FINAL PROJECT REPORT
MULTIVARIATE DATA ANALYSIS OF LIVER CANCER SUBTYPES

Raphaël Romand-Ferroni
CentraleSupélec
raphael.romandferroni@student-cs.fr

ABSTRACT

Liver cancer, characterized by its heterogeneity and high mortality rate, ranks as the third most common cause of cancer-related death globally. Despite advancements in imaging technologies, significant challenges remain in the early diagnosis and subtype differentiation of liver cancer, particularly in patients with advanced stage disease. This study leverages a comprehensive dataset of 147 hepatic tumors from real patients to explore the utility of radiomic analysis in distinguishing between the main liver cancer subtypes: **hepatocellular carcinoma** (CCK), **cholangiocarcinoma** (CHC), and **mixed tumors** (Mixtes). By employing multivariate analysis techniques - PCA, RGCCA nurtured by the work of Tenenhaus et al. (2014), and a reformulated version of Parallel Factor Analysis (PARAFAC) in the context of irregular sampling of the data as introduced in Sort et al. (2023) - a large set of extracted radiomic features and 3D data at various phases, this project aims to enhance the diagnostic accuracy and potentially guide personalized treatment strategies based on phenotypic and proteogenomic information inferred from imaging data. The link to the GitHub repository can be find here: https://github.com/LaFerraille/Liver-cancer-detection

1 INTRODUCTION

Liver cancer presents a profound challenge to public health systems worldwide due to its increasing incidence and the complexity in its clinical management. The disease is often diagnosed at a late stage where surgical options are limited, emphasizing the necessity for improved non-invasive diagnostic techniques. This project addresses the critical need for precise characterization of liver cancer subtypes through the use of advanced imaging and data analysis techniques. Radiomics, which extracts quantitative measures from images, provides a promising approach to capture the tumor's phenotypic characteristics in a non-invasive manner, offering a potential pathway for early diagnosis and subtype differentiation.

Three primary subtypes of liver cancer—Hepatocellular Carcinoma, Cholangiocarcinoma, and Mixed Tumors—exhibit distinct prognoses and require different management strategies. This research utilizes contrast-enhanced imaging data acquired at various phases - **arterial** (ART), **portal** (PORT), **venous** (VEIN) and **late** (TARD) - to investigate whether these subtypes can be effectively differentiated based on radiomic profiles. Ultimately, this project seeks to establish a methodological framework for the enhanced differentiation of liver cancer sub-types, driving forward the fields of oncological imaging and precision medicine.

2 DATA EXPLORATION

The dataset comprises 147 hepatic tumor instances collected from patients, reflecting a real-world scenario of liver cancer heterogeneity. Each tumor has been characterized through multiple imaging phases, capturing a wide array of radiomic features, including first-order statistics, shape-based features, and several textural descriptors derived at different contrast phases. These features encapsulate

critical information about the tumor's density, signal profile, cellular structure, and morphological irregularities.

2.1 Patient Demographics and Clinical Information

A first file contains essential demographic and clinical information, including unique identifiers for each patient, gender, age at the time of liver disease diagnosis, dates of MRI, surgical operations, relapses, and death. The dataset also includes alpha foetoprotein levels, which are particularly relevant for diagnosing Hepatocellular Carcinoma (CHC) as demonstrated in Figure 4.

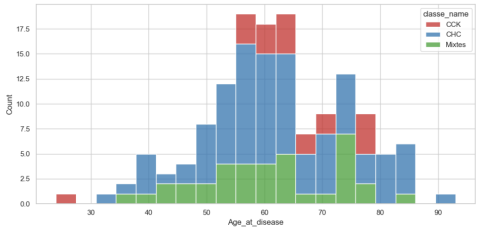

Figure 1: Distribution of Age Among Liver Cancer Patients by Subtype

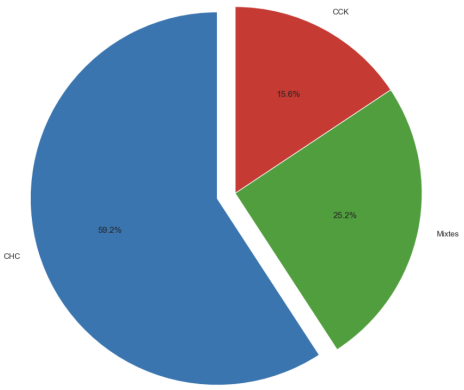

Figure 2: Distribution of Liver Cancer Subtypes

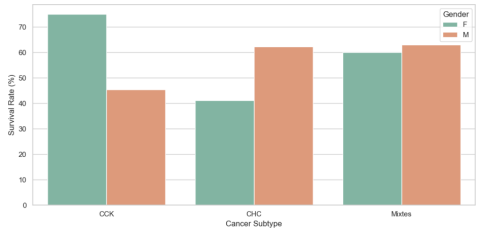

Figure 3: Percentage of Survivors in Each Cancer Subtype by Gender

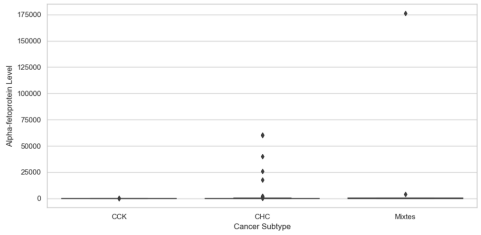

Figure 4: Alpha-fetoprotein Levels by Liver Cancer Subtype

The Figure 1 shows that liver carcinoma subtypes are not proper to one group of age in particular. They all occur mostly around 60 years old for patients. As illustrated in Figure 2, the Cholangiocarcinoma (CHC) subtype is the most represented in this cohort of patient. The fact that the level of alpha-foetoprotein is highly correlated with the presence of CHC will greatly help us in detecting this type of carcinoma as depicted in Figure 4. Eventually, we show in Figure 3 that there is a bias in the gender distribution of carcinoma subtype surival rates: men are more prone to survive when affected by CHC and mixed tumors but that is not the case for CCK which kills women more than men.

We also want to see what the descriptive features that are the most correlated to each of the subtype carcinoma. To do this, it involved transforming the categorical data 'Gender' into numerical binary columns and cleaning the dataset by removing NaN values and outliers (one outlier for Alpha-foetoprotein that exacerbated mixed tumors correlation). We thus computed these correlations in 5 for 99 patients with complete data.

As expected for CHC subtype (middle plot), the alpha-foetoprotein has the highest positive correlation with the presence of this carcinoma. The left plot indicates to the survival rate is positively correlated with the presence of CCK subtype, where it's the distant relapse for the mixed tumors.

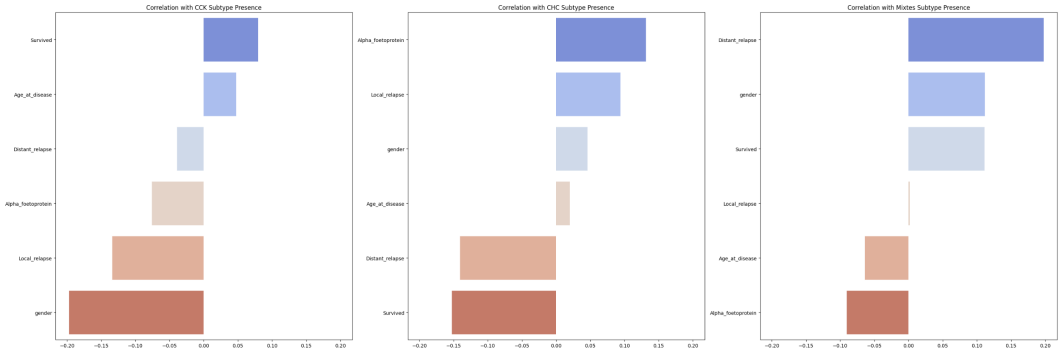

Figure 5: Correlation with each Carcinoma Subtype Presence

2.2 Radiomic Feature Overview

2.2.1 Global Radiomic

The radiomic data is in two files and includes first order features, shape-based features, and several groups of texture features such as GLCM, GLRLM, GLSZM, GLDM, and NGTDM. These features capture basic statistical, geometric, and textural properties of the tumors, providing a multi-faceted view of the tumor environment.

For the global radiomic data - which contains theoretically four injection phases for every patients - we compute the density function for the firstorder features by subtype and phase of injection. Particularly here in the case of Variance we see different behaviours depending on the carcinoma subtypes in Figure 6: CCK and mixed tumors exhibit similar density properties across the different phases after the injection of the contrast product, but it's not the case for CHC which exhibit thinner tails and less kurtosis across phases. Plus, CHC variance density appears to be centered around zero with a sharp peak, indicating that the values don't deviate much from zero. The presence of these peaks suggests that most of the tumors may have a similar texture characteristic regarding time. The same applies for texture features e.g with coarseness of the image in Figure 7 where CCK density progressively goes over CHC density over time.

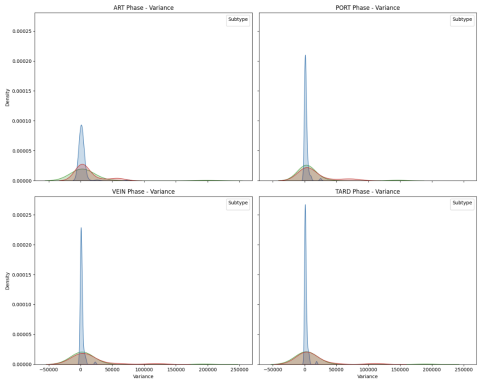

Figure 6: Density Plot of First Order Variance by Subtype and Phase

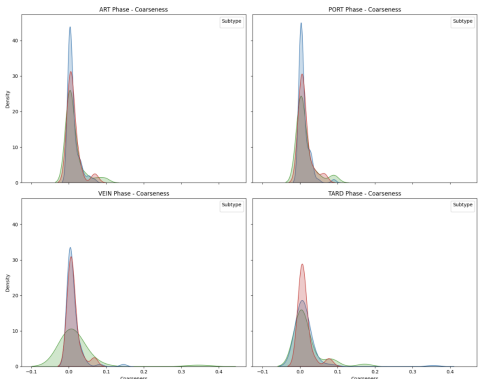

Figure 7: Density Plot of Texture Coarseness by Subtype and Phase

To perform these plots, we must ensure that every patient ID is represented by it's four phases. We get rid of "incomplete" patient ID for computation purposes (e.g we keep only **117** patients). We also get rid of diagnostic columns that are not useful here.

2.2.2 MULTISLICE RADIOMIC

For the multislice data we add the slice dimension whose range can vary from 1 to 90 per injection phase and we spot that the number of slices may not be consistent for every phases at each patient ID. We thus first filter on the patient ID with an equal number of slices across each injection phase. We get **49** patients left. Then, as illustrated in Figure 9, the ranges may vary across each patient ID. To enhance the dataset's utility for specific analyses, we implemented a function to extract only the central ten slices from the range of each patient ID, focusing on the middle segment to potentially capture the most representative data from each patient's scans. To perform this, we must have first filtered on the patient having a range of slices above 10 as illustrated in the complementary cumulative distribution plot 8. The horizontal red dotted line represent the number of patient that we keep in the last consistent dataset e.g **33** here.

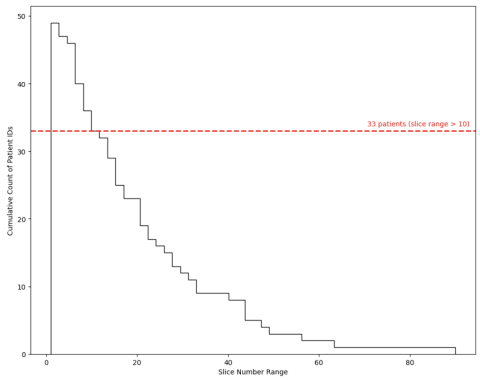

Figure 8: Complementary Cumulative Distribution of Slice Number Ranges

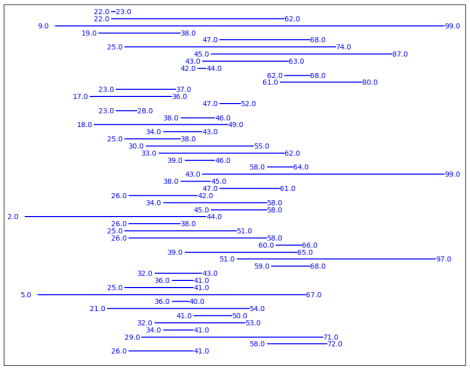

Figure 9: Slice Number Ranges for Each Patient

Given the high-dimensional nature of our dataset, it becomes prudent to consider dimensionality reduction techniques. That's we are going to investigate in the next sections.

3 DATA PREPROCESSING

To perform classical Machine Learning prediction for this mutliclass classification task, we had to perform some data processing to ensure data consistency.

We developed two functions to transform our datasets into matrices of features (X) and labels (y), tailored for both a global dataset and a multislice dataset. In the global dataset, initially containing 107 features, each record contains multiple observations under different injection time. Thus, each original feature generates four new features, one for each injection time. The global feature matrix (X) contains $107(originalfeatures) \times 4(injectiontime) = 428$ new features. Similarly, the multislice feature matrix eventually contains $93(originalfeatures) \times 10(slices) \times 4(injectiontime) = 3720$ new features.

3.1 NAIVE RESULTS

We trained 5 different models (Logistic Regression, Random Forest, XGBoost, KNN, and a final Voting Classifier with hard voting condition) on the global feature matrix (X). After scaling of the data, here are the results obtained in 10

The Logistic Regression underperform the random baseline of 0.46 and both the Random Forest and Voting Classifier obtain accuracy scores above 0.6.

The high dimensionality of the feature matrices in our study presents a challenge for these traditional machine learning models. Here the curse of dimensionality may impact negatively the performance of our models due to overfitting and long computation times.

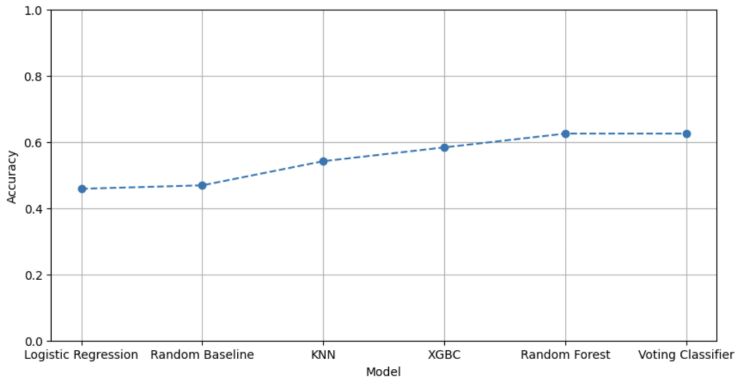

Figure 10: Comparison of Model Accuracies

3.2 Principal Component Analysis (PCA)

Let's work here with the global feature matrix of dimension 117×428. The PCA biplot on Figure 11 shows the distribution of liver cancer subtypes (each defined by a color) based on their location in the first two principal components. Principal Component 1 (PC1) accounts for 39.40% of the variance, and Principal Component 2 (PC2) accounts for 21.48% of the variance. The majority of points are clustered around the center of the plot, which suggests that most of the variability can be explained within the central region of the PCA space. Regarding the green point (mixed tumor labeling) in the extreme bottom right, it's an outlier in terms of its position relative to the rest of the data.

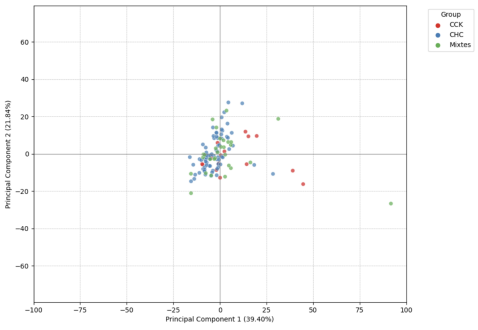

Figure 11: PCA Factorial Plan of Global Features

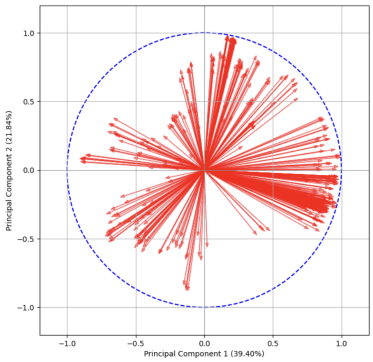

Figure 12: PCA Loading Plot of Global Features

Here the loading plot on 12 provides a visual representation of how each feature in the original dataset contributes to the principal components created by the PCA process. Features that point in similar directions are positively correlated, while those in opposite directions are negatively correlated. For the top 6 features we can see that the most important features are 'original_firstorder_Range' and that their is strong correlation between features of the same type but different time occurence (e.g. ART, PORT, VEIN or TARD).

Regarding the optimal number of components to choose, we can empirically choose the number of **15** components here based on the elbow method applied to the scree plot 13. Four components here add up to a substantial cumulative variance, around 90%.

Adding PCA to the training pipeline, we get these new results highlighted in Figure 14. We see that the best model outperform (66% of accuracy) the previous best model without PCA (62.5%) as well as for the third and fourth best models.

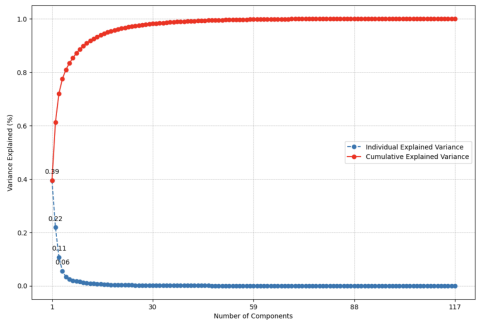

Figure 13: PCA Scree Plot of Global Features

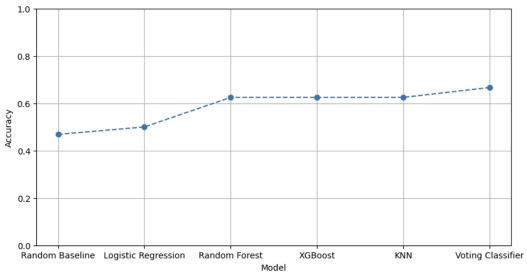

Figure 14: Comparison of Model Accuracies with PCA

4 Regularized Generalized Canonical Correlation Analysis (RGCCA)

4.1 RGCCA

In our analysis, we employ the Regularized Generalized Canonical Correlation Analysis (RGCCA) method, introduced by Tenenhaus et al. (2014). This approach allows us to consider multiple sets of variables simultaneously, instead of analyzing each feature individually. Specifically, we operate on J data blocks $X_1, \ldots, X_J$ where $J = 7$. Each $n \times p_j$ matrix X_j represents a block of p_j variables measured for n subjects. RGCCA seeks to find, for each block X_j, a linear combination $y_j = X_j a_j$ that best summarizes the information relevant to the analysis. The objectives of RGCCA are twofold: (i) each block component y_j should capture a substantial amount of the variance within its own block, and (ii) components of connected blocks should be highly correlated. These conditions ensure that the block components not only represent their own data well but also share information in a structured and meaningful way. RGCCA is defined as the following optimization problem:

$$\underset{\mathbf{a}_1,\ldots,\mathbf{a}_J}{\text{maximize}} \sum_{j,k=1}^{J} c_{jk} g\left(\text{cov}\left(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k\right)\right) \text{ s.t. } \left(1-\tau_j\right)\text{var}\left(\mathbf{X}_j\mathbf{a}_j\right) + \tau_j \left\|\mathbf{a}_j\right\|^2 = 1$$

where g is a continuous convex function that we set here to x^2.

Here we study specifically the consistent high-dimensional multislice feature matrix of shape 31×3720. The multislice feature matrix can be divided in two distinct block (it doesn't have the shape information): the first-order information and the texture information. The correlation inside each block is illustrated in Figure 15 where $X1$ is the block corresponding to the firstorder information per slice and $X2$ the texture information (GLCM, GLRLM, GLSZM, GLDM, and NGTDM) per slice.

We thus take $J = 3$ where the third block is our subtype carcinoma labeling encoded with 0 and 1. We append this third block consisting of three columns to our big feature matrix. Plus, we are looking for components y_1 and y_2 that explain well their own block X_1 and X_2 and that are correlated with y_3 which is not suppose to explain its own block. Consequently, we define τ as follows: $\tau_1 = 1, \tau_2 = 1, \tau_3 = 0$. The design matrix which encode the structural scheme between the three block will be a diagonal matrix with zero everywhere except for the position $X_1 <-> X_3$ and $X_2 <-> X_3$.

Using the R package (RGCCA), we plot the corresponding factorial plan on Figure 16. It's still hard to distinguish between the three classes, even more between CCK and CHC.

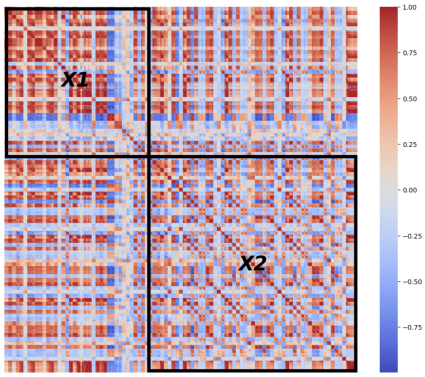

Figure 15: Correlation matrix of the multislice features

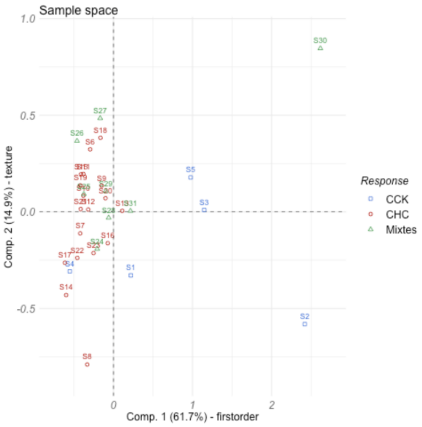

Figure 16: RGCCA Factorial Plan

4.2 Sparse RGCCA

The very high dimensional block settings of the multislice feature matrix make us looking for some sparsity in the data. We now apply the RGCCA to the $J = 3$ blocks but now introducing a sparsity vector. The elements of the sparsity vector of size are as follows in our case: $s_1 = 0.8$, $s_2 = 0.1$ and $s_3 = 1$. Each element of the sparsity vector is identical across the constraints applied to the block weight vectors associated to block X_j:

$$\forall k, \|a_{j,k}\|_1 \leq \text{sparsity}[j]\sqrt{p_j}.$$

Here we constrain much more the texture block which contain a huge amount of features to some sparsity.

We obtain the following factorial plan for the sparse method of RGCCA on Figure 18. Here we can clearly distinguish and separate the two sets of classes CCK and CHC but it's still hard to clearly distinguish mixed tumors (Mixtes) and CHC. A SVM model could be applied to this last scatterplot to have a straight line separating the two classes.

5 Multiway Tensor Rank Decomposition CP/PARAFAC

The paper by Sort et al. (2023) proposes an extension of the PARAFAC model to accommodate data that spans across three axes: individual, magnitude, and time. This framework allows for the description of each individual through a collection of longitudinal variables, each characterized by its unique temporal sampling.

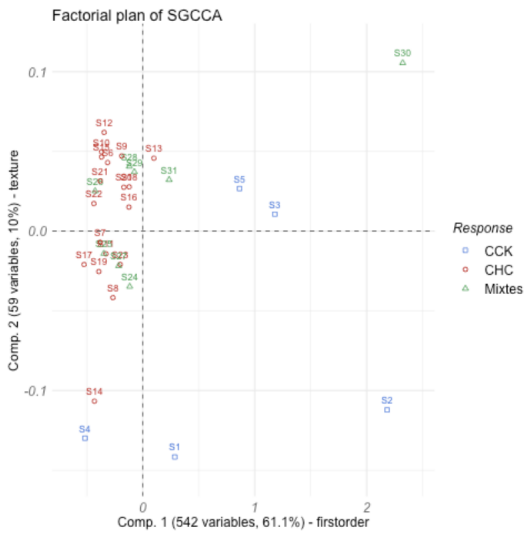

Figure 17: Sparse RGCCA Factorial Plan

Here, even with irregular and sparse sampling like in our case with inconsistent number of injection phases and slices across the data, the estimation of factors and canonical functions can remain robust, provided that the sample size of individuals is sufficiently large.

This functional rendition of PARAFAC hinges on a novel formulation that depends solely on covariance functions, as opposed to raw data. These functions are inferred through techniques of aggregation and smoothing.

We decide here to work solely on the global feature matrix, we set $i \in \{1, \ldots, 147\}$ denoting the index of all the individuals, and $k \in \{1, \ldots, 107\}$ denotes the index of features. For a specific individual and variable, let $t_{i,k}$ represent the vector whose elements are the $q_{i,k}$ time points of observation. For example, $t_{i,k}$ might be a vector of length 1, 2, 3, or 4, corresponding to the respective phases: ART, PORT, VEIN, and TARD. We thus convert our global feature matrix into a three-way tensor of shape $(ID \times features \times phases)$. We conduct a tensor factorization to find latent structures within the high-dimensional dataset. Setting rank to 2 we find the following plot:

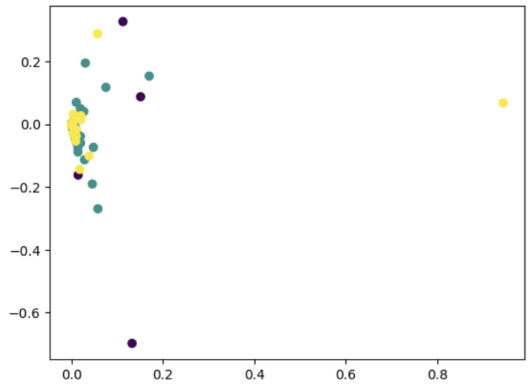

Figure 18: PARAFAC Factorial Plan

Here again the visualization doesn't provide any advancements adding information to the study that we performed with RGCCA and Sparse RGCCA.

6 CONCLUSION

In conclusion, despite the great heterogeneity of the liver carcinoma subtypes, we demonstrated the potential of PCA, RGCCA, and PARAFAC methods in extracting meaningful patterns from high-dimensional imaging data. PCA helped reduce dimensionality and revealed intrinsic data structures, while RGCCA exploited the relational information between different feature sets and cancer subtypes to improve interpretability. Sparse RGCCA, in particular, provided clear subgroup separation, paving the way for more targeted analyses. Although the PARAFAC approach did not significantly enhance the differentiation power beyond RGCCA methods, it offered valuable insights into the way we treat interactions of features and the possiblity to overcome irregular and sparse sampling in data.

REFERENCES

Lucas Sort, Fabien Girka, Laurent Le Brusquet, and Arthur Tenenhaus. Décomposition parafac pour données longitudinales. July 2023. URL `https://hal.science/hal-04257688v1/document`.

Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jerome Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3): 569–583, 2014. doi: 10.1093/biostatistics/kxu001.