVCG.2020.2990336.

[CD15] Wengen Chen and Vasken Dilsizian. PET Assessment of Vascular Inflammation and Atherosclerotic Plaques: SUV or TBR? *Journal of Nuclear Medicine*, 56(4):503–504, 2015. doi:10.2967/JNUMED.115.154385.

[CH74] Tadeusz Caliński and Joachim Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics – Theory and Methods*, 3:1–27, 1974. doi:10.1080/03610927408827101.

[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.

[CLTC$^+$21] Giuseppe Cutaia, Giuseppe La Tona, Albert Comelli, Federica Vernuccio, Francesco Agnello, Cesare Gagliardo, Leonardo Salvaggio, Natale Quartuccio, Letterio Sturiale, Alessandro Stefano, Mauro Calamia, Gaspare Arnone, Massimo Midiri, and Giuseppe Salvaggio. Radiomics and Prostate MRI: Current Role and Future Applications. *Journal of Imaging*, 7(2):34, 2021. doi:10.3390/jimaging7020034.

[CV22] Patricio Cerda and Gael Varoquaux. Encoding High-Cardinality String Categorical Variables. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1164–1176, 2022. doi:10.1109/TKDE.2020.2992529.

[CZD19] Zhangyu Cheng, Chengming Zou, and Jianwei Dong. Outlier Detection Using Isolation Forest and Local Outlier Factor. In *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, pages 161—-168. Association for Computing Machinery, 2019. doi:10.1145/3338840.3355641.

[DAB$^+$17] Laurent Dercle, Samy Ammari, Mathilde Bateson, Paul Blanc Durand, Eva Haspinger, Christophe Massard, Cyril Jaudet, Andrea Varga, Eric Deutsch, Jean-Charles Soria, and Charles Ferté. Limits of radiomic-based entropy as a surrogate of tumor heterogeneity: ROI-area, acquisition protocol and tissue site exert substantial influence. *Scientific Reports*, 7(1):7952, 2017. doi:10.1038/s41598-017-08310-5.

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

[DLH11] Ove Daae Lampe and Helwig Hauser. Interactive visualization of streaming data with Kernel Density Estimation. In *Pacific Visualization Symposium*, pages 171–178, 2011. doi:10.1109/PACIFICVIS.2011.5742387.

[DMGG15] Andrea De Mauro, Marco Greco, and Michele Grimaldi. What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, Volume 1644(1):97–104, 2015. doi:10.1063/1.4907823.

[DoNM22] MedUni Wien. Department of Nuclear Medicine. PET-MRI Imaging Scan. https://radnuk.meduniwien.ac.at/nuklearmedizin/, 2022. Last access on the 16th of December 2022.

[DWM+98] Anthony V. D'Amico, Richard Whittington, S. Bruce Malkowicz, Delray Schultz, Kenneth Blank, Gregory A. Broderick, John E. Tomaszewski, Andrew A. Renshaw, Irving Kaplan, Clair J. Beard, and Alan Wein. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. *Journal of the American Medical Association (JAMA)*, 280(11):969–974, 1998. doi:10.1001/JAMA.280.11.969.

[DXLL09] Lian Duan, Lida Xu, Ying Liu, and Jun Lee. Cluster-based outlier detection. *Annals of Operations Research*, 168(1):151–168, 2009. doi:10.1007/s10479-008-0371-9.

[EMK+21] Mateus Espadoto, Rafael Messias Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(3):2153–2173, 2021. doi:10.1109/TVCG.2019.2944182.

[ESA+21] Absalom Ezugwu, Amit Shukla, Moyinoluwa Agbaje, Adán José-García, Oyelade Olaide, and Ovre Agushaka. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, 33, 2021. doi:10.1007/s00521-020-05395-4.

[fDCP22] Centers for Disease Control and Prevent. Prostate Cancer. https://www.cdc.gov/cancer/prostate/index.htm, 2022. Last access on the 22nd of March 2023.

[GDKB17] Ievgeniia Gutenko, Konstantin Dmitriev, Arie E. Kaufman, and Matthew A. Barish. AnaFe: Visual Analytics of Image-derived Temporal Features – Focusing on the Spleen. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(1):171–180, 2017. doi:10.1109/TVCG.2016.2598463.

[GGAM12] Theresia Gschwandtner, Johannes Gaertner, Wolfgang Aigner, and Silvia Miksch. A Taxonomy of Dirty Time-Oriented Data. *International Cross-Domain Conference and Workshop on Availability, Reliability, and Security (CD-ARES)*, pages 58–72, 2012. doi:10.1007/978-3-642-32498-7_5.

[GMS+15] Laura Garrison, Juliane Müller, Stefanie Schreiber, Steffen Oeltze-Jafra Helwig Hauser, and Stefan Bruckner. DimLift: Interactive Hierarchical Data Exploration Through Dimensional Bundling. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(6):2908–2922, 2015. doi:10.1109/TVCG.2021.3057519.

[Gro20] ProPharma Group. The Role of Clinical Data and Clinical Data Science. https://www.propharmagroup.com/thought-leadership/clinical-data-and-clinical-data-science, 2020. Last access on the 22nd of March 2023.

[HCL+19] Fang Huang, Yinjie Chen, Li Li, Ji Zhou, Jian Tao, Xicheng Tan, and Guangsong Fan. Implementation of the parallel mean shift-based image segmentation algorithm on a GPU cluster. *International Journal of Digital Earth*, 12(3):328–353, 2019. doi:10.1080/17538947.2018.1432709.

[HDH+18] Chih-Yang Hsu, Mike Doubrovin, Chia-Ho Hua, Omar Mohammed, Barry L. Shulkin, Sue Kaste, Sara Federico, Monica Metzger, Matthew Krasin, Christopher Tinkle, Thomas E. Merchant, and John T. Lucas. Radiomics Features Differentiate Between Normal and Tumoral High-Fdg Uptake. *Scientific Reports*, 8(1):1–11, 2018. doi:10.1038/s41598-018-22319-4.

[HHD+21] Erling A. Hoivik, Erlend Hodneland, Julie A. Dybvik, Kari S. Wagner-Larsen, Kristine E. Fasmer, Hege F. Berg, Mari K. Halle, Ingfrid S. Haldorsen, and Camilla Krakstad. A radiogenomics application for prognostic profiling of endometrial cancer. *Communications Biology*, 4(1):1363–1374, 2021. doi:10.1038/s42003-021-02894-5.

[HK20] John T. Hancock and Taghi M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):28, 2020. doi:10.1186/s40537-020-00305-w.

[HLY+23] Qiuyuan Hu, Ke Li, Conghui Yang, Yue Wang, Rong Huang, Mingqiu Gu, Yuqiang Xiao, Yunchao Huang, and Long Chen. The role of artificial intelligence based on PET/CT radiomics in NSCLC: Disease management, opportunities, and challenges. *Frontiers in Oncology*, 13:1–12, 2023. doi:10.3389/fonc.2023.1133164.

[HNHP07] David J. Hernandez, Matthew E. Nielsen, Misop Han, and Alan W. Partin. Contemporary evaluation of the D'amico risk classification of prostate cancer. *Urology*, 70(5):931–935, 2007. doi:10.1016/j.urology.2007.08.055.

[HSC+07] Patricia Harnden, Mike D Shelley, Bernadette Coles, John Staffurth, and Malcom D Mason. Should the Gleason grading system for prostate cancer be modified to account for high-grade tertiary components? A systematic review and meta-analysis. *The Lancet Oncology*, 8:411–419, 2007. doi:10.1016/S1470-2045(07)70136-5.

[IAI+17] Mariarosaria Incoronato, Marco Aiello, Teresa Infante, Carlo Cavaliere, Anna Maria Grimaldi, Peppino Mirabelli, Serena Monti, and Marco Salvatore. Radiogenomic Analysis of Oncological Data: A Technical Survey. *International Journal of Molecular Sciences (IJMS)*, 18(4), 2017. doi:10.3390/IJMS18040805.

[IIC$^+$13] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. A Systematic Review on the Practice of Evaluating Visualization. *Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[KBZ$^+$21] Md. Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G. Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415, 2021. doi:10.1093/BIB/bbz170.

[KCH$^+$03] Won Kim, Byoung-Ju Choi, Eui Hong, Soo-Kyung Kim, and Doheon Lee. A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, 7:81–99, 2003. doi:10.1023/A:1021564703268.

[KMSZ09] Daniel A. Keim, Florian Mansmann, Andreas Stoffel, and Hartmut Ziegler. Visual Analytics. In *Encyclopedia of Database Systems*, pages 3341–3346. Springer US, 2009. doi:10.1007/978-0-387-39940-9_1122.

[LBI$^+$12] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(9):1520–1536, 2012. doi:10.1109/TVCG.2011.279.

[LL17] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Conference on Neural Information Processing Systems (NIPS)*, 2017. doi:10.48550/arXiv.1705.07874.

[LLD$^+$17] Philippe Lambin, Ralph T.H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E.C. de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T.H.M. Larue, Aniek J.G. Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12):749–762, 2017. doi:10.1038/nrclinonc.2017.141.

[LLWF21] Hongfu Liu, Jun Li, Yue Wu, and Yun Fu. Clustering With Outlier Removal. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2369–2379, 2021. doi:10.1109/TKDE.2019.2954317.

[LS09] Vinata B. Lokeshwar and Marie G. Selzer. Hyaluronidase: Both a tumor promoter and suppressor. In Robert Stern, editor, *Hyaluronan in Cancer Biology*, pages 189–206. Academic Press, 2009. doi:10.1016/B978-012374178-3.10011-0.

[LSKS10] Alexander Lex, Marc Streit, Ernst Pieter Christiaan Kruijff, and Dieter Schmalstieg. Caleydo: Design and Evaluation of a Visual Analysis Framework for Gene Expression Data in its Biological Context. In *Proceedings of IEEE Pacific Visualization Symposium*, pages 57–64. IEEE, 2010. doi:10.1109/PACIFICVIS.2010.5429609.

[LSS+12] Alexander Lex, Marc Streit, Hans-Jörg Schulz, Christian Partl, Dieter Schmalstieg, Peter Park, and Nils Gehlenborg. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum (CGF)*, 31(3):1175–1184, 2012. doi:10.1111/j.1467-8659.2012.03110.x.

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *International Conference on Data Mining*, volume 8, pages 413–422. IEEE, 2008. doi:10.1109/ICDM.2008.17.

[Lu20] Haw-minn Lu. Quasi-orthonormal Encoding for Machine Learning Applications. In *Proceedings of the 19th Python in Science Conference*, pages 11–17. SCIPY, 2020. doi:10.25080/Majora-342d178e-002.

[LXNR19] Ruijiang Li, Lei Xing, Sandy Napel, and Daniel Rubin, editors. *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.

[LY15] Feng-Mei Lu and Zhen Yuan. PET/SPECT molecular imaging in clinical neuroscience: recent advances in the investigation of CNS diseases. *Quantitative Imaging in Medicine and Surgery*, 5(3), 2015. doi:10.3978/j.issn.2223-4292.2015.03.16.

[Mac22] Ruairi J Mackenzie. DNA vs. RNA – 5 Key Differences and Comparison. https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719, 2022. Last access on the 1st of February 2023.

[MAF14] Peshawa Muhammad Ali and Rezhna Faraj. Data Normalization and Standardization: A Technical Report. *The Machine Learning Lab*, 1(1):1–6, 2014. doi:10.13140/RG.2.2.28948.04489.

[MAM+23] Julie Malet, Julien Ancel, Abdenasser Moubtakir, Dimitri Papathanassiou, Gaëtan Deslée, and Maxime Dewolf. Assessment of the Association between Entropy in PET/CT and Response to Anti-PD-1/PD-L1 Monotherapy in Stage III or IV NSCLC. *Life*, 13(4):1051, 2023. doi:10.3390/life13041051.

[Med23] John Hopkins Medicine. Radical Prostatectomy. https://www.hopkinsmedicine.org/health/

treatment-tests-and-therapies/radical-prostatectomy, 2023. Last access on the 22nd of March 2023.

[MHHWM18] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018. doi:10.1109/ISEMANTIC.2018.8549751.

[MML+20] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to Radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, 2020. doi:10.2967/jnumed.118.222893.

[MSB+15] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. Comparison of different Methods for Univariate Time Series Imputation in R. *Computing Research Repository (CoRR)*, 2015. doi:10.48550/arXiv.1510.03924.

[MSO+20] Juliane Müller, Matthaeus Stoehr, Alexander Oeser, Jan Gaebel, Marc Streit, Andreas Dietz, and Steffen Oeltze-Jafra. A visual approach to explainable computerized clinical decision support. *Computers & Graphics (CAG)*, 91:1–11, 2020. doi:10.1016/j.CAG.2020.06.004.

[Mun09] Tamara Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 15(6):921–928, 2009. doi:10.1109/TVCG.2009.111.

[MV98] J.B. Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998. doi:10.1016/S1361-8415(01)80026-8.

[MWH+20] Eric Mörth, Kari Strøno Wagner-Larsen, Erlend Hodneland, Camilla Krakstad, Ingfrid S. Haldorsen, Stefan Bruckner, and Noeska Natasja Smit. RadEx: Integrated Visual Exploration of Multiparametric Studies for Radiomic Tumor Profiling. *Computer Graphics Forum (CGF)*, 39(7):611–622, 2020. doi:10.1111/CGF.14172.

[Nar20] Naveen Naidu Narisetty. Chapter 4 - Bayesian model selection for high-dimensional data. In *Principles and Methods for Data Science*, volume 43 of *Handbook of Statistics*, pages 207–248. Elsevier, 2020. doi:10.1016/bs.host.2019.08.001.

[NCI21] National Cancer Institute NCI. Cancer Background Information. https://www.cancer.gov/, 2021. Last access on the 16th of December 2022.

[Ng17] Sok Choo Ng. Principal component analysis to reduce dimension on digital image. In *International Conference on Advances in Information Technology*, volume 111, pages 113–119, 2017. doi:10.1016/j.procs.2017.06.017.

[NGA+09] Issam El Naqa, Perry W. Grigsby, Aditya Apte, Elizabeth Kidd, Eric Donnelly, Divya Khullar, Summer Chaudhari, Deshan Yang, Martin Schmitt, Richard Laforest, Wade L. Thorstad, and Joseph O. Deasy. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognition*, 42(6):1162–1171, 2009. doi:10.1016/j.patcog.2008.08.011.

[NHG19] S. Nusrat, T. Harbig, and Nils Gehlenborg. Tasks, Techniques, and Tools for Genomic Data Visualization. *Computer Graphics Forum*, 38(3):781–805, 2019. doi:10.1111/cgf.13727.

[NNH+14] Quang Vinh Nguyen, Guy Nelmes, Mao Lin Huang, Simeon Simoff, and Daniel Catchpoole. Interactive Visualization for Patient-to-Patient Comparison. *Genomics & Informatics*, 12(1):21–34, 2014. doi:10.5808/GI.2014.12.1.21.

[OH21] Oyeyemi Osho and Sungbum Hong. 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*, 10:294–308, 2021. IJERTV10IS040188.pdf.

[oPC23] ZERO The End of Prostate Cancer. Genomics Background Information. `https://zerocancer.org/`, 2023. Last access on the 30th of January 2023.

[oSA23] American College of Surgeons ACS. Cancer Staging Systems. `https://www.facs.org/quality-programs/cancer-programs/american-joint-committee-on-cancer/cancer-staging-systems/`, 2023. Last access on the 22nd of March 2023.

[PPL+19] Andreas S. Panayides, Marios S. Pattichis, Stephanos Leandrou, Costas Pitris, Anastasia Constantinidou, and Constantinos S. Pattichis. Radiogenomics for Precision Medicine With a Big Data Analytics Perspective. *IEEE Journal of Biomedical and Health Informatics*, 23(5):2063–2079, 2019. doi:10.1109/JBHI.2018.2879381.

[PWKJ08] Bernd J. Pichler, Hans F. Wehrl, Armin Kolb, and Martin S. Judenhofer. Positron Emission Tomography/Magnetic Resonance Imaging: The Next Generation of Multimodality Imaging? *Seminars in Nuclear Medicine*, 38(3):199–208, 2008. doi:10.1053/j.semnuclmed.2008.02.001.

[QLN+19] Zhonglin Qu, Chng Wei Lau, Quang Vinh Nguyen, Yi Zhou, and Daniel R Catchpoole. Visual Analytics of Genomic and Cancer Data: A Systematic Review. *Cancer Informatics*, 18:1–19, 2019. doi:10.1177/1176935119835546.

[RAW+16] Alexander Rind, Wolfgang Aigner, Markus Wagner, Silvia Miksch, and Tim Lammarsch. Task Cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. *Information Visualization*, 15(4):288–300, 2016. doi:10.1177/1473871615621602.

[Raw19] Prashanth Rawla. Epidemiology of Prostate Cancer. *World Journal of Oncology*, 10:63–89, 2019. doi:10.14740/wjon1191.

[RD00] Erhard Rahm and xx g Do, Hong. Data Cleaning: Problems and Current Approaches. *Data Engineering Bulleting*, 23(4):3–13, 2000. betterevaluation.org/sites/default/files/data-cleaning.pdf.

[RG19] Frédéric Ros and Serge Guillaume. A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications*, 128:96–108, 2019. doi:10.1016/j.eswa.2019.03.031.

[ROC+17] Sylvain Reuzé, Fanny Orlhac, Cyrus Chargari, Christophe Nioche, Elaine Limkin, Franşois Riet, Alexandre Escande, Christine Haie-Meder, Laurent Dercle, Sébastien Gouy, Irène Buvat, Eric Deutsch, and Charlotte Robert. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. *Oncotarget*, 8(26):43169–43179, 2017. doi:10.18632/oncotarget.17856.

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

[RTS+19] Sarah Rice, Maria Tselepi, Antony Sorial, Guillaume Aubourg, Colin Shepherd, David Almarza, David Deehan, Louise Reynard, and John Loughlin. Prioritization of PLEC and GRINA as Osteoarthritis Risk Genes Through the Identification and Characterization of Novel Methylation Quantitative Trait Loci. *Arthritis and Rheumatology*, 71, 2019. doi:10.1002/art.40849.

[Rux06] Graeme D. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4):688–690, 2006. doi:10.1093/beheco/ark016.

[RvdHD+15] Renata Georgia Raidou, Uulke A van der Heide, Cuong Viet Dinh, Ghazaleh Ghobadi, Jesper Follsted Kallehauge, Marcel Breeuwer, and Anna Vilanova. Visual Analytics for the Exploration of Tumor Tissue

Characterization. *Computer Graphics Forum (CGF)*, 34(3):11–20, 2015. doi:10.1111/CGF.12613.

[RZ19] Fakhitah Ridzuan and Wan Mohd Nazmee Zainon. A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science*, 161:731–738, 2019. doi:10.1016/j.procs.2019.11.177.

[Sch22] Matthew Schmitz. Understanding the D'Amico Classification System for Prostate Cancer. `https://www.verywellhealth.com/damico-classification-system-for-prostate-cancer-2782233`, 2022. Last access on the 22nd of March 2023.

[SCR19] Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, 7:1–12, 2019. doi:10.1177/2050312118822912.

[SDE+16] John R. Srigley, Brett Delahunt, Lars Egevad, Hemamali Samaratunga, John Yaxley, and Andrew J. Evans. One is the new six: The International Society of Urological Pathology (ISUP) patient-focused approach to Gleason grading. *Canadian Urological Association Journal*, 10:339–341, 2016. doi:10.5489/cuaj.4146.

[SEK03] Michael Steinbach, Levent Ertöz, and Vipin Kumar. 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.

[SGB+21] Camilla Scapicchio, Michela Gabelloni, Andrea Barucci, Dania Cioni, Luca Saba, and Emanuele Neri. A deep look into radiomics. *La Radiologia medica*, 126, 2021. doi:10.1007/s11547-021-01389-x.

[SGPLB13] Michael P Schroeder, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Visualizing multidimensional cancer genomics data. *Genome Medicine*, 5(1):9, 2013. doi:10.1186/gm413.

[SGTB13] Julien Soler, Laurent Gaubert, Fabien Tencé, and Cédric Buche. Data Clustering and Similarity. In *International Florida Artificial Intelligence Research Society Conference*, volume 26, pages 492–495. Association for the Advancement of Artificial Intelligence (AAAI), 2013. dblp.org/rec/conf/flairs/SolerTGB13.

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

[SJG+22] Sanjay Saxena, Biswajit Jena, Neha Gupta, Suchismita Das, Deepaneeta Sarmah, Pallab Bhattacharya, Tanmay Nath, Sudip Paul, Mostafa M. Fouda, Manudeep Kalra, Luca Saba, Gyan Pareek, and Jasjit S. Suri. Role of Artificial Intelligence in Radiogenomics for Cancers in the Era of Precision Medicine. *Cancers*, 14(12):2860, 2022. doi:10.3390/cancers14122860.

[SKT+19] Lara Schneider, Tim Kehl, Kristina Thedinga, Nadja Liddy Grammes, Christina Backes, Christopher Mohr, Benjamin Schubert, Kerstin Lenhof, Nico Gerstner, Andreas Daniel Hartkopf, Markus Wallwiener, Oliver Kohlbacher, Andreas Keller, Eckart Meese, Norbert M. Graf, and Hans-Peter Lenhof. ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification. *Bioinformatics*, 35:5171–5181, 2019. doi:10.1093/bioinformatics/btz302.

[SNHS13] Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. A Design Space of Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013. doi:10.1109/TVCG.2013.120.

[SRM+22] Mamello Sekhoacha, Keamogetswe Riet, Paballo Motloung, Lemohang Gumenku, Ayodeji Adegoke, and Samson Mashele. Prostate Cancer Review: Genetics, Diagnosis, Treatment Options, and Alternative Approaches. *Molecules*, 27(17):5730, 2022. doi:10.3390/molecules27175730.

[SRY+21] Lin Shui, Haoyu Ren, Xi Yang, Jian Li, Ziwei Chen, Cheng-Yi Cheng, Hong Zhu, and Pixian Shui. 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.

[Su10] Li-Ming Su. *Early Diagnosis and Treatment of Cancer Series: Prostate Cancer*. W.B. Saunders, 2010. doi:10.1016/B978-1-4160-4575-5.50013-X.

[SWC+09] Jonathan Sterne, Ian White, John Carlin, Michael Spratt, Patrick Royston, Michael Kenward, Angela Wood, and James Carpenter. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal (BMJ)*, 338:b2393, 2009. doi:10.1136/bmj.b2393.

[TFH11] Cagatay Turkay, Peter Filzmoser, and Helwig Hauser. Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011. doi:10.1109/TVCG.2011.178.

[THFM14] Peter Trebuňa, Jana Halčinová, Milan Fil'o, and Jaromir Markovic. The importance of normalization and standardization in the process of clustering. In *International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, volume 12, pages 381–385. IEEE, 2014. doi:10.1109/SAMI.2014.6822444.

[TLM08] Du-Yih Tsai, Yongbum Lee, and Eri Matsuyama. Information Entropy Measure for Evaluation of Image Quality. *Journal of Digital Imaging*, 21:338–347, 2008. doi:10.1007/s10278-007-9044-5.

[TLS$^+$14] Cagatay Turkay, Alexander Lex, Marc Streit, Hanspeter Pfister, and Helwig Hauser. Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. *IEEE Computer Graphics and Applications*, 34(2):38–47, 2014. doi:10.1109/MCG.2014.1.

[TZH$^+$20] Lennart Tautz, Hannu Zhang, Markus Hüllebrand, Matthias Ivantsits, Sebastian Kelle, Titus Kuehne, Volkmar Falk, and Anja Hennemuth. Cardiac radiomics: an interactive approach for 4D data exploration. *Current Directions in Biomedical Engineering (CDBME)*, 6(1):1–6, 2020. doi:10.1515/CDBME-2020-0008.

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

[VDM14] Laurens Van Der Maaten. Accelerating T-SNE Using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014. jmlr.org/papers/v15/vandermaaten14a.html.

[VVTT20] Alexandros Vamvakas, Katerina Vassiou, Dimitra Tsivaka, and Ioannis Tsougos. Decision support systems in breast cancer. *Cancers (Basel)*, pages 319–327, 2020. doi:10.1016/B978-0-12-819178-1.00031-9.

[WBKA21] Tamsin Wesley, Stuart Berzins, George Kannourakis, and Nuzhat Ahmed. The attributes of plakins in cancer and disease: perspectives on ovarian cancer progression, chemoresistance and recurrence. *Cell Communication and Signaling*, 19(1):55, 2021. doi:10.1186/s12964-021-00726-x.

[WHO22] World Health Organization WHO. World Cancer Statistics. `https://gco.iarc.fr/`, 2022. Last access on the 16th of December 2022.

[WLH$^+$22] Le Wang, Bin Lu, Mengjie He, Youqing Wang, Zongping Wang, and Lingbin Du. Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019. *Frontiers in Public Health*, 10, 2022. doi:10.3389/fpubh.2022.811044.

[XT15] Dongkuan Xu and Yingjie Tian. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2:165–193, 2015. doi:10.1007/s40745-015-0040-1.

[XWY$^+$21] Ruizhi Xiang, Wencan Wang, Lei Yang, Shiyuan Wang, and Chaohan Xu. A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics*, 12, 2021. doi:10.3389/fgene.2021.646936.

[YJY+17] Lina Yu, Hengle Jiang, Hongfeng Yu, Chi Zhang, Josiah Mcallister, and Dandan Zheng. iVAR: Interactive visual analytics of radiomics features from large-scale medical images. In *International Conference on Big Data*, pages 3916–3923. IEEE, 2017. doi:10.1109/BigData.2017.8258398.

[ZCP+21] Mario Zanfardino, Rossana Castaldo, Katia Pane, Ornella Affinito, Marco Aiello, Marco Salvatore, and Monica Franzese. MuSA: a graphical user interface for multi-OMICs data integration in radiogenomic studies. *Scientific Reports*, 11(1):1550, 2021. doi:10.1038/s41598-021-81200-z.

[ZLVL20] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. *Radiology*, 295(2):328–338, 2020. doi:10.1148/RADIOL.2020191145.

[ZZ21] YanPing Zhao and XiaoLai Zhou. K-means Clustering Algorithm and Its Improvement Research. *Journal of Physics: Conference Series*, 1873(1):012074, 2021. doi:10.1088/1742-6596/1873/1/012074.