Informatics

# Stratifizierung und Visualisierung von Hepatozellulärem Karzinom Patienten

BACHELORARBEIT

zur Erlangung des akademischen Grades

**Bachelor of Science**

im Rahmen des Studiums

**Medizinische Informatik**

eingereicht von

**Mark-Christian Bront**
Matrikelnummer 11928262

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Assistant Prof. Dr. Renata Georgia Raidou

Wien, 30. September 2023

| | |
|---|---|
| Mark-Christian Bront | Renata Georgia Raidou |

TU WIEN Informatics

# Stratification and Visualization of Hepatocellular Carcinoma Patients

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

**Bachelor of Science**

in

**Medical Informatics**

by

**Mark-Christian Bront**
Registration Number 11928262

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr. Renata Georgia Raidou

Vienna, 30th September, 2023

Mark-Christian Bront     Renata Georgia Raidou

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

# Erklärung zur Verfassung der Arbeit

Mark-Christian Bront

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, 30. September 2023

Mark-Christian Bront

# Acknowledgements

First and foremost I wish to thank my supervisor for agreeing to supervise this thesis guiding me in the process of developing the framework and giving me pointers for interesting ideas which I later implemented in the notebook.

I also want to thank my colleagues to whom I've presented the idea and gave their feedback on "what would be interesting and cool".

Last but not least I want to thank my parents for allowing me to study at TU Wien and supporting me along the way.

# Kurzfassung

Ziel dieser Arbeit ist es, die Eignung von radiomischen Merkmalen von Patienten mit hepatozellulärem Karzinom (HCC) Patienten für die Stratifizierung und Visualisierung der Leber und des Tumors. Nachdem wir die Merkmale berechnet haben, skalieren wir diese in einem gemeinsamen Bereich von 0 bis 1000. Wir führen ein Clustering der Leber- und Tumormerkmale durch, nachdem sie mit verschiedenen verschiedene Techniken zur Dimensionalitätsreduktion. Wir diskutieren und vergleichen die untersuchten Methoden zur Dimensionalitätsreduktion, um eine sinnvolle Stratifizierung der Patienten basierend auf ihren leber- und tumorspezifischen Daten. Schließlich kombinieren wir diese Cluster-Deskriptoren mit dem entsprechenden klinischen Datensatz der verschiedene gesundheitliche und demografische Merkmale enthält, die die Eigenschaften der Patienten beschreiben. Mit Hilfe von Visualisierungen untersuchen wir mögliche Zusammenhänge zwischen den Merkmalen und unterstützen die Ableitung von Erkenntnissen über die zugrunde liegenden Datenmuster zu unterstützen. Diese abgeleiteten Erkenntnisse sollen Ärzten und Klinikern in Zukunft bei der Schätzung der bei der Abschätzung der Gesamtüberlebensrate für neue Patienten. Zu den erwähnten Erkenntnissen gehört, dass dass die Altersgruppe eines Patienten seine Gesamtüberlebensrate drastisch beeinflussen kann. Unter Zusammenfassend lässt sich sagen, dass wir eine Möglichkeit für Ärzte und Kliniker geschaffen haben, ihre Patienten anhand von Merkmalen aus den volumetrischen Daten und den klinischen Daten zu stratifizieren und weitere Erkenntnisse zu gewinnen, die für die Schätzung des Gesamtüberlebens dieser Patienten hilfreich sind.

# Abstract

This thesis aims to investigate the suitability of radiomics features of Hepatocellular Carcinoma (HCC) patients for the stratification and visualization of the liver and the tumour. After we compute the features, we scale these into a common range of 0 to 1000. We perform clustering on the liver and tumour features after they are reduced via different dimensionality reduction techniques. We discuss and compare the investigated dimensionality reduction methods used to derive a meaningful stratification of the patients based on their liver and tumour-specific data.

Finally, we combine these cluster descriptors with the corresponding clinical dataset containing diverse health and demographic features describing the patient's characteristics. We employ visualizations to investigate possible connections between features to support the derivation of insights with regard to the underlying data patterns. These derived insights are anticipated to help doctors and clinicians in the future when estimating the overall survival metric for new patients. Among the mentioned insights, we found that the age group of a patient can drastically influence their overall survival metric. In conclusion, we created a way for doctors and clinicians to stratify their patients based on features extracted from theirvolumetric data and their clinical data, and derive further insights helpful for estimating the overall survival of these patients.

# Contents

CHAPTER 1

# Introduction

Interpretable visualizations and communication of insights derived from them are the key components of a high-quality diagnosis. As such, understanding data also plays a key role in the first place. In this thesis, we present the steps of gaining knowledge into the stratification of Hepatocellular Carcinoma patients employing features derived from Computed Tomography (CT) scans. We use these to cluster patients, combining new variables derived from the field of radiomics with existing clinical data. Lastly, we summarize the available data via plots, thus gaining new insights. We anticipate that the resulting visuals will help clinicians, data scientists, and potentially laypeople to understand the data. By having these interpretable visuals, we could improve the diagnosis process and support better predictions about the patients.

The first step towards a diagnosis is data collection. In this case, a clinician may consider electronic health records (EHRs) as data sources. The Cancer Imaging Archive (TCIA) [CVS$^+$13] provides us with the clinical data of the patients from the original study of Morshid et al. [MEK$^+$19]. We use the dataset of Mowad et al. [MFM$^+$21]. The dataset contains lots of discrete variables that describe the patient's health history and demographics. For new patients, certain features may not be readily available because the data has not been entered correctly or has not been measured via other diagnostics up to that point. As such, because of missing data, we can not perform a comprehensive diagnosis for those new patients. The dataset also has Computed Tomography (CT) scans together with the segmentations of the liver, vessels and tumour. Researchers can use these volumes as a basis for extracting new variables or features through the use of radiomics, as Lambin et al. [LRVL$^+$12] described. The research question of this thesis is: How can visualization support the derivation of insights when stratifying a cohort of HCC patients?

In this thesis, we aim to investigate the suitability of radiomics features of Hepatocellular Carcinoma (HCC) patients for their stratification. We consider the result of this stratification, i.e., the patient clusters, as derived features that we combine with the corresponding clinical data and later visualize to provide new insights about the diagnosis and prognosis of the patients. By doing so, we make insights visible, and these insights have the potential to help doctors make better decisions in treatment planning and estimating overall survival.

We present notebooks [Git], which represent the implementation of the aforementioned methodology, as a proof of concept towards an exploratory analysis of radiomics data and other features extracted from medical records. We intended the notebooks to be used in analysing HHC patients, however, other researchers can apply the same approaches to other afflictions and illnesses; therefore offering a starting point for visual investigation of other patient cohorts.

CHAPTER 2

# Related Work

Many researchers have investigated the usage of radiomics for extracting features over the years and have obtained promising results. Lambin et al. [LRVL$^+$12] initiated the overall journey of radiomics in 2012. They noted the ways medical imaging had progressed up until that point as advancements in imaging hardware and agents, standardised protocol procedures, and advancements in imaging analysis through the use of computer-assisted detection (CAD) systems. They describe the overall workflow and lay the groundwork for quantitative analysis. The fundamental hypothesis of radiomics is to find prognostic markers using features available in imaging data. They also suggest linking radiomics to radio-genomics, as genomic features may be expressed and linked to captured images. In 2012 Kumar et al. [KGB$^+$12] also investigated the use and challenges of radiomics based on CT scans of non-small-cell lung cancer patients, and found that these "*can be used to build descriptive and predictive models relating image features to phenotypes or gene–protein signatures*" [KGB$^+$12]. They also divided radiomics into multiple processes, such as:

- acquisition and reconstruction of images
- segmentation and rendering of images
- feature extraction and feature qualification
- databases and data sharing of data
- ad hoc informatics analyses

Each of the processes faces challenges of its own. For example, image acquisition and reconstruction face the challenge of varying image resolution between different hardware producers, which are then used to acquire medical images. They noted that such issues can lead to difficulties when comparing results across multiple institutions, which use different scanners. Segmentation challenges include the high interreader variability when done manually. Even in the case of semiautomatic procedures like region-growing-based algorithms [AB94], these are dependent on an operator to specify a seed point in the volume of interest, thus introducing once again interobserver variability into the segmentations.

Various papers have investigated features that may be useful for the purposes of classifying tumours. One example would be the recent work of Lysdahlgaard [Lys22], who conducted a comparison of radiomics features extracted from CT images of healthy liver tissue and tumour tissue. From here, we may find features relating to the tumour shape, intensity values, and texture. Specifically, they follow the categories of semantic features like size, shape, and location. Agnostic features can be split into two subcategories: morphological of the first order (e.g., volume, size, lesion diameter) or higher order (e.g., Minkowski functionals, fractal dimensions), and the statistical ones. Under the subcategory of statistical features, there exist first-order ones like mean, median, standard deviation, kurtosis, skewness; second-order textures like the gray-level co-occurrence matrix (GLCM), gray-level neighborhood difference matrix (GLNDM), gray-level run length matrix (GLRLM), gray-level size zone (GLSZM); and lastly higher-order like the Wavelet and Laplacian of Gaussian transformations. The taxonomy for the statistical features can be found in Figure 2.1.

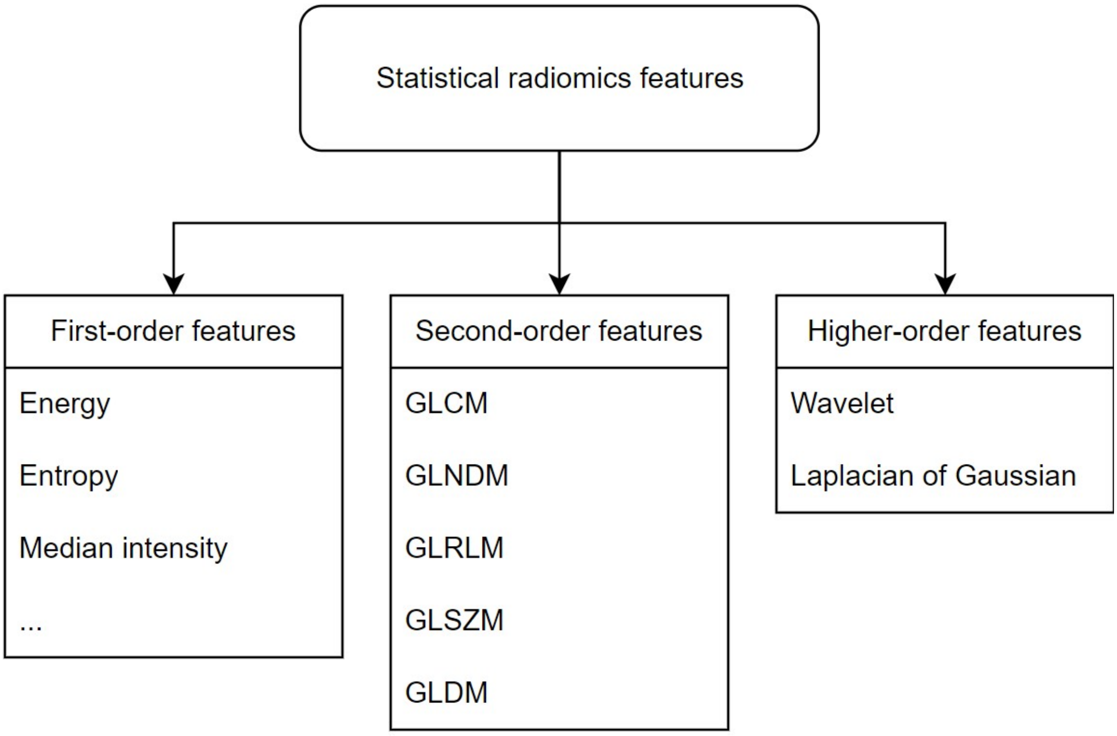


Figure 2.1: Taxonomy of statistical radiomics features.

Moreover, Yeung et al. [YCW+20] extracted the gross tumour volume with the help of co-registration with diagnostic tri-phasic contrast CT, which is the mask applied to the CT volumes for further feature extraction. In the same year, Rela et al. [RSR20]authored a review of liver tumour segmentations and their classification, and also looked into the feature extraction step. Here they also make use of texture, shape, and size features as seen in previous work. The gray-level co-occurrence matrix is also mentioned alongside other statistical features like mean, standard deviation and kurtosis. conducting a state-of-the-art report regarding radiomics, Wei et al. [WJG+20] have shown the developments of radiomics and future opportunities. Finally, Fu et al. [FWZ+19] used "peel-off" features under the hypothesis that tumours grow in size from the inside to the outside. These features operate on layers of the tumour, and according to Fu et al., the entropy measured based on this approach demonstrated relevance for predicting outcomes without further tumour development after resection.

When diagnosing a patient and choosing the correct treatment, one must also consider tumour metastases as a key factor. In this sense, Fiz et al. [FVG+20] made a systematic review of 32 papers regarding the radiomics of liver metastases. From the investigated papers, CT was used the most (n=19), followed by magnetic resonance imaging (MRI) (n=8), positron emission tomography (PET) (n=3) and lastly, multiple imaging modalities such as CT combined with MRI and MRI combined with PET respectively (n=2). They found that these radiomics features provide accurate predictions of chemotherapy treatments and the patient's survival. All of the features found are also available in the Pyradiomics [VGFP+17] library, which is publicly available.

Echegaray et al. [EGS+15] have investigated the robustness of certain features with respect to the delivered segmentation. They conducted this study because they were also interested in obtaining core samples of tumours by extracting said features from a maximal circle within the delineations Among the features used were also the Haralick features based on the gray-level co-occurrence matrix. Building upon this work are Bakr et al. [BES+17], who investigate microvascular invasions in HCC patients. They employed delta radiomics, which they obtained as an absolute difference and the ratio of all radiomics features from all pairs of imaging phases. Going even further with the idea of higher-order spectra, Acharya et al. [ASR+12] investigate fatty liver disease via ultrasound imaging. Ultrasound-derived features are out of scope for this thesis and we do not discuss them further.

Machine learning and artificial intelligence have experienced large developments over the past years and hardware that can support research has become more accessible than before. Using them for developing systems that help in diagnosing patients has also become more accessible. Ahmad et al. [ADQY19] presented the use of deep learning methods, such as deep Neural Networks [GBC16] and Convolutional Neural Networks [GBC16], for segmenting the liver volumes.Training such models comes with a high computational cost and time investment. However, once training is complete, the proposed model can speed up the diagnostics process and help experts in diagnosing patients and making decisions. They tested their proposed Convolutional Neural Network model for liver segmentation (CNN-LivSeg) on the MICCAI-SLiver07 dataset [HVGS+09] and achieved a mean accuracy of 0.9725, specificity of 0.9904, and sensitivity of 0.9652. As our data already has segmentations of the liver annotated by experts, we will not need such a model for our purposes. However, it is still worth mentioning that methods exist for extracting segmentations, which can be used in the feature extraction pipeline as masks if needed.

Liver classification has also been investigated by Gunasundari et al. [GA12], where Fast Discrete Curvelet Transform (FDCT), biorthogonal wavelet transform and gray-level co-occurrence matrix were used to classify volumes based on their texture. Even before the explosion of computational resources, Chen et al. [CCC+98] made an automatic system in 1988 that segments the liver and afterwards distinguishes healthy livers from those having hepatoma or hemageoma.

Mörth et al. [MWLH+20] investigated a visual analytics approach for visualizing radiomics data, similar to the one of Raidou et al. [RvdHD+15]. A difference however is that Mörth et al. use a 1D t-distributed Stochastic Neighbor Embedding (t-SNE) [HR02] on their high-dimensional radiomic tumour features in order to pair this with a clinical meaningful parameter, building a 2D visualization of the patient cohort. Mörth et al. have also investigated image-centric cohort visualization approaches of Steenwijk et al. [SMB+10], where linked views are employed, connecting scatterplots and parallel coordinate plots to imaging data of each patient. The work of Klemm et al. [KOJL+14] takes an epidemiological approach using segmentations and hypothesis formulation via model-based visualizations. Jönnson et al. [JBF+19] use imaging data as well as clinical data in their interactive visual environment without using radiomic features. Preim et al. [PL20] conducted a survey, where they provided an overview of the field of visual analytics for the public health sector. Angelelli et al. [AOH+14] focus on a visual analytics tool based on heterogeneous cohort data in order to generate and validate hypotheses. Eckelt et al. [EAZ+19] present TourDino, which aims to help users verify hypotheses and confirm insights gained from the statistical analysis of tabular data. Bladder Runner is another visual analytics tool of Raidou et al. [RCMA+18], which enables the investigation of individual patients as well as the cohort of prostate cancer radiotherapy patients. Bernard et al. [BSM+15] took a data-centred approach and built a tool for providing an overview of large patient sets, which also includes guidance and overall reduces the time needed for the analytical workflow. None of these approaches is suitable or readily extensible to HCC patient cohorts.

CHAPTER 3

# Methodology

In this section, we discuss the overall procedures applied in the pipeline of this thesis and the reasoning behind them. We go over the data preprocessing step, where the organ and tumour volumes are extracted from the dataset shown in the "Clean data and choose subset" step in Figure 3.1. Afterwards, we discuss the chosen radiomics features which are extracted from the aforementioned volumes and the clinical features from the provided clinical dataset as seen in the steps "Radiomics extractor" for the volumetric data processing and "Choose subset of features" for the clinical data in Figure 3.1. Possible procedures for scaling (Figure 3.1, step "Normalization") and dimensionality reduction(Figure 3.1, step "Dimensionality reduction") are also discussed with regard to their upsides and downsides. These are important to consider for our application, as the data may be transformed in an unforeseen manner, reducing the amount of information retained and hindering our capability to find key connections between features that may have led to new insights. Clustering of the patient cohort (Figure 3.1, step "Agglomerative Clustering") based on the reduced radiomics features is done in order to synthesise new descriptors for the liver cluster and tumour cluster correspondence. It is once again important to understand the clustering process and its characteristics as these may change the cluster labels assigned to each patient, leading to different groupings in the later visualization. Last but not least, visualizing the results (Figure 3.1, step "Visualization") provides new insights and presents our data. In doing so, the application aids us in seeing possible connections between features and synthesized cluster labels. Each step of the pipeline is discussed in further detail in its own section. The overall pipeline is shown in Figure 3.1.

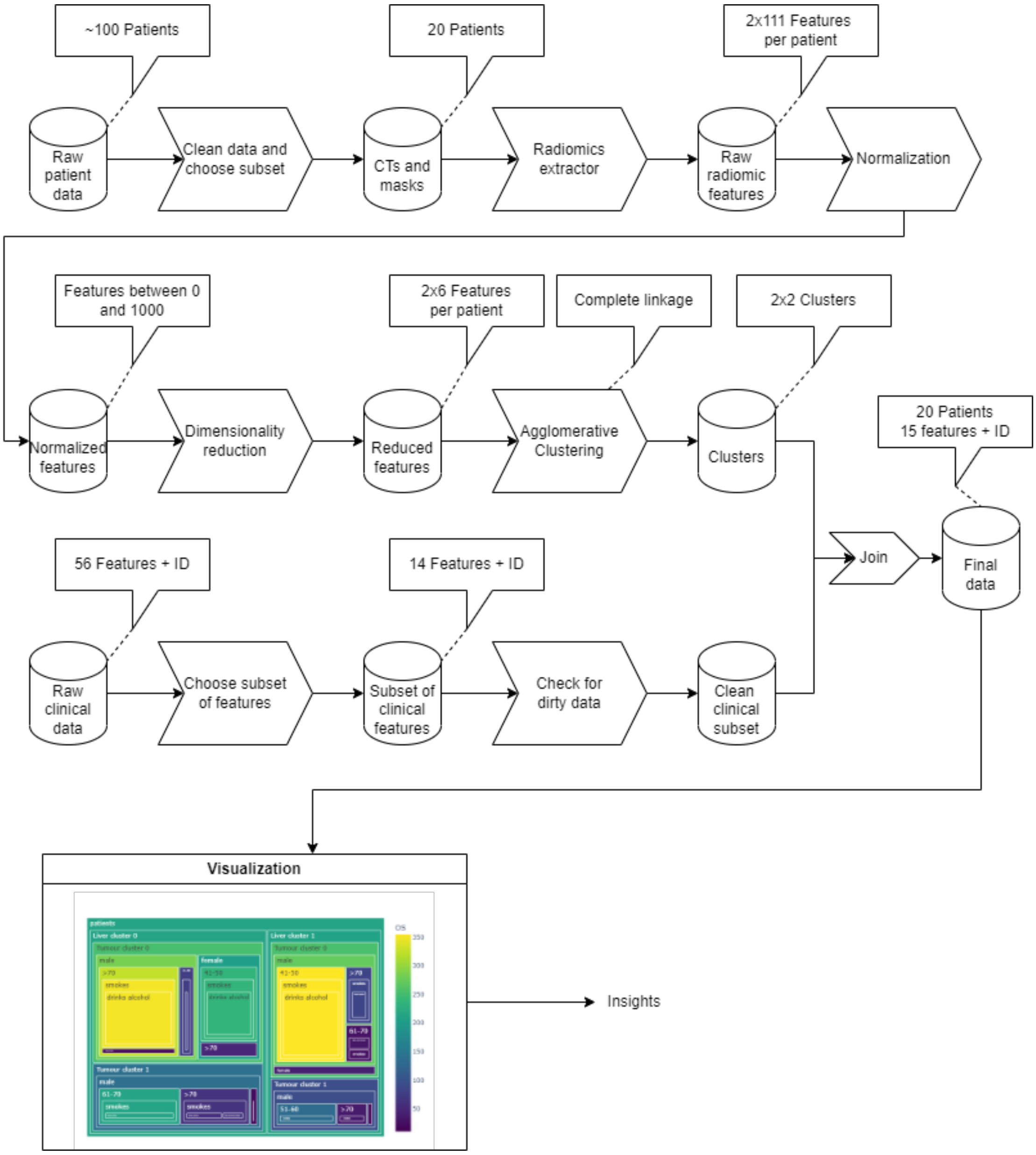


Figure 3.1: Pipeline used for data processing and insight derivation. The first branch handles the volumetric data, which starts with the "Clean data and choose subset" step and ends with the "Agglomerative clustering" step. The second branch handles the clinical data, which starts with the "Choose subset of features" step and ends the the "Check for dirty data" step. The results of both branches are joined based on the patient IDs and then visualized in the "Visualization" step, resulting into "Insights".

## 3.1 Preprocessing

We acquired the original data [MFM$^+$21] from The Cancer Imaging Archive [CVS$^+$13], which corresponds to the original publication by Morshid et al. [MEK$^+$19]. The preprocessing mainly consisted of extracting the liver and tumour segmentations from the dataset and their corresponding CT scans. This task was done in a semi-automated way as the filenames were not always standardised and as such needed inspection for each patient to ensure correspondence. There are in total 105 patients, from which we selected the first 20 where the extraction was possible. The approach is, though, not limited to 20 patients and scales well throughout all the steps of the pipeline for more than 20 patients. Another part of the preprocessing step required finding the corresponding rows of the patients in the clinical dataset and perform an overall sanity check for missing values.

## 3.2 Feature Extraction

To begin with, we need to extract features from our CT scans and stratify the cohort of patients with regard to these extracted features and available clinical information. The feature extraction process enables a deeper understanding of each patient because after this process we can identify the volume, surface area, and roundness of the shapes, all of which may contain valuable information for recognising patterns of patients. This should facilitate us to derive insights about their evolution and overall survival. The computed radiomics features can be found on the readily available Pyradomics documentation [Rad] and are as follows:

- First Order Features [Fir]
- Shape Features (3D) [Shab]
- Gray Level Co-occurrence Matrix (GLCM) Features [Graa]
- Gray Level Size Zone Matrix (GLSZM) Features [Grad]
- Gray Level Run Length Matrix (GLRLM) Features [Grac]
- Neighbouring Gray Tone Difference Matrix (NGTDM) Features [Nei]
- Gray Level Dependence Matrix (GLDM) Features [Grab]

The only features not included are the 2D shape features [Shaa] and the removed features within pyradiomics [Rem].

In addition to the CT scans and segmentations, a spreadsheet with clinical data regarding the patients containing demographic (such as age and sex), behavioural (smoking and drinking habits) and many more scores related to the tumour's staging have been provided. Regarding the clinical data, there are a total of 56 feature columns and one TCIA ID column present. From those 56 features, we chose 14, which had relevant information to us. These were checked for missing values. We grouped these features as shown in the following List 3.2.

- Overall survival (OS): number of weeks
- Clinical predisposition:
  - Hepatitis: none, HCV/HBV, or both
  - Family history of cancer (fhx_can) and family history of liver cancer (fhx_livc)
- Demographic
  - Age: years
  - Sex: 1 = Male 2 = Female
- Behaviour
  - Smoking: 0 = does not smoke, 1 = does smoke
  - Alcohol: 0 = does not drink, 1 = does drink
- Classfification
  - Child Pugh classification score of cirrhosis mortality severity (CPS): A,B,C
  - Cancer of the liver italian program score (CLIP_Score)
  - Okuda
  - Tumour, lymph node and metastasis (TNM) staging
  - Barcelona clinic liver cancer staging (BLCL): 0, A, B, C, D

## 3.3 Scaling and Testing for Normality

After extracting features of both the patient's liver and tumour, each variable is distributed in its own value range. It is generally recommended as stated by Singh et al. [SS20] and Patro et al. [PS15] to normalize all features (variables) into the same value range before further clustering. Thus there exist two main options:

- Standardization [PS15]
- Min-max normalization [PS15]

Both have their advantages and disadvantages. Standardization (also known as z-score normalization) assumes that the variables are normally distributed. If that is the case, then any sample within the distribution can be transformed according to its mean and standard deviation with the following formula:

$$
\begin{aligned}
&X' = \frac{X - \mu}{\sigma} \\
&X' \text{is the Z-score normalized values} \\
&\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ or mean value} \\
&\sigma = \sqrt{\frac{\sum_{i=1}^{n} X_i - \mu}{n-1}} \text{ or standard deviation of X}
\end{aligned}
$$

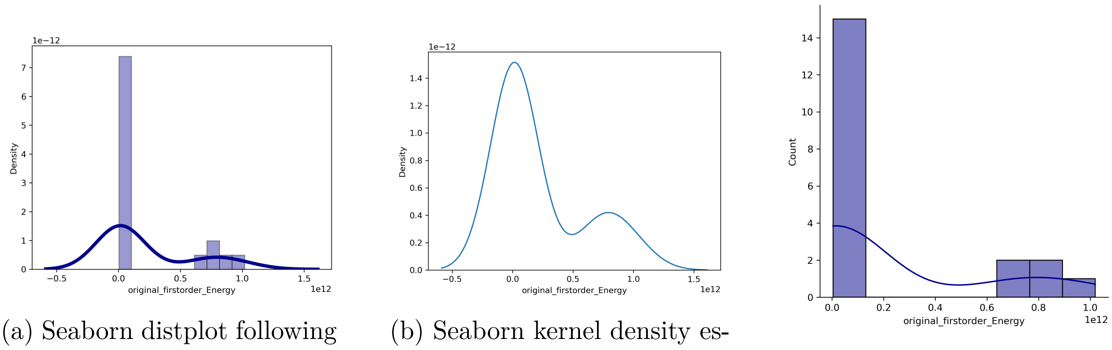
This all hinges on the assumption that the feature follows a normal distribution. If that turns out to be wrong then we can go for normalization. Here we define the maximum and minimum value of the new range to which the variable should be mapped to. The min-max normalization formula used in our implementation will be the following:

$$
\begin{aligned}
X' &= \frac{X - X_{min}}{X_{max} - X_{min}} \cdot (B - A) + A \\
X' &\text{ is the min-max normalized values} \\
X_{min} &\text{ is the minimum value of X} \\
X_{max} &\text{ is the maximum value of X} \\
A &\text{ is the minimum value of the new range} \\
B &\text{ is the maximum value of the new range}
\end{aligned}
$$

Now to test our normality assumption we may use histograms and plots of the probability density function. There are guides and tutorials for this such as on Towards Data Science [Koe] and documentation pages of packages like Seaborn [Seab], [Seaa]. Kohersen's guide [Koe] was taken as a reference for plotting the kernel density estimation as seen in Figure 3.2a. From there we followed the documentation of Seaborn and its `kde-` and `displot`, resulting in Figures 3.2b and 3.2c respectively. Each plot shows the kernel distribution estimation of the extracted first-order energy radiomics feature, which does not follow a normal distribution but rather a bimodal one and as such we would need to normalize our data. We also tested normality with the help of the implementation provided by scipy [Scid], which is based on the works of D'Agostino et al. [DP73] [DIA71] and Shapiro et al. [SW65]. We tested against the standard p-value of 0.05, as a result, 40 features do not follow a normal distribution.

## 3.4 Dimensionality Reduction

The next step is to perform dimensionality reduction. As there are 111 features for a liver and a tumour per patient, it becomes more and more computationally intensive to



(a) Seaborn distplot following Kohersen's guide.

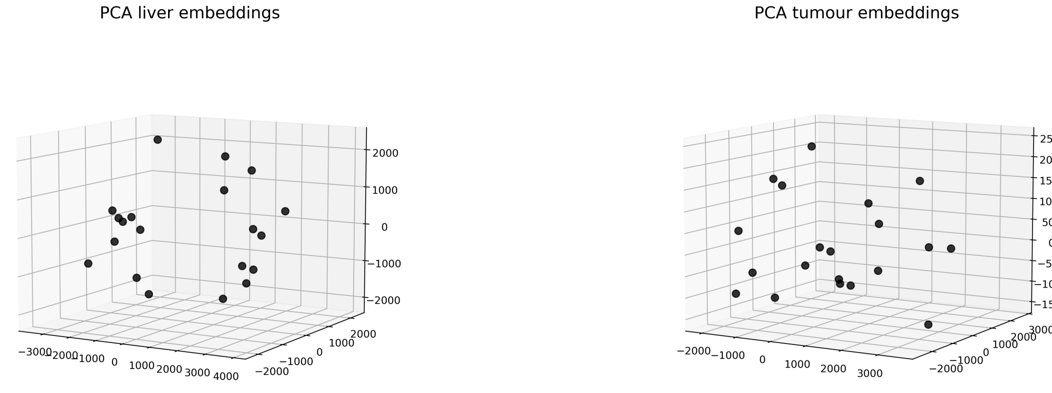
(b) Seaborn kernel density estimation plot.

(c) Seaborn displot.

perform clustering on such "wide" datasets. This leads to the curse of dimensionality found in the data mining context to be precise, as shown by Verleysen et al. [VF05]. Thus, to train a complex model we need a rather high amount of samples to attain a desirable validity. While one could argue that with today's technology, it would not matter that much for real-time execution for a small number of patients, this changes drastically with an increase in the number of patients and features. Another important note regarding dimensionality reduction is that high-dimensional data is often difficult to visually encode and interpret.

### 3.4.1 PCA: Principal Component Analysis

Principal component analysis (PCA) [WEG87] is a method in which linear combinations of variables are made such that the variance of data is maximised along the axes of said linear combinations, also known as principal components. This is done to maximise the information gained from the raw data. By doing so, we can reduce the dimensionality with some loss in reconstruction, but gain a more efficient representation of the data. In that sense, PCA is applied in order to keep the most information from the data to enable faster computation of organ representation. We used the implementation provided by the Scikit-learn package [PVG$^+$11]. In Figures 3.3a and 3.3b we see the liver and tumour data being embedded into the new space of a 3D PCA with the use of the first three principal components that each result after fitting the models to the data. Three principal components are chosen at first to better visualize the embeddings, but later on, we will investigate how many components are needed to retain a certain amount of variance.



(a) Embeddings of liver data

(b) Embeddings of tumour data

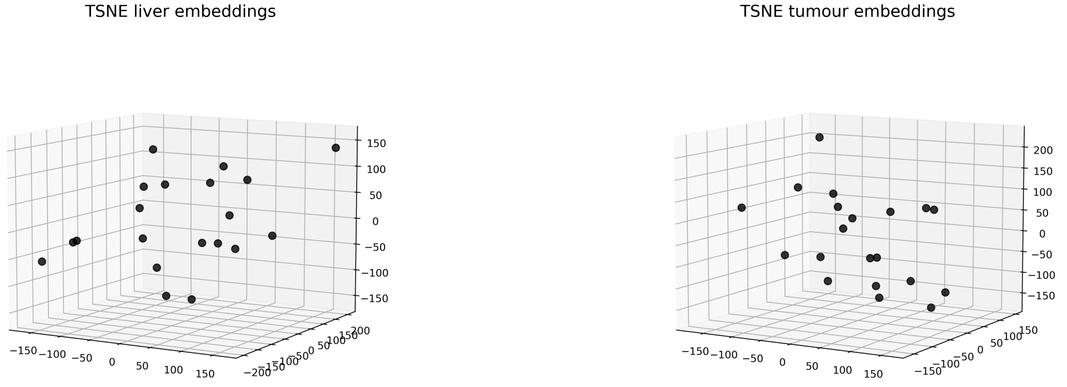
Figure 3.3: Embeddings using the first three principal components of PCA.

### 3.4.2 t-SNE: t-Distributed Stochastic Neighbor Embedding

Hinton and Roweis [HR02] presented the algorithm for Stochastic Neighbour Embedding (SNE) as an alternative to PCA. Van der Maaten and Hinton [VdMH08] then improved over the original SNE, which reduces the tendency to map points at the centre of the projection. In t-SNE, the first step involves computing the symmetric similarities of data points from their Euclidean distances. These similarities represent the probabilities $p_{ij}$ that a data point $x_i$ would select data point $x_j$ as its neighbour, assuming they were chosen from a Gaussian distribution centred at $x_i$. When $i = j$, the conditional probability is set to 0 due to only being interested in pairwise similarities. The same is done for the low-dimensional counterparts $y_i$ and $y_j$ resulting in conditional probabilities $q_{ij}$, however, using a Student-t distribution. The algorithm tries to find abstract mappings such that the mismatch between $p_{ij}$ and $q_{ij}$ is minimized. The algorithm minimizes the cost function, which is the sum of Kullback-Leibler divergences [VdMH08] over all data points, using a gradient descent procedure:

$$cost = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The user-defined perplexity value is used to find the variance $\sigma_i$ of each Gaussian distribution centred around data point $x_i$. The key difference is that t-SNE attempts to preserve local information about neighbourhoods rather than the global approach PCA has of finding linear combinations that maximize variance. A drawback to this method is that we can at most obtain three new axes due to the algorithm limitation. Once again the implementation from the Scikit-learn package [PVG$^+$11] was used. In Figures 3.4a and 3.4b we see the liver and tumour embeddings using three of the resulting axes from t-SNE.



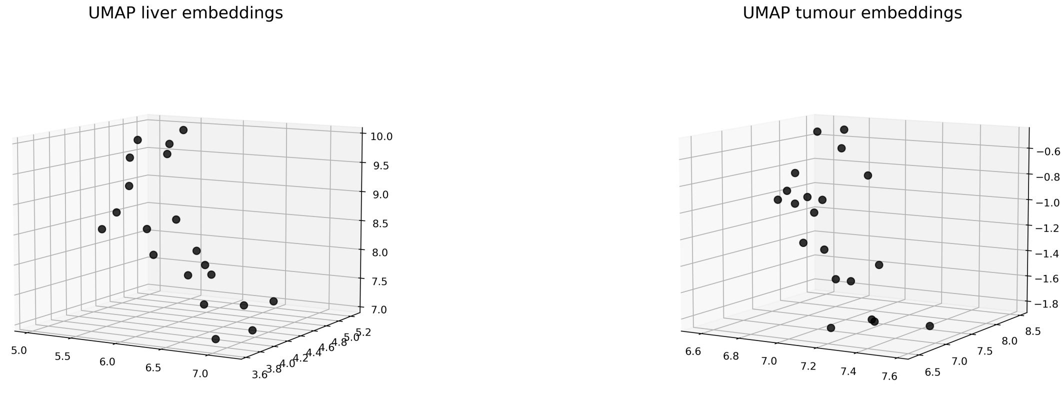
(a) Embeddings of liver data.

(b) Embeddings of tumour data.

Figure 3.4: Embeddings of volumetric data using the first three axes of t-SNE.

### 3.4.3 UMAP: Uniform Manifold Approximation and Projection

McInnes et al. [MHM18] present the uniform manifold approximation and projection technique for dimension reduction. The procedure can be seen as a mix between PCA and t-SNE when speaking about the preservation of linearities. UMAP in its essence is based upon topological data analysis and simplexes [MHM18], which are used to represent the data points and capture the fundamental topology of the data set. This can be seen as building an abstract neighbourhood graph of the data set and projecting it into a lower-dimensional space. Seen from a high abstraction level, UMAP learns the representation of a manifold, which approximates the data set by using local fuzzy simplicial set representations of the data. Following that, UMAP optimizes the manifold representation such that the cross entropy between the lower and higher dimensional space is minimized. In this minimization process, we can imagine that there are attractive and repulsive forces between the points of the edges spanned in the graph [How]. Depending on the set parameters it can prioritize the preservation of global linearities or local ones. There are numerous advantages to UMAP such as it being faster than t-SNE [MHM18] [Per], scaling better with embedding dimension and being able to transform new data [Tra]. In Figures 3.5a and 3.5b we see the embeddings of the liver and tumour data as projected in the 3D space using the first three axes.



(a) Embeddings of liver data.

(b) Embeddings of tumour data.

Figure 3.5: Embeddings of volumetric data using the first three axes of UMAP.

### 3.4.4 Evaluation of Explained Variance

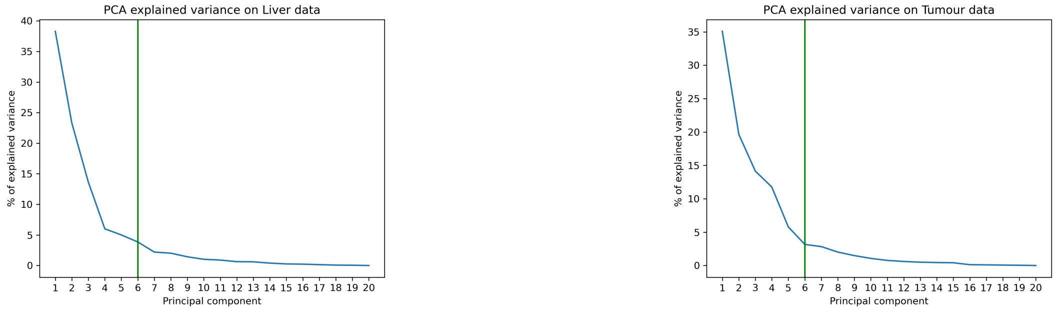
When comparing these three different approaches to dimensionality reduction we need a metric. PCA offers the variance explained metric for its principal components. Unfortunately, this is not universally applicable to t-SNE and UMAP directly from their respective implementation, as they operate on a different basis and their aim is not to maximize variance.

The method of visualizing explained variance is called a Scree plot, first introduced by Cattell [Cat66]. Figure 3.6 displays the following ratio of the eigenvalues $\lambda_i$ for each principal component:

$$\%\ of\ explained\ variance = \frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i} \cdot 100$$
$$\lambda_i = i\text{-}th\ eigenvalue$$

In Figure 3.6 we see how much variance is explained by a particular principal component from the PCA application to our data. The first component manages to explain about 38% of the total variance, thus containing the most information of all the principal components.

We prioritize the preservation of at least 90% of the variance and, as such, we can take a look at the cumulative explained variance in Figure 3.7. Hereby we are adding up the individual percentages in descending order because the goal is to also minimize the amount of principal components used. We decided to use the first 6 components and thereby go from 111 radiomics features down to only 6 reduced features, while also preserving at least 90% of the variance. Of course, this would only apply if we were to use the PCA reduction, but we decided to use UMAP as the final tool for this, as it is the

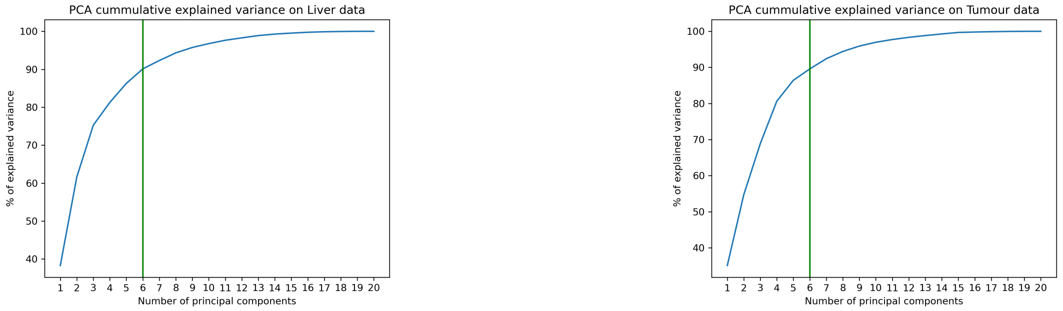


(a) Based on liver data.

(b) Based on tumour data.

Figure 3.6: Percentage of explained variance by each principal component by data provided. The green line shows the number of principal components needed for at least 90% variance preservation.

newest of the three methods. Diaz-Papkovic et al. [DPATG21] concluded that UMAP has shown wide adoption among the scientific community and that it is a viable option for processes like clustering, giving us another reason to choose UMAP.



(a) Based on liver data.

(b) Based on tumour data.

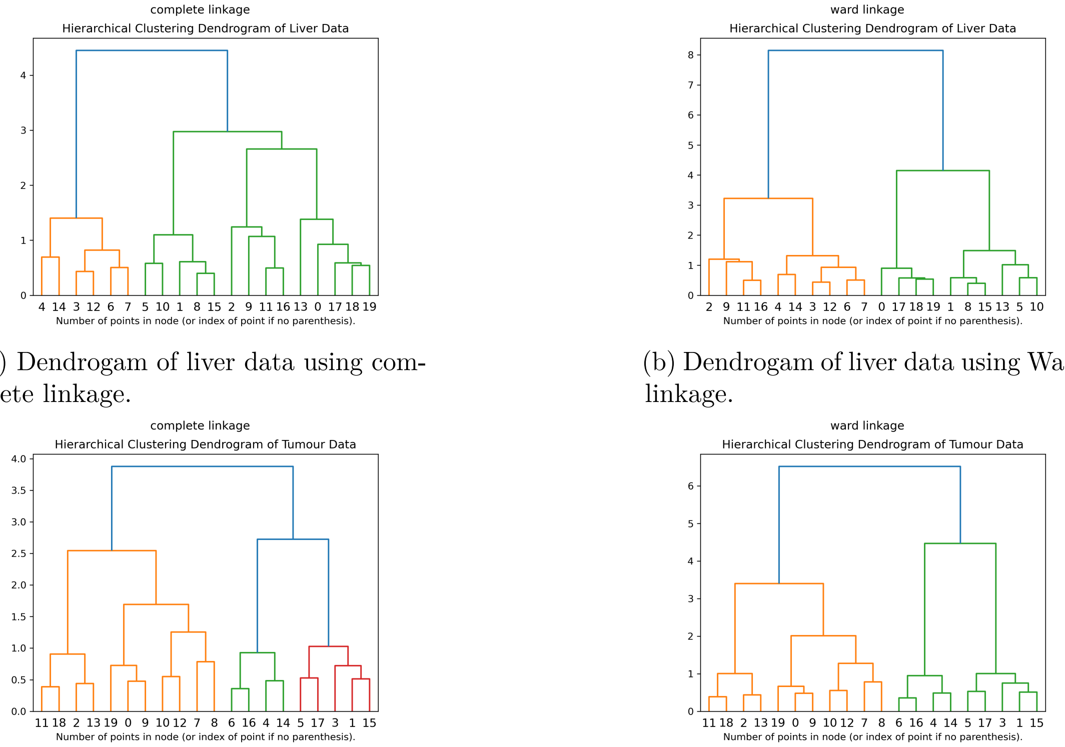
Figure 3.7: Percentage of cumulative explained variance from the PCA approach. The green line shows the number of principal components needed for at least 90% variance preservation.

## 3.5 Clustering and Merging of Data Sets

We perform the clustering of patients based on the resulting dimensionality reduction embedding. Sklearn [PVG+11] once again offers a wide range of options regarding clustering [Scib] with an overview of each method, including use cases and scalability. We choose to use agglomerative clustering [Scic], as we are interested in how clusters may form and which merge together. Agglomerative clustering starts out by placing each data point in its own cluster and step by step merges the nearest ones together until one single node remains [Mül11]. When using agglomerative clustering [Scia] we need to set a couple of parameters:

- `n_clusters` defines how many clusters need to be found. This has been set to `None`.
- `metric` defines the metric used to compute the linkage distances. `'euclidean'` was used in order to allow for a direct comparison of the linkages, as the `ward` linkage only works when paired with the `euclidean` distance metric.
- `linkage` defines the linkage criterion used during the merging step of clusters. We tested with `ward` and `complete` linkages to see the difference in cluster formation.
- `distance_threshold` defines the linkage distance threshold after which clusters will not be merged and was set to 0 because we want to compute the whole tree.

Klemm et al. [KLR+13] investigated the different linkage methods and concluded that the `ward` method was biased towards building clusters, which have a similar size, and that `complete` produced compact clusters. Böröndy et al. [BFR22] also looked into how different setting of the clustering affects cluster formation, and found that there is only a slight difference in the formed clusters. We therefore can say that the selected settings need to be investigated on a case-by-case basis with respect to the data set used and a generalization is not possible. We chose the `complete` linkage criterion to be used. In figure 3.8 we can see a comparison between the `complete` and `ward` linkage used in hierarchical clustering on the liver and tumour data resulting from the UMAP reduction.



(a) Dendrogam of liver data using complete linkage.

(b) Dendrogam of liver data using Ward linkage.

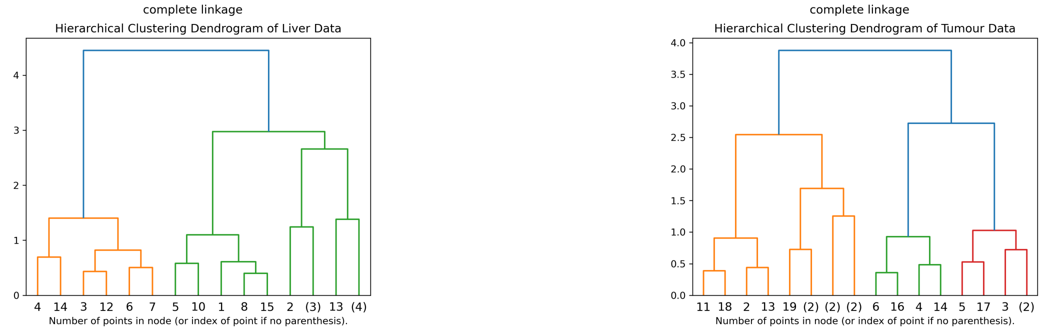
(c) Dendrogam of tumour data using complete linkage.

(d) Dendrogam of tumour data using Ward linkage.

Figure 3.8: Comparison of `complete` and `ward` linkages in hierarchical clustering.

In the previous examples, we were able to see the indices of patients displayed on the x-axis. This works fine when there is a low amount of data, however, when working with larger datasets it would be recommended to only keep the top few levels as those tend to be more relevant as seen in Figure 3.9. Afterwards, we merged the data based on the TCIA patient IDs. We investigated how the clusters would form when the un-reduced dataset is used. In Figure 3.10 we see the dendrograms for the respective datasets. In Figures 3.10b and 3.10a we see that the clustering remained the same for the liver data when looking at the level where only 2 clusters are taken into consideration. Regarding the liver data in Figures 3.10d and 3.10c some changes in cluster structure are made visible when looking at the indices and clusters correspondence. For example in Figure 3.10d the patient with index 17 does not get merged into the same cluster as the one with ID 9 until the very end, as opposed to what is shown in Figure  3.10c.

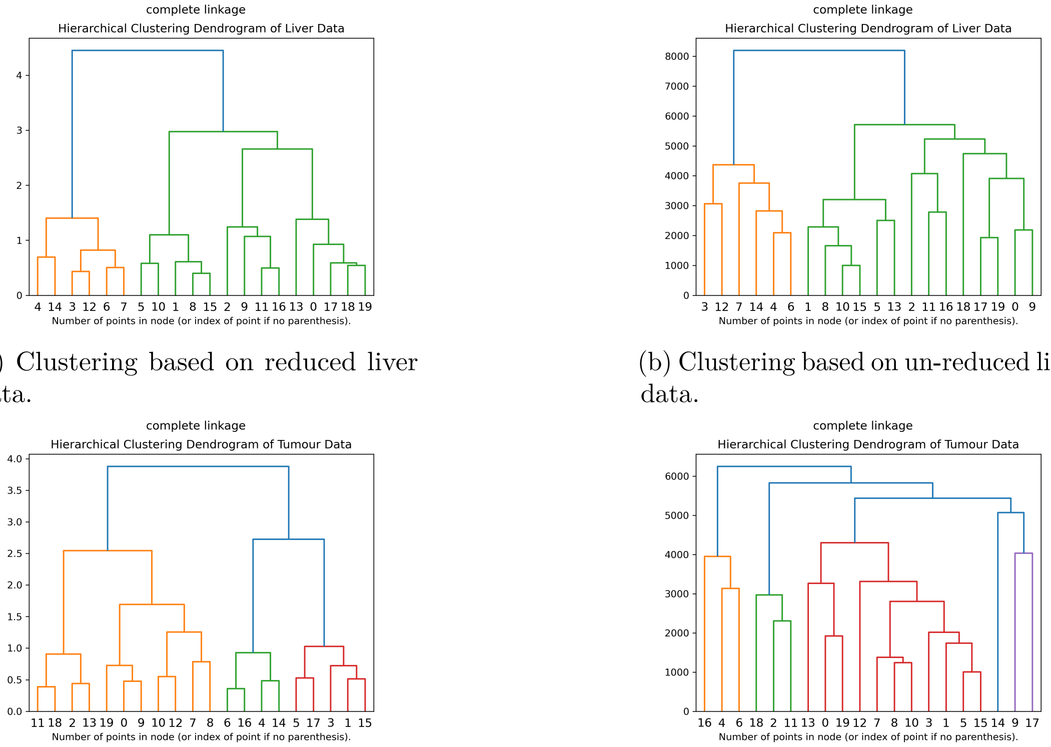
To summarize, we investigated three dimensionality reduction algorithms, PCA, t-SNE and UMAP, compared these and decided to employ UMAP as a dimensionality reduction tool to reduce the size of our data. Afterwards, we investigated agglomerative clustering and compared the resulting clusters of the patient's livers and tumours based on the reduced data from UMAP, using the `ward` and `complete` linkage.



(a) Dendrogram of liver data showing the top 3 levels.

(b) Dendrogram of tumour data showing the top 3 levels.

Figure 3.9: Truncated dendrograms where only the top 3 levels are displayed. The number of patients within an undisplayed subcluster is displayed in the parenthesis. Numbers without parentheses represent the indices of patients.



(a) Clustering based on reduced liver data.

(b) Clustering based on un-reduced liver data.

(c) Clustering based on reduced tumour data.

(d) Clustering based on un-reduced tumour data.

Figure 3.10: Comparison of clustering results based on the original and reduced features.

## 3.6 Visualization

Visualizations not only offer a more comprehensible view of the data to the researchers but to anyone who has access to them. The goal is to enable users to investigate the data set and gain insights into the groups forming through cohort stratification. Parallel coordinate plots allow us to visualize higher dimensional points in a 2D plot, making relationships between points visible. Heatmaps enable us to investigate potential correlations between specific features of the data set.We employ Sankey diagrams to investigate possible sets of patients in the data set and follow their paths. Sunburst charts and treemaps also help in investigating sets of patients and building hypotheses about the cohort. These visualizations aim to uncover connections between data points, the features present in the clinical data set, and the derived liver and tumour descriptors By including these visualizations, we are getting closer to a visual analytics solution, which if applied correctly, will help doctors in investigating their patient cohorts. In the next few subsections, we present different methods of visualizing our data.

### 3.6.1 Parallel Coordinates Plot

Inselberg et al. [ID90] present parallel coordinates plots as a tool for representing n-dimensional data on a 2-dimensional plane. According to the authors, these plots can have superficial similarities to Nomography [Bee24], however, the main difference is that parallel coordinate plots enable the representation of multidimensional relations. Inselberg et al. also note that at that time, the fields of robotics, statistics, and computational geometry, as well as other fields, had an increased interest in this method of visualization. A possible use case for this plot was found in the domain of air traffic control, where the detection of conflicts between aircraft was handled. We chose parallel coordinate plots as the first approach to visualize the stratified patient cohort. As can be seen in Figure 3.11 not much insight is gained due to the overlaps of the lines created by having discrete features. Due to this overlapping, we can not deduce how many lines pass through each level on the feature's axis. At most, we could infer from the colour of the lines which paths may have higher overall survival duration. Therefore, we looked into other possible visualizations that can provide us with a sense of how many patients are following a specific path.

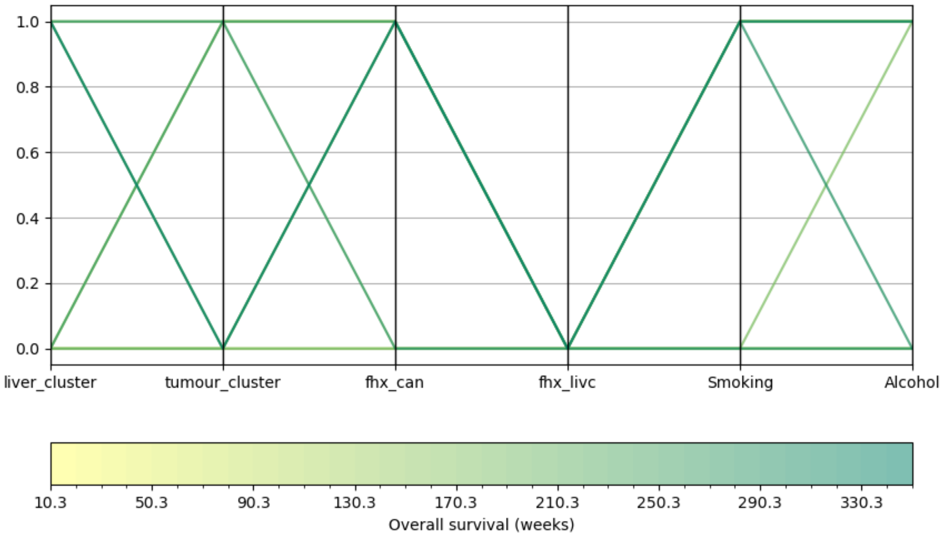


Figure 3.11: Parallel coordinates plot of selected features.

### 3.6.2 Sankey Plot

Sankey et al. [San96,KS98] first introduced Sankey diagrams to show the thermal efficiency of steam engines. Riehmann et al. [RHF05] shows a general example of how Sankey diagrams can be used. These plots can also be employed in the visualization of material flow management as shown by Schmidt [Sch08a, Sch08b]. Schmidt [Sch08b] investigates some of the main use cases of Sankey plots and finds that these can be used for visualizing flows and paths of resources, like materials and energy, within a system. Thus, Sankey plots can visualize the paths of patients within a hospital or other healthcare institutions, as shown by Lamer et al. [LLP+20]. Lamer et al. [LLP+20] also mention that these diagrams can be employed to "*visualize the trajectory of a population that experienced an event*". In our case, the population is our cohort of patients and the corresponding event is their examination. They also mention that investigating these patient flows may contribute towards generating hypotheses about patient care. According to their results, the Sankey diagrams help in detecting atypical patient flows, investigating and validating hypotheses, and are easy to understand for end-users. We see the results mentioned above as further motivation to use Sankey diagrams in order to visualize and provide an overview of patient paths. In Figure 3.12 presents the set of patients along all features.

We must take into consideration visual overlap due to multiple features having rather long descriptions. The representation within the notebook allows us to drag the paths and reposition them interactively to resolve clutter, as seen in Figure 3.13. The plot shown in Figure 3.13 can still be quite hard to navigate and understand. We suggest splitting it

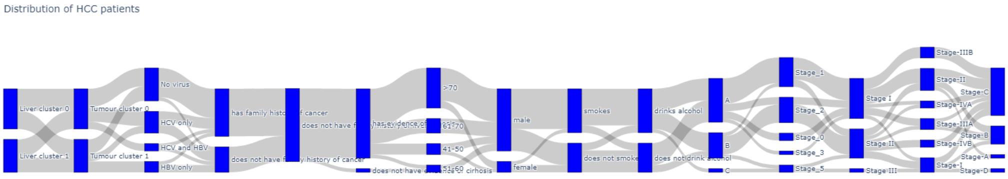


Figure 3.12: Sankey diagram containing all combined features.

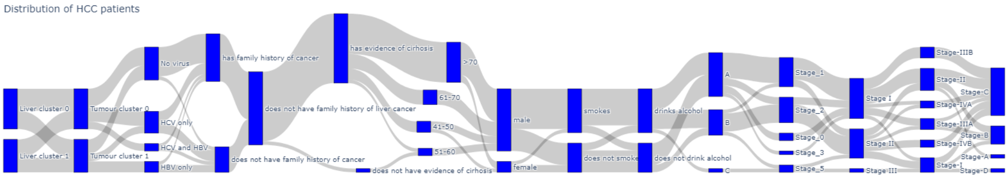


Figure 3.13: Sankey diagram where some paths have been dragged such that the descriptions of the boxes are visible.

up into several plots, each containing a part of the original. An example of such a plot can be seen in Figure 3.14. From the simple visual plot, we are already capable of getting

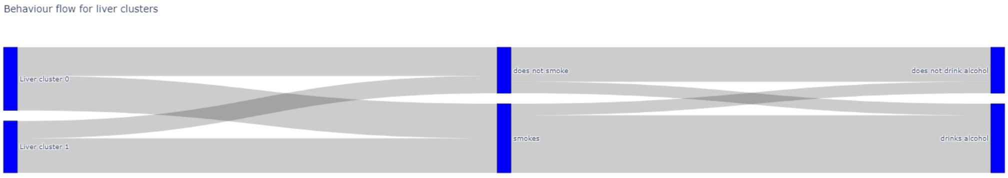
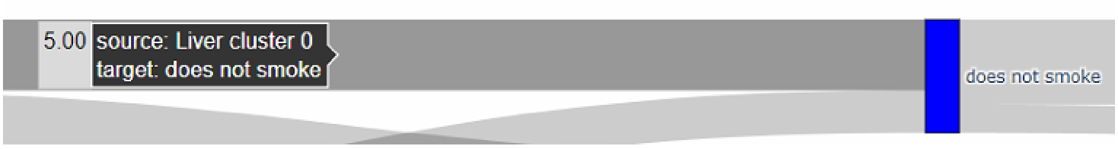


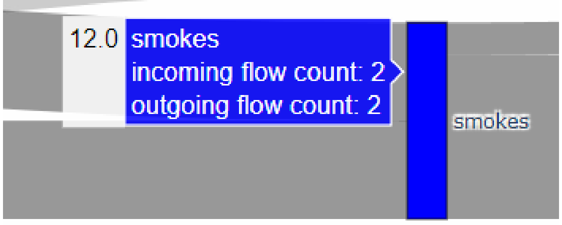
Figure 3.14: Sankey plot depicting the flow of patients based on the liver cluster, drinking and smoking behaviour.

some information about the sets of patients and what paths may look like. However, the Sankey plot offers further information and the concrete number of individuals that are within one set when hovering over the grey flow path as seen in Figure 3.15. When hovering over the blue nodes we will get information about the number of incoming paths and outgoing paths as seen in Figure 3.16. This feature is useful when there are multiple paths that overlap and thus would lead to false assumptions about the data flow within the chart.



5.00 source: Liver cluster 0
target: does not smoke

Figure 3.15: Additional information provided by the Sankey chart when hovering over a flow path. "source" depicts the starting node of the path, whereas "target" the terminal node. The number "5.00" shows us how many patients (or entities) are within this path.

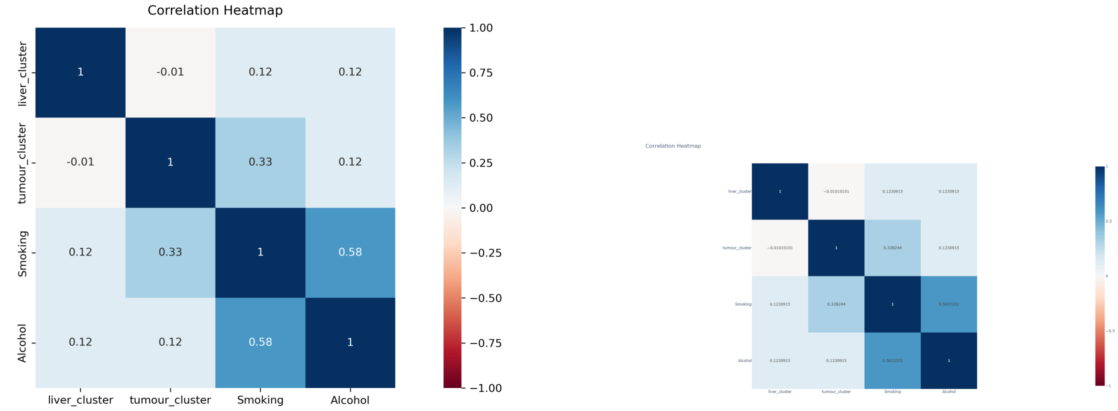


12.0 smokes
incoming flow count: 2
outgoing flow count: 2

Figure 3.16: Additional information provided by the Sankey chart when hovering over a node of the chart. the incoming and outgoing flow counts are also represented, in this case, both are 2. The number "12.0" describes how many patients are within that node.

### 3.6.3 Heatmaps

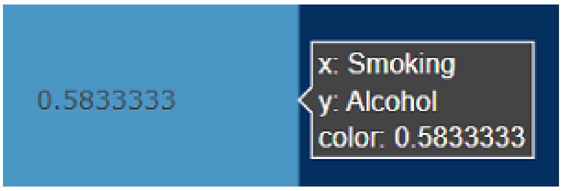
We employ heatmaps in order to check how various variables may correlate to each other. Heatmaps originate in the information visualization domain from the work of Toussaint Luoa [Lou73]. Heatmap visualizations, as seen in papers of Wilkinson et al. [WF09] or Eisen et al. [ESBB98] can be used for showing gene expression across different conditions. Each function has its advantages and disadvantages. In Figure 3.17a we see the basic heatmap containing the values of the correlation matrix and the colour bar. In Figure 3.17b the colour bar is further away and it is not as friendly for printing, but shows more digits within each cell. By using Plotly we also gain on-hover information like before as seen in Figure 3.18. Figures 3.19a and 3.19b represent the covariance heatmaps via the use of the beforementioned packages.



(a) Seaborn package.

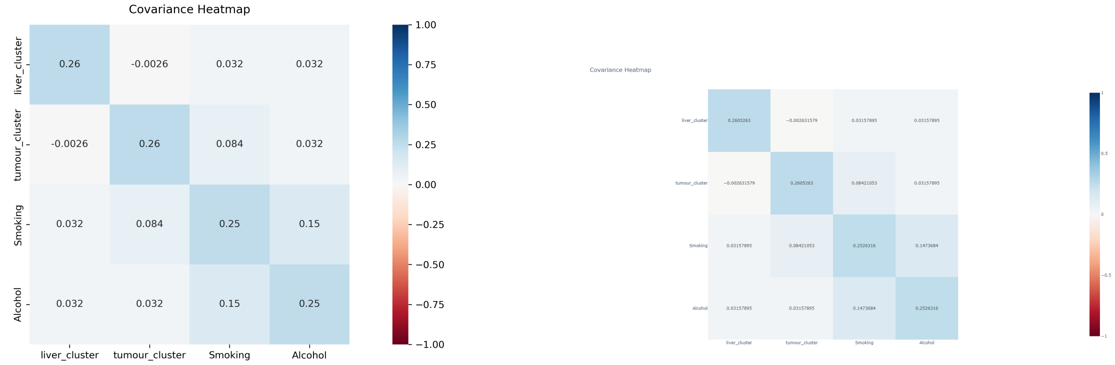
(b) Plotly package.

Figure 3.17: Correlation heatmaps of selected features using the Seaborn and Plotly package.


0.5833333

x: Smoking
y: Alcohol
color: 0.5833333

Figure 3.18: Information provided in the Plotly heatmap when hovering over a cell. x and y are the coordinates of the matrix, in this case, 0.5833333 is the correlation value between "Smoking" and "Alcohol".



(a) Covariance heatmap of selected features using the Seaborn package.

(b) Covariance heatmap of selected features using the Plotly package.

### 3.6.4 Sunburst Chart

Sunburst charts are a method of visualizing hierarchical data in a circular manner. Here, the plot arranges the different hierarchies radially, beginning at the centre with the root and expanding outwards with each ring being another step in the hierarchy. Stasko et al. [SCGM00] conducts a comparison between the sunburst and treemap chart types, where the capability of finding files with the help of these charts was evaluated. They conclude that the sunburst chart provided more help, as finding files was done faster and more correctly than with the treemap.

In Figure 3.20 we can see the cohort of patients represented in a sunburst chart. Here the colour of the circle segments represents the overall survival duration measured in weeks. By clicking on any given segment, the chart will update the view and a zoomed-in version is displayed. For instance, by clicking on tumour cluster 0 in the liver cluster 0 group we obtain the chart seen in Figure 3.21. This way, we can explore our data in more detail and also see the bigger picture by having the zoomed-out version. It is possible to go back to any layer by clicking on the inner-most circle.

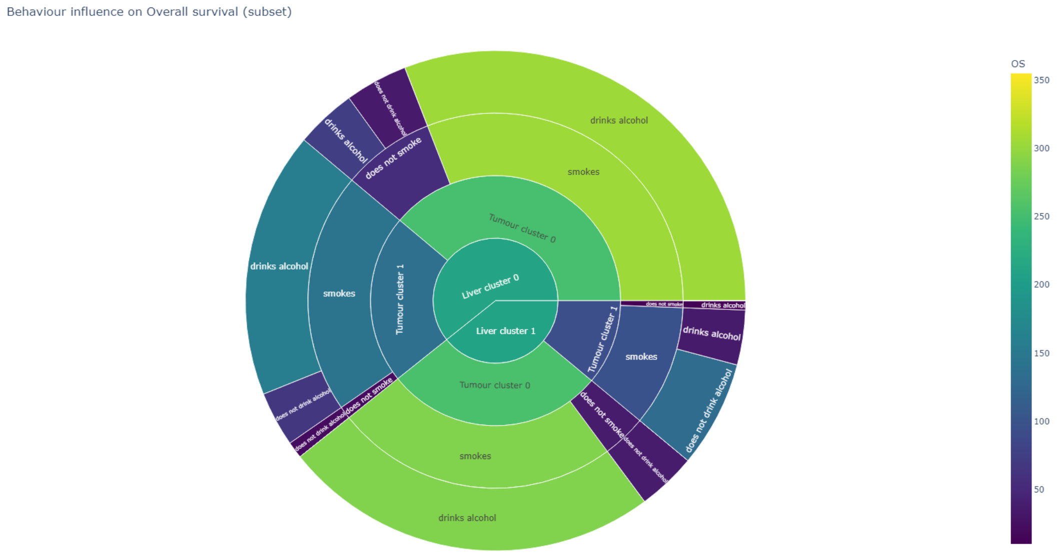


Figure 3.20: Sunburst chart including the liver cluster and tumour cluster descriptors as well as the smoking and drinking habits of the patients. The colours correspond to the average overall survival (in weeks) of the groups formed.

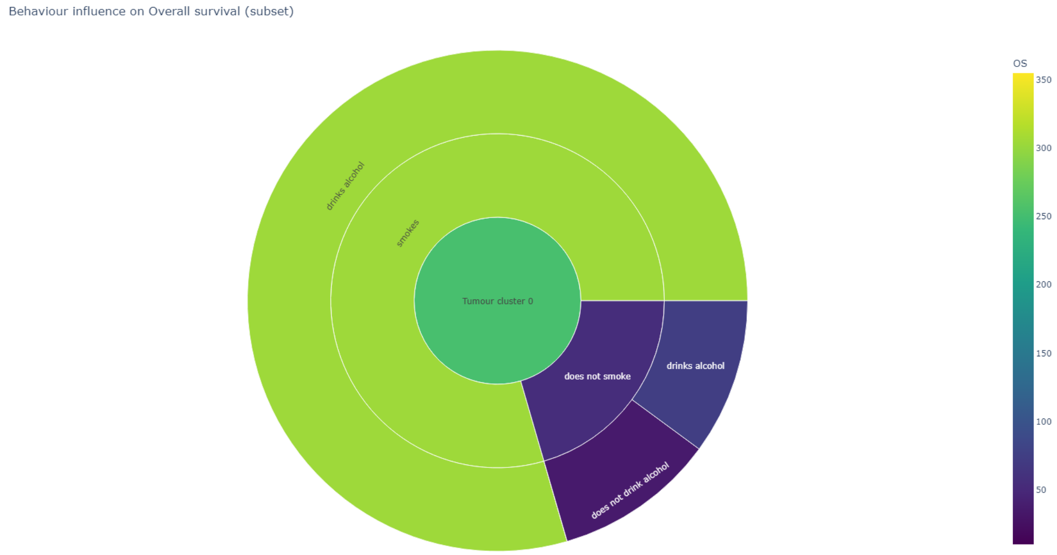


Figure 3.21: Sunburst chart including the tumour cluster descriptor as well as the smoking and drinking habits of the patients in liver cluster 0.

### 3.6.5 Treemaps

Ben Shneiderman [Shn92] first introduced treemaps as a modality to visualize hierarchical data similar to the sunburst chart. In Figure 3.22 we see a similar view to the sunburst chart in Figure 3.20. However, the difference is that the age groups and sex are added to see how this influences the cohort of patients. Once again it is possible to obtain zoomed-in versions by clicking on a desired group. As an example when clicking on the group with the descriptors of liver cluster 0 and tumour cluster 0, we obtain the view presented in Figure 5.3.

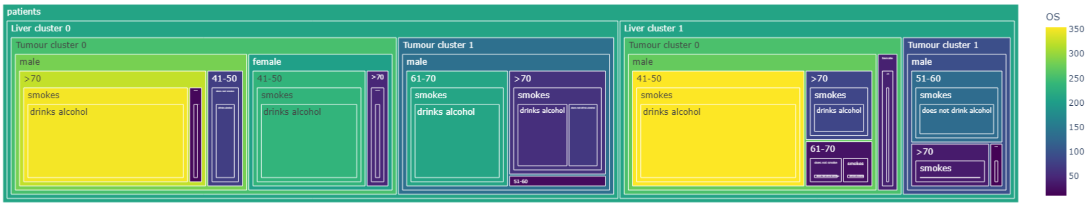


Figure 3.22: Treemap chart including the liver cluster and tumour cluster descriptors as well as the smoking and drinking habits, age groups and sex of the patients. The colours correspond to the average overall survival (in weeks) of the groups formed.

# CHAPTER 4

# Implementation

We envisioned the implementation to follow our presented methodology and based it on the research of related work, which makes use of radiomics. We found that most of the features are already available via the Pyradiomics [VGFP$^+$17] package, and therefore used it for feature extraction. Pyradiomics, according to its authors, aims to "*establish a reference standard for Radiomic Analysis, and provide a tested and maintained open-source platform for easy and reproducible Radiomic Feature extraction*". Reproducibility is very important because we do not want feature values to change at random, impacting any and all processing steps in the following pipeline.

For visualizing the distributions of our data, we have used various plots implemented in the Seaborn [WBO$^+$17] package. We needed this to determine how the data points were scattered in space and also get an intuition about their distribution. We also tested the normality of each feature with the help of the implementation provided by scipy [Scid]. Dimensionality reduction used algorithms implemented in the Scikit-learn [PVG$^+$11] and UMAP [MHM18] packages. As seen in the chapter for clustering 3.5, we also used packages like Scipy [VGO$^+$20], and Scikit-learn [PVG$^+$11] for clustering our data and visualizing those results. We used Plotly [Inc15] and Seaborn [WBO$^+$17] to investigate the structure of our cohort and draw conclusions regarding groups of patients as seen in Chapter 3.6. We made the Sankey plots with the help of the Plotly library and their documentation [San]. For heatmaps, we used the respective functions found in the Seaborn [Heaa] and Plotly [Heab] packages. We use the Plotly implementation of the sunburst chart [Sun] as it offers interactions like going deeper into the hierarchy and making it possible to go down different paths while investigating the cohort. For Treemaps, we chose to once again use the Plotly version [Tre] as it is quite easy to use and offers interactive capabilities when viewed directly in the notebook or as an HTML file. The implementation itself is available on GitHub [Git].

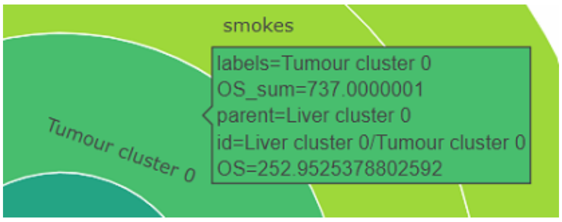
# CHAPTER 5

# Results

As seen in Section 3.6 where we have taken a look at different visualization options, some interesting patterns arise regarding the groupings of patients and how their overall survival rating changes. For example, when looking at the sunburst plots in Figure 3.21, we could say that the patients of liver cluster 0 and tumour cluster 0 who drink alcohol and smoke have higher overall survival than those who do not.

Sankey diagrams, as seen in Figure 3.13, can offer an idea of how the paths of patients may change with respect to the observed feature. This can help in identifying features that could contribute to improving the separation of groups (e.g., sex, age group) or those which do not (e.g., family history of liver cancer fhx_livc).

From sunburst charts, we conclude that the patients that are in liver cluster 0 and tumour cluster 0 who neither smoke nor drink alcohol have a lower overall survival duration than those who would smoke and drink but have the same categorisation in regards to their liver and tumour cluster, which is quite intriguing. In this case, we are interested in what the concrete average value of the overall survival duration is, then we can hover over the desired segment and also find the path which leads to it as seen in Figure 5.1 is of about 252.95 weeks for the group who are in the liver cluster 0 and tumour cluster 0. From

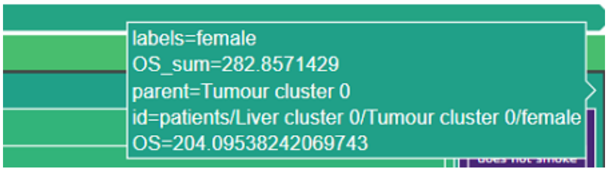


labels=Tumour cluster 0
OS_sum=737.0000001
parent=Liver cluster 0
id=Liver cluster 0/Tumour cluster 0
OS=252.9525378802592

Figure 5.1: Additional information provided by the sunburst chart when hovering over a segment.

sunburst charts, we conclude that the patients that are in liver cluster 0 and tumour cluster 0 who neither smoke nor drink alcohol have a lower overall survival duration than those who would smoke and drink but have the same categorisation in regards to their liver and tumour cluster, which is quite intriguing. In this case, we are interested in what the concrete average value of the overall survival duration is, then we can hover over the desired segment and also find the path which leads to it as seen in Figure 5.1 is of about 252.95 weeks for the group who are in the liver cluster 0 and tumour cluster 0.

Treemaps offer a more rectangular view of the tabular data but offer the same functionality as the sunburst plots. In Figure 3.22 additional information like the age group and sex were used to look for other possible variables that may explain the overall survival scores of the patients. It turns out that the age group has a large contribution when brought into the mix of displayed variables. For instance, in Figure 5.3, looking only at liver cluster 0 tumour cluster 0 patients, we can notice a difference in the survival score between male and female patients. This difference is now further expanded when looking at the male patients and their age groups, where those above the age of 70 have a much higher survival score than those aged 41-50. The opposite behaviour can be seen for female patients, those aged 41-50 have higher scores than those above 70 years of age. From this, we may conclude that the visualizations aided us in discovering unforeseen connections between variables. Similar to the sunburst chart, the treemap chart also offers on-hover information seen in Figure 5.2 about the groups the same way as in Figure 5.1. From the information provided in Figure 5.2 we may say that the overall survival duration of female patients who are in the liver cluster 0 and tumour cluster 0 is on the average of 204.095 weeks. However, when looking at Figure 5.3, we can see that we also have a strong split between the 41-50 and the >70 age groups with vastly different values for overall survival, one having around 200 and the other close to 50 weeks.



labels=female
OS_sum=282.8571429
parent=Tumour cluster 0
id=patients/Liver cluster 0/Tumour cluster 0/female
OS=204.09538242069743

Figure 5.2: Additional information provided by the treemap chart when hovering over a segment.

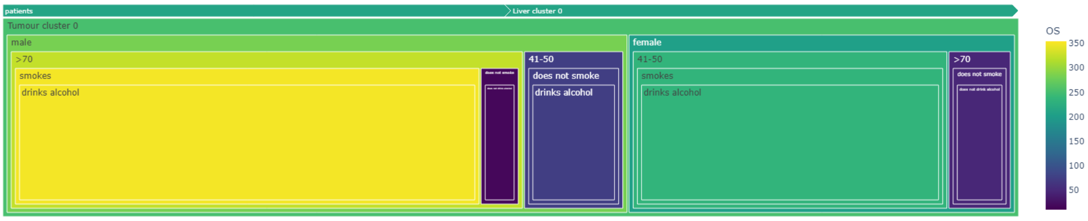


Figure 5.3: Treemap chart where the subgroup of "liver cluster 0" and "tumour cluster 0" have been selected.

CHAPTER 6

# Conclusions and Future Work

To summarize, this thesis investigated the stratification of Hepatocellular Carcinoma patients. We described the radiomics feature extraction procedure used to obtain relevant numerical descriptors of our patients, which were then min-max normalized for the dimensionality reduction step via UMAP. Clustering was then discussed based on the reduced features in order to explain the derived liver and tumour cluster descriptors, which are mapped to the patient's clinical data. The merged data set was visualized via multiple different plots and insights were gained regarding the cohort.

Data scientists and clinicians looking to investigate the cohort of patients may find the graphics and visual plots useful. We can use Plotly Dash [SH19] in order to investigate other possible visualization methods provided in Dash and potentially offer another way of interaction besides Jupyter Notebooks. As the interests of patients and clinicians may not necessarily align, we could also find it interesting to see if there may be a need for a dedicated version for patients to help them better understand their health circumstances.

Another possibility is to integrate treatment data into the visualizations such that decisions could be supported by the measured changes in the patient's anatomy and metabolism. This would further improve the groupings that result from the visualizations, as we could see paths towards a more likely recovery or higher survival chances.

In general, one can apply these visualizations regardless of the originating dataset, as long as the data can be brought in a tabular form of features. This allows for repurposing the implementation for various data exploration applications like the following:

1. Investigating other medical affiliations via the use of radiomics. Lung deterioration due to some viral infections, smoking habits, and other pollutants like asbestos.

2. Kindeys may also be investigated to see possible kidney stone formations and how these may occur. Here, ultrasound as well as CT scans can be of help for acquiring the volumetric data.

3. Going outside the medical sector, we may also be interested in the energetic sector regarding power distribution and usage, where Sankey diagrams are known to be used. The same idea can be applied to investigating finances and following the inflow and outflow of capital. In this case, the features can be the power source (i.e., hydroelectric, solar, wind, nuclear power), the consumer type (i.e., metallurgic industry, residential buildings, education facilities etc.), or even the nodes within the distribution chain (i.e, power transformers, high-voltage power lines).

# List of Figures

# Bibliography

[AB94] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[ADQY19] Mubashir Ahmad, Yuan Ding, Syed Furqan Qadri, and Jian Yang. Convolutional-neural-network-based feature extraction for liver segmentation from ct images. In *Eleventh International Conference on Digital Image Processing (ICDIP 2019)*, volume 11179, pages 829–835. SPIE, 2019.

[AOH$^+$14] Paolo Angelelli, Steffen Oeltze, Judit Haász, 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.

[ASR$^+$12] U Rajendra Acharya, S Vinitha Sree, Ricardo Ribeiro, Ganapathy Krishnamurthi, Rui Tato Marinho, João Sanches, and Jasjit S Suri. Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm. *Medical physics*, 39(7Part1):4255–4264, 2012.

[Bee24] RD Beetle. S. brodetsky, a first course in nomography. 1924.

[BES$^+$17] Shaimaa Bakr, Sebastian Echegaray, Rajesh Shah, Aya Kamaya, John Louie, Sandy Napel, Nishita Kothary, and Olivier Gevaert. Noninvasive radiomics signature based on quantitative analysis of computed tomography images as a surrogate for microvascular invasion in hepatocellular carcinoma: a pilot study. *Journal of Medical Imaging*, 4(4):041303–041303, 2017.

[BFR22] Ádám Böröndy, Katarína Furmanová, and Renata Georgia Raidou. Understanding the impact of statistical and machine learning choices on predictive models for radiotherapy. In *# PLACEHOLDER_PARENT_METADATA_VALUE#*, volume 2022, pages 65–69, 2022.

[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.

[Cat66] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[CCC$^+$98] E-Liang Chen, Pau-Choo Chung, Ching-Liang Chen, Hong-Ming Tsai, and Chein-I Chang. An automatic diagnostic system for ct liver image classification. *IEEE transactions on biomedical engineering*, 45(6):783–794, 1998.

[CVS$^+$13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

[DIA71] R DIAgostino. An omnibus test of normality for moderate and large sample sizes. *Biometrika*, 58(34):1–348, 1971.

[DP73] RALPH D'agostino and Egon S Pearson. Tests for departure from normality. empirical results for the distributions of $b^2$ and $\sqrt{b}$. *Biometrika*, 60(3):613–622, 1973.

[DPATG21] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, 2021.

[EAZ$^+$19] Klaus Eckelt, Patrick Adelberger, Thomas Zichner, Andreas Wernitznig, and Marc Streit. Tourdino: A support view for confirming patterns in tabular data. In *EuroVA@ EuroVis*, pages 7–11, 2019.

[EGS$^+$15] Sebastian Echegaray, Olivier Gevaert, Rajesh Shah, Aya Kamaya, John Louie, Nishita Kothary, and Sandy Napel. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using ct images of hepatocellular carcinoma. *Journal of Medical Imaging*, 2(4):041011–041011, 2015.

[ESBB98] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[Fir] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.firstorder`. First Order Features.

[FVG$^+$20] Francesco Fiz, Luca Viganò, Nicolò Gennaro, Guido Costa, Ludovico La Bella, Alexandra Boichuk, Lara Cavinato, Martina Sollini, Letterio S Politi, Arturo Chiti, et al. Radiomics of liver metastases: a systematic review. *Cancers*, 12(10):2881, 2020.

[FWZ+19] Sirui Fu, Jingwei Wei, Jie Zhang, Di Dong, Jiangdian Song, Yong Li, Chongyang Duan, Shuaitong Zhang, Xiaoqun Li, Dongsheng Gu, et al. Selection between liver resection versus transarterial chemoembolization in hepatocellular carcinoma: a multicenter study. *Clinical and Translational Gastroenterology*, 10(8), 2019.

[GA12] S Gunasundari and M Suganya Ananthi. Comparison and evaluation of methods for liver tumor classification from ct datasets. *International journal of computer applications*, 39(18):46–51, 2012.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Git] `https://github.com/mbront/BA-HCC-Liver-Clustering`. GitHub repository of the implementation.

[Graa] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glcm`. Gray Level Co-occurrence Matrix (GLCM) Features.

[Grab] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.gldm`. Gray Level Dependence Matrix (GLDM) Features.

[Grac] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glrlm`. Gray Level Run Length Matrix (GLRLM) Features.

[Grad] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glszm`. Gray Level Size Zone Matrix (GLSZM) Features.

[Heaa] `https://seaborn.pydata.org/generated/seaborn.heatmap.html`. Heatmaps in Python via Seaborn.

[Heab] `https://plotly.com/python/heatmaps/`. Heatmaps in Python via Plotly.

[How] `https://umap-learn.readthedocs.io/en/latest/how_umap_works.html#`. How UMAP works.

[HR02] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

[HVGS+09] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009.

[ID90] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the first IEEE conference on visualization: visualization90*, pages 361–378. IEEE, 1990.

[Inc15] Plotly Technologies Inc. Collaborative data science, 2015.

[JBF+19] Daniel Jönsson, Albin Bergström, Camilla Forsell, Rozalyn Simon, Maria Engström, Anders Ynnerman, and Ingrid Hotz. A visual environment for hypothesis formation and reasoning in studies with fmri and multivariate clinical data. In *Eurographics Workshop on Visual Computing for Biology and Medicine*, 2019.

[KGB+12] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.

[KLR+13] Paul Klemm, Kai Lawonn, Marko Rak, Bernhard Preim, Klaus D Tönnies, Katrin Hegenscheid, Henry Völzke, and Steffen Oeltze. Visualization and analysis of lumbar spine canal variability in cohort study data. In *VMV*, pages 121–128, 2013.

[Koe] Will Koehrsen. `https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0`. Histograms and Density Plots in Python.

[KOJL+14] Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive visual analysis of image-centric cohort study data. *IEEE transactions on visualization and computer graphics*, 20(12):1673–1682, 2014.

[KS98] Alex BW Kennedy and H Riall Sankey. The thermal efficiency of steam engines. report of the committee appointed to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam engines: With an introductory note.(including appendixes and plate at back of volume). In *Minutes of the Proceedings of the Institution of Civil Engineers*, volume 134, pages 278–312. Thomas Telford-ICE Virtual Library, 1898.

[LLP+20] Antoine Lamer, Gery Laurent, Sylvia Pelayo, EL Amrani, Emmanuel Chazard, and Romaric Marcilly. Exploring patient path through sankey diagram: a proof of concept. *Studies in health technology and informatics*, 270, 2020.

[Lou73] Toussaint Loua. *Atlas statistique de la population de Paris.* J. Dejey & cie, 1873.

[LRVL$^+$12] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[Lys22] S Lysdahlgaard. Comparing radiomics features of tumour and healthy liver tissue in a limited ct dataset: A machine learning study. *Radiography*, 28(3):718–724, 2022.

[MEK$^+$19] Ali Morshid, Khaled M Elsayes, Ahmed M Khalaf, Mohab M Elmohr, Justin Yu, Ahmed O Kaseb, Manal Hassan, Armeen Mahvash, Zhihui Wang, John D Hazle, et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence*, 1(5):e180021, 2019.

[MFM$^+$21] AW Moawad, D Fuentes, A Morshid, AM Khalaf, MM Elmohr, A Abusaif, JD Hazle, AO Kaseb, M Hassan, A Mahvash, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation [data set]. *The Cancer Imaging Archive*, 2021.

[MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[Mül11] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[MWLH$^+$20] Eric Mörth, Kari Wagner-Larsen, Erlend Hodneland, Camilla Krakstad, Ingfrid S Haldorsen, Stefan Bruckner, and Noeska N Smit. Radex: Integrated visual exploration of multiparametric studies for radiomic tumor profiling. In *Computer Graphics Forum*, volume 39, pages 611–622. Wiley Online Library, 2020.

[Nei] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.ngtdm`. Neighbouring Gray Tone Difference Matrix (NGTDM) Features.

[Per] `https://umap-learn.readthedocs.io/en/latest/performance.html#performance-comparison-of-dimension-reduction-impleme` Performance Comparison of Dimension Reduction Implementations.

[PL20] Bernhard Preim and Kai Lawonn. A survey of visual analytics for public health. In *Computer Graphics Forum*, volume 39, pages 543–580. Wiley Online Library, 2020.

[PS15] SGOPAL Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Rad] `https://pyradiomics.readthedocs.io/en/latest/features.html`. Radiomic Features.

[RCMA+18] Renata G Raidou, Oscar Casares-Magaz, Artem Amirkhanov, Vitali Moiseenko, Ludvig P Muren, John P Einck, Anna Vilanova, and M Eduard Gröller. Bladder runner: Visual analytics for the exploration of rt-induced bladder toxicity in a cohort study. In *Computer Graphics Forum*, volume 37, pages 205–216. Wiley Online Library, 2018.

[Rem] `https://pyradiomics.readthedocs.io/en/latest/removedfeatures.html`. Removed Features.

[RHF05] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 233–240, 2005.

[RSR20] Munipraveena Rela, Nagaraja Rao Suryakari, and P Ramana Reddy. Liver tumor segmentation and classification: A systematic review. *2020 IEEE-HYDCON*, pages 1–6, 2020.

[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. In *Computer Graphics Forum*, volume 34, pages 11–20. Wiley Online Library, 2015.

[San] `https://plotly.com/python/sankey-diagram/`. Sankey Diagrams in Python via Plotly.

[San96] Henry Riall Sankey. The thermal efficiency of steam-engines.(including appendixes). In *Minutes of the Proceedings of the Institution of Civil Engineers*, volume 125, pages 182–212. Thomas Telford-ICE Virtual Library, 1896.

[SCGM00] John Stasko, Richard Catrambone, Mark Guzdial, and Kevin McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies*, 53(5):663–694, 2000.

[Sch08a] Mario Schmidt. The sankey diagram in energy and material flow management. *Journal of Industrial Ecology*, 12(1):82–94, 2008.

[Sch08b] Mario Schmidt. The sankey diagram in energy and material flow management: part ii: methodology and current applications. *Journal of industrial ecology*, 12(2):173–185, 2008.

[Scia] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering`. Scikit-learn agglomerative clustering.

[Scib] `https://scikit-learn.org/stable/modules/clustering.html`. Scikit-learn clustering.

[Scic] `https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering`. Scikit-learn hierarchical clustering.

[Scid] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html#scipy.stats.normaltest`. Scipy normal test documentation.

[Seaa] `https://seaborn.pydata.org/generated/seaborn.displot.html`. Seaborn displot.

[Seab] `https://seaborn.pydata.org/generated/seaborn.kdeplot.html`. Seaborn kernel density estimate plot.

[SH19] Shammamah Hossain. Visualization of Bioinformatics Data with Dash Bio. In Chris Calloway, David Lippa, Dillon Niederhut, and David Shupe, editors, *Proceedings of the 18th Python in Science Conference*, pages 126 – 133, 2019.

[Shaa] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.shape2D`. Shape Features (2D).

[Shab] `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.shape`. Shape Features (3D).

[Shn92] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, jan 1992.

[SMB$^+$10] Martijn D Steenwijk, Julien Milles, M Buchem, J Reiber, and Charl P Botha. Integrated visual analysis for heterogeneous datasets in cohort studies. In *IEEE VisWeek workshop on visual analytics in health care*, volume 3, 2010.

[SS20] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

[Sun] `https://plotly.com/python/sunburst-charts/`. Sunburst Charts in Python via plotly.

[SW65] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[Tra] `https://umap-learn.readthedocs.io/en/latest/transform.html#transforming-new-data-with-umap`. Transforming New Data with UMAP.

[Tre] `https://plotly.com/python/treemaps/`. Treemap Charts in Python via Plotly.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[VF05] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005.

[VGFP$^+$17] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[VGO$^+$20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[WBO$^+$17] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[WF09] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[WJG+20] Jingwei Wei, Hanyu Jiang, Dongsheng Gu, Meng Niu, Fangfang Fu, Yuqi Han, Bin Song, and Jie Tian. Radiomics in liver diseases: Current progress and future opportunities. *Liver International*, 40(9):2050–2063, 2020.

[YCW+20] Cynthia SY Yeung, CL Chiang, Natalie SM Wong, SK Ha, KS Tsang, Connie HM Ho, B Wang, Venus WY Lee, Mark KH Chan, and Francis AS Lee. Palliative liver radiotherapy (rt) for symptomatic hepatocellular carcinoma (hcc). *Scientific Reports*, 10(1):1254, 2020.