The analysis and preprocessing of raman spectra of glioma cells

Author: Joel Sjöberg 38686
Masters thesis in Computer Science
Supervisor: Luigia Petre
The Faculty Of Science And Engineering
Åbo Akademi University
2021

Contents

Foreword

Backword

Remember Hofstadters law: "It always takes longer than you expect, especially when you take into account Hofstadters law". Thank you to fellow: **Fellow0, Fellow1, Fellow2, Fellow3, ..., FellowN** where $N \to \infty$

Abstract

Errors are left in as an exercise to the reader

Chapter 1

Introduction

The mammalian brain contains so-called neurons and glial cells. Historically it was believed that the brain contained ten times more glial cells than neurons, but recent studies suggest the number of neurons are equal to the number of glial cells [1]. Glial cells were previously thought to be insignificant in terms of the brains computational functionality, only lending structural support to the neurons. Recent studies have disputed this and suggest their contribution to the nervous system is greater than once thought, though their actual function is still a matter of speculation. Glioma is a type of brain cancer which manifests within the glial cells and disrupts brain functions. The survivability of the cancer is extremely poor with a life expectancy of a few months without treatment to a few years depending on the patients health, the tumor type and grade; rarely do patients survive for five years [2, 3, 4]. Gliomas are categorized depending on their glial-cell of origin. There are four main types of glial cells (also called neuroglia or simply glia): oligodendrocytes, astocytes, ependymal cells and microglia. Oligendroglioma originates from oligodendrocytes, astocytoma from astocytes and ependymomas originate from ependymal cells. Furthermore, astocytoma-types may develop into glioblastoma multiforme (GBM), the most aggressive form of brain cancer; this may even communicate with microglia to increase tumor growth [5]. It is also possible for GBM to develop from other brain cells [2]. This cancer is particularly aggressive due to its quick reappearance in the brain only a short period after surgery [3]. The heterogeneity of GBM-cells further complicates the healing process, due to poor response to targeted treatments [6].

The World Health Organization (WHO) has defined four levels (or "grades") of cancer severity used to describe the cancer aggressiveness and tumor growth. Grades I and II are considered low-grade and grades III and IV are considered high-grade. Glioblastoma is cathegorized as a grade IV cancer [4, 7]; these grades are used to determine an appropriate prognosis and line of treatment. A study by Vi-

gneswaran et al. [7] suggested these grades could be divided further to better describe the features of the tumor and express versions with poor prognosis. This suggestion is also supported by Hirose et al. [8]. Ceccarelli et al. [9] introduce alternative subdivisions of these classes, which show promise in expanding knowledge about glioma tumors and aid in treatment selection. Such evaluations require in-depth knowledge about the tumor tissue and further examination, which may last for weeks after extraction. Ceccarelli et al. define the subdivisions by six distinct classes, labeled LGm1-6. Their analysis showed IDH mutations in LGm1-3, as the name suggests, IDH mutations refers to mutations in the IDH1 or IDH2 genes whihc encode for the enzymes IDH1 and IDH2. These mutations are shown to be significant in a variety of cancers, including glioma [10]. Furthermore, LGm4-6 were IDH wild-type, where a majority of tumors could be labeled as glioblastoma. IDH wild-type refers to IDH genes with no mutations, but they are often correlated with poor prognosis in high-grade glioma. These clusters are reinforced by the results produced by Vigneswaran et al. The process of determining a prognosis and a line of treatment has great promise in improving patient outcome and is essentially influenced by classifying tumors into these subdivisions.

This thesis is the result of a project whose purpose is to optimize the categorization process, based on a deep learning model capable of predicting tumor-types in a matter of minutes. The project relies on tissue from tumors extracted from 45 patients and scanned using Raman spectroscopy. Raman spectroscopy was invented by Chandrasekhara Venkata Raman and measures the vibrations of molecules by spectral analysis. This method can be executed fairly quickly and can provide chemical information from the spectral light. A laser emits a ray unto the tumor tissue, causing the energy level of the molecules within to change, which in turn changes their vibration. This vibration is gathered by the instrument and provides information regarding the molecular properties of the material [11, 12]. This spectra is the data which the model uses as training and testing data. The choice of using Raman spectra in this way is due to the method's success in previous studies of Raman spectra using machine learning [13, 14]. The use of Raman spectra is further motivated by Liu et al. [15], whose work show promise for deep learning models trained on raw Raman spectra. The advantage of this method in the context of multilabel classification seem considerable, when compared to other machine learning methods such as Support Vector Machines, Random Forest and K-nearest neighbor [15].

This thesis aims to analyze the spectra extracted from all patient samples in an attempt to automate outlier detection. The samples are examined by statistical methods designed specifically for that purpose. Hierarchical clustering and K-

means clustering are applied to the samples to divide the spectra into subsets which we find identifies many outliers. These results are examined in contrast to the results of a criterion for finding outliers in the data by the provider, after which the method most suitable for his purpose is used to remove the outliers. Following the removal of the outliers, we present a preprocessing pipeline which will be used to prepare the data for machine learning applications such as Artificial Neural Networks or Support Vector Machines. The features which best divide the data into the six LGm classes are extracted using f-classif. These features drastically reduces the size of the spectra which are analyzed for prognosis which in turn reduces the examination time. The thesis then aims to provide a clear way of preparing Raman spectra for machine learning applications and provides the most important features those spectra consist of. Suggestions and a discussion for how these methods may be improved, which alternative methods could be tested instead and eventual limits to this project are also given for future consideration.

The thesis is structured as follows, Chapter 2 presents the preliminary background for the statistical methods used in the project, along with the necessary mathematical definitions by which these methods are defined. Among these are the notion of the mean and standard deviation which are central to methods of analysis we use, this includes the analysis of variance (ANOVA) which is applied by the f-classif method for extracting features from the data. Understanding the underlying definitions and consequences is necessary to validate and confirm the results and as such, the chapter also presents the definition of supervised and unsupervised learning with short definitions on machine learning terminologies. The formal definitions of K-means clustering and Hierarchical clustering is given. In Chapter 3, we discuss the exploration methods in detail, to give further understanding of the data on which this project is based. The chapter begins by introducing the concrete shape of the data. Feature selection is applied to the data by the f-classif method and the results are examined. The majority of this chapter is based on the visual analysis of the outlier detection. This is done by applying the statistical methods and the clustering methods to the samples. The results of each method are analyzed in contrast to the criterion defined by the dataprovider in greater detail to form an argument for or against the method in question. The chapter ends by removing the outliers by the optimal method and performing feature selection once more on the data devoid of outliers. In Chapter 4, we present our suggestion as well as arguments for the preprocessing pipeline to prepare the data for machine learning. An Artificial Neural Network is created and trained on the curated data. The performance of the architecture is measured and presented. The thesis is concluded in Chapter 6, where we discuss improvements and suggestions to out methods. We also provide sugges-

tions for alternative methods used for feature selection and preprocessing for future study and tests.

Chapter 2

Theoretical Background

In this chapter we present the necessary mathematical concepts on which this thesis is based. The concepts are considered fundamental for understanding the methods applied in the project. It begins by covering necessary mathematical theory required to understand data representations and handling within Machine Learning (ML). It then proceeds by defining common concepts in machine learning and the subject of supervised and unsupervised learning.

2.1 Mathematical Foundations

2.2 Machine Learning

TODO: Rewrite this section to better match contents, might even be unnecessary if DL is not mentioned in the end Machine learning is the practice of computing models for relationships between sets of data. The field has garnered significant interest within academia and industry alike due to the promising result in applications for which deterministic algorithms have proven difficult or impossible to make. There are two paradigms for learning: Supervised learning (using labeled data to approximate models) and unsupervised learning (finding patterns within the data itself).

Models are used to great length within many scientific domains. Though each domain has defined this term differently, the definitions in the context of machine learning shall be used. In this context, a model is a data structure made out of constant parameters which may be performed on any input vector x to produce a prediction y

Definition: A model is an approximation of a desired function f which produces relevant results based on human definitions.

Mathematically a model may be represented as a collection of numbers M which may in turn be used to compute f for any given example. In the context of machine learning a set of parameters may be tuned during a learning process (or training process). These parameters are combined with samples of data through some mathematical procedure to effectively model a distribution from which the data was extracted. The equation below is an example of a n-dimensional object.

$$x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n$$

2.2.1 K-means Clustering

Clustering is an unsupervised learning method whose primary use is in grouping sets of data. In this thesis we consider the *K-means clustering* algorithm. The following is a formal definition of *K - means clustering* as defined by MacQueen [16]. Given a set of N-dimensional points (where $N \in \mathbb{N}$) E_N and a desired amount of partitions k of E_N, partition the elements of E_N into k sets, the subsets are stored in a superset S such that $S = \{S_1, S_2, ...S_k\}$. The partitioning of E_N is performed by randomly initializing k N-dimensional points as randomly selected points within E_N. We define the set V with elements v where v_i is the i:th cluster center where $i \in [0,k]$. The partitioning of the elements $x \in E_N$ into their respective partition S_i is performed by computing the closest cluster center $\forall_{x \in E_N}$. Let T_i be the set of $x \in E_N$ such that the distance from the element to the relevant cluster is minimal, T_i is defined by formula (2.1).

$$T_i = \{x : x \in E_N | (|x - v_i| \leq |x - v_j|)\} 0 \leq j \leq k \wedge j \in \mathbb{N} \tag{2.1}$$

For centers who share equal distance to any given x the cluster with the smallest index is chosen as the containing set. Let S_i^c denote the complement of set S_i, then the partitions $S_i \in S$ are defined by formula (2.2).

$$S_i = T_i \cap \bigcap_{j=1}^{(i-1)} S_j^c \tag{2.2}$$

A consequence to this definition is that outliers have a potential to drastically change the quality of the cluster outcomes [17]. To remedy this and the stochastic nature of the initialization process, the method is run several times on the same dataset, yielding the resulting clusters with minimal inertia. The problem *K-means clustering* attempts to solve is proven to be NP-hard [17, 18] but the algorithm itself has a time complexity of $O(n^2)$ [19].

2.2.2 Hierarchical clustering

Hierarchical clustering is a clustering method which has a deterministic process. Each cluster formed is based on the entire dataset in contrast to *K-means* which approximates clusters by performing small changes to the cluster centers. The method produces clusters by iteratively combining the closest clusters according to the given linkage criterion (defined in the sections below). The two primary strategies for forming clusters are *agglomerative* and *divisive*. Agglomerative clustering initializes one cluster for each data point and combines them in a hierarchy according to the linkage criterion until all clusters are part of the hierarchy. Divisive strategies initializes one universal cluster for all data points and proceeds to separate the points into distinct clusters according to the linkage criterion. The method proceeds until all data points are separated to their own cluster within the unifying hierarchy. The project described in this thesis uses the *agglomerative* strategy. All strategies depend on the distance measure and linkage criterion [20].

2.2.2.1 Distance metrics

Let u and v be vectors of the same dimension n. The *Euclidean distance* (also called *L2-distance*) measure can be used to measure distance between the vectors in euclidean space. The *Euclidean distance* is calculated by equation (2.3).

$$d(u,v) = \sqrt{\Sigma_i (u_i - v_i)^2} \tag{2.3}$$

The *Manhattan distance* (also called *L1-distance*) metric is also a viable alternative if the distance is to be measured in blocks. The distance is akin to finding a shortest path among blocks and is therefore calculated as expressed in equation (2.4).

$$d(u,v) = \Sigma_i |u_i - v_i| \tag{2.4}$$

Cosine similarity measures similarity between vector angles and suits situations where certain vectors are expected to be similar. Should the vectors be sizable in terms of dimensionality, this method will yield varying results, especially if the elements have significant variance in each dimension. It is calculated as expressed in equation (2.5).

$$d(u,v) = \frac{\Sigma_i u_i v_i}{\sqrt{\Sigma_i u_i^2}\sqrt{\Sigma_i v_i^2}} \tag{2.5}$$

2.2.2.2 Linkage Criteria

In order to measure distance between clusters it is essential to know between which points the distance should be measured, since clusters often consist of several points. Linkage criteria describes the method for determining how the distance metric will be applied. SKlearn define four criteria in the documentation [21]: Single linkage, complete linkage, average linkage and ward linkage. Depending on which criterion is applied the results may differ considerably, it is therefore vital to have a formal understanding of their application and consequences.

Single linkage goes through each pair of clusters measuring the distance among all points within one with respect to the other. The distance between these clusters is determined to be the distance between the two closest points. Let U be the elements in the first cluster and V be the elements of the second. The distance between the first and the second cluster is defined formally in equation (2.6).

$$d(U,V) = \forall_{u,v \in U,V} min(d(u,v)) \tag{2.6}$$

Single linkage tend to produce trivial results, forging a hierarchy in a chain where individual elements slowly merge with the bigger cluster. In contrast complete linkage considers the largest distance between two points for every pair of clusters. The distance between two clusters then become the distance between the points which are the furthest apart. Formally expressed in equation (2.7).

$$d(U,V) = \forall_{u,v \in U,V} max(d(u,v)) \tag{2.7}$$

By considering the largest possible distance between two clusters it bypasses the setback by single linkage, allowing more clusters to form before merging into one unifying cluster. Average linkage calculates the average between all elements for every pair of clusters and merges the ones possessing to the minimal average distance. Formally described by equation (2.8).

$$d(U,V) = \frac{1}{|U||V|} \sum_{u}^{U} \sum_{v}^{V} d(U_u, V_v) \tag{2.8}$$

Ward linkage represents distance by how much the summed square would increase by merging them. The method aims to merge the clusters such that the within cluster variance is minimal. Let c_a be the center of cluster a, then ward linkage is expressed formally by equation (2.9) [22].

$$d(U,V) = \frac{|U||V|}{|U| - |V|} ||c_U - c_V||^2 \tag{2.9}$$

2.2.3 Feature Selection

In many cases the data available contains numerous features, which often helps to building sufficient classifiers as the model may find non-trivial patterns among the features. To avoid expanding the dependence on large datasets, it is often necessary to strip the data of certain features which possess minimal correlation to other features or which lack that correlation entirely [23]. Features that possess the necessary expressive information are not always trivial, there are several ways in which they may be found. **TODO: Expand this, it's been 3 weeks already!**

2.3 Deep Learning

The field of Artificial Intelligence is founded on the notion of designing algorithms for solving problems. The field encountered tremendous progress in **[FIND YEAR, AI FOUNDATIONS]** referred to by **[NAME]** as the "look ma, no hands" era of Artificial Intelligence. One such method which have proven useful for these tasks is the practice of approximating models through Artificial Neural Networks.

2.3.1 Artificial Neural Networks

Artificial Neural Networks ("ANNs") have been used to great success during the 20th century [**Source Here**]. With the use of ANNs several fields including Natural Language Processing, Encoding and Image classification have undergone revolutionary leaps in performance regarding optimization due to the predictive power of these networks [**Source Here**]. At the same time they are heavily criticized for their complexity, yielding a structure much more akin to a so called *"black box"* than a reliable and deterministic method for prediction[**Source Here**]. This complexity is due to numerous different structural typologies available at present and an awesome number of tuned parameters which are modified with the goal of minimizing the predictive error [**Source Here**].

A consequence of this is hard skepticism in regards to the correctness of their functionality within practical use. While these models have shown great promise when compared to their human counterparts, the question remain whether or not perfect performance can be yielded from the constructed models.

Definition 1. Training an ANN is allowing minuscule changes through the randomly initialized structure in order to approximate a collection of nested functions

$$f_n(f_{n-1}(...f_1(X)))$$

2.3.2 Computational Representations

The initial purpose of ANNs was to create a computational model of the human cortex which took the form of the McCulloch, Pitts neuron. The multilayer perceptron (MLP) introduced in **[year here]** formed the basic structure which would become ANNs.

Chapter 3

Data Exploration

Deep learning models require tremendous amounts of data, to ensure the model works well; the data must also possess sufficient characteristics to approximate the sample population from which it was extracted. To satisfy this requirement we examine the data in attempt to remove outliers and determine whether the data is sufficient for classification. Moreover, tumors have been shown to be heterogeneous [24], which may be problematic for a classifier as heterogeneous samples lack in shared characteristics. In this chapter the data available to the project is examined in greater detail; details for how the Raman spectra were prepared is given to document the preprocessing of spectra for future use. First we describe the mathematical representation of the samples. Since the number of samples is too small to use in a deep learning model, we explain how each sample may be separated into individual spectra; this separation yields a drastic increase in the number of available training examples. We then explain how to balance the data; an unbalanced dataset would likely introduce bias in machine learning models, rendering their desired predictive capabilities uncertain. We achieve this by duplicating underrepresented sample classes in the dataset. Furthermore, this balancing is performed to maintain majority and minority classes, thus retaining some distributional information from the original dataset. The quintessential purpose of this chapter is to analyze the data using different methodologies for detecting spectra which has been altered due to non-tumor material affecting the spectra (henceforth denoted outliers). As a starting point for this analysis, we have defined the frequency criterion. The criterion is used to separate spectra from outliers in each sample. The outliers captured by this criterion is compared to outliers detected by other outlier detection techniques such as the Standard Deviation Test (SDT) and the Interquartile Range Method (IRM). The unsupervised machine learning methods K-means clustering and Hierarchical clustering are also used to detect outliers in each sample. The results from each method are then compared in effort to select the one which best separates outliers

from the tumor spectra after which the data is curated by that method.

Another point of interest in this project is the identification of representative frequencies within the spectra. Each spectra belongs to a tumor which can be categorized by six different classes. The hypothesis states certain frequencies should be sufficient in determining which class the tumor belongs to. Feature selection is used for this purpose, representing each frequency within the spectra as distinct features. However, in order to extract such features reliably, the data must be devoid of outliers. Should outliers exist within the dataset, the features given by the methods used will be influenced and may yield conflicting results with the ground truth. To prepare for this the data is plotted for visual inspection using various methods to be discussed in the following section on feature selection. It is confirmed by the provider that the majority of samples include outliers. These outliers are influenced by a variety of other materials found on the tissues surface e.g. spectra of blood drops, plastic which may be reflected form underneath thin tissue or necrotic tissue, which is shown to affect the spectral signal significantly. Using the extracted features, a model can be trained on the data faster which can be essential many machine learning methods require a significant amount of computing resources, data and time. The features are extracted before and after the removal of problematic spectra. This is done to compare the impact of the outliers in feature selection. We expect the features vary significantly different subsets of the data available. Feature selection is not a good strategy for machine learning in this particular case if the features extracted from different subsets vary tremendously. In this case there are clear signs the machine learning models will fail to emphasize features which best represent the whole dataset; which is problematic since the model should have sufficiently accurate predictions for unseen data.

3.1 Data Representation

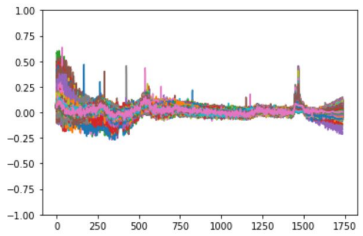
The data consists of the Raman-spectra extracted from the tissue of glioma tumors from 45 patients. Multiple samples of tissue were extracted from the same patient in some cases, yielding several samples for the respective patient. To maintain separation among the patients, the samples are sorted by their respective patient of origin. This is necessary due to the heterogeneity of each tumor. The data will be separated into three separate datasets. These sets are referred to as the training set, the validation set and the test set. Due to the expected heterogeneity, all datasets will consist of unique patients to avoid scenarios in which the model overfits to a patients tumor sample. This structure also allows for easier handling of the number of patients in each sample class, allowing for analysis on each class separately from

the others.

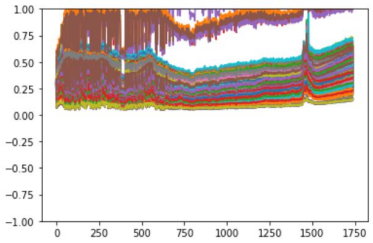
There is also large variation with regard to the sample shape within the data. Each sample is a 3-dimensional array of shape $(w, h, 1738)$ where w and h are the width and height of the sample, respectively. This formalization is necessary, as width and height have non-zero variance among different samples. The shape is a result of how the tissue was scanned. In each case the tissue was placed inside the Raman spectrometer and scanned successively from side to side. This makes it possible to display each sample as an image, by substituting the third dimension (denoted above by 1738) by a color value denoted by which class the spectra belongs to. The number 1738 is constant through all samples and represents the length of a modified Raman-spectra. The spectra has been modified by the data provider; the data consists spectra resulting of a linear mapping from the extracted spectra. This was done to omit unnecessary frequencies and allow focus on parts of the spectra which we believe is sufficient for this project. Furthermore, each element inside these arrays is a real number without clear bounds. The largest absolute element found within the complete dataset is 79427.0625. Some values are negative, which is confirmed by the providers to have significance for the projects purpose. The project aims to prepare these spectra for use in machine learning methods. Predictions should be performed on individual spectra extracted form the dataset. We choose this strategy since each spectrum is independent of all other surrounding spectra, ideally sharing in some characteristics from the other spectra belonging to the LGm class we wish to predict i.e. one spectrum should describe which LGm class the sample belongs to. There are six distinct LGm classes as defined by Ceccarelli et al. [9], denoted LGm1 - 6. The model will take as input one vector of shape $(1, 1738)$ and produce one vector of shape $(1, 6)$. This strategy is inspired by Liu et al.[15], who managed to get satisfactory performance by training a model on raw spectra i.e. spectra without preprocessing or outlier removal. Restructuring of the data to represent all samples as a list of spectra rather than 3-dimensional matrices yields a dataset with more than $300,000$ datapoints, which better suits deep learning tasks.

Initially, we choose to examine each sample by plotting the spectral lines in a plot. This allows us to examine the general shape and visually deduce if any common patterns are present in the samples and identify problematic samples. As expected from analogue measurements, significant amounts of noise are present in each plot. Despite this, many spectra share in some general characteristics with a few spikes appearing on an mostly flat spectral line. An example of the plots created for this examination be seen in Figure 3.1.

In Figure 3.1 is shown two collections of spectra, the spectra belonging to sam-



(a) Spectra from patient HF-1293. The rest of the samples available share in this pattern with some deviations



(b) Spectra from patient HF-1887. The frequencies tilt towards the upper part of the plot. The example is decidedly removed from the analysis.

Figure 3.1: Examples of samples drawn from the data, HF-1293 display a common pattern across all samples, HF-1887 is removed due to the skewed baseline. The range of the values are normalized between -1 and 1 to easier display the spectra.

ple HF-1293 are shown in Figure 3.1 (a) and spectra from sample HF-1887 are shown in 3.1 (b). The spectra form HF-1293 follow a general pattern visible in the vast majority of sample plots. Sample HF-1887 is an example of one sample which we deem too sporadic for this project, we remove it due to the skewed baseline of all spectra in the sample after confiding with the data provider, who agrees with our decision. After visual examination, we can confirm that some characteristics are present, but they are not sufficiently different among the different LGm classes to be classified by a human being in this case. Machine learning is needed to detect precise differences among the samples.

Another reason to remove samples are due to their size. The analysis we aim to perform requires considerable computational power and memory space. Many of the samples available have a manageable size, as they consist of approximately 3600 spectra i.e. width and height are approximately 60 and 60 respectively. In case the samples are too big for use we may simply extract a random sub sample from the larger sample and analyze those. But applying that method at this stage would potentially ruin the form of the samples which is used to evaluate the outlier detection methods in a later section. Fortunately only one sample suffers from this problem. Sample HF-3097, shown in Figure 3.2, shows a concerning number of spikes in contrast to the other samples and is the only sample which require considerable memory space (the number of spectra exceed 40000).

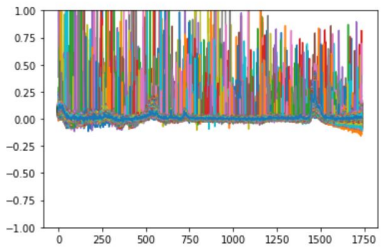


Figure 3.2: Sample HF-3097 from LGm2, the spectra spikes are extremely sporadic. The range of the values are normalized between -1 and 1 to easier display the spectra.

The values inside the spectra are also considerably bigger than the values found in the other samples. The methods SDT and IQM require statistical constants be extracted on the entire dataset, including this sample would affect the standard deviation and the mean of the frequencies tremendously (even if the constants were evaluated within the scope of the sample itself). K-means clustering is also a unsupervised machine learning model which can be trained on data before predictions are performed on unseen data. Since the aim is to use the training set for training the model, inclusion of the sample would affect the performance considerably. If the sample is put in the test set for testing, the noise in the signals would affect predictions considerably, which would affect testing accuracy due to what appears to be noise in the frequencies. To avoid these statistical issues and to reduce the computational requirements, the sample is discarded following the data providers approval.

3.2 Organization and Balance

The model will, as a consequence of it's learning-algorithm, become biased towards certain predictions. As the model encounters frequent examples of a certain class, the connections which produce such predictions will strengthen. Over exposure to examples of a certain class will force the model to associate features with that class, redirecting focus from classes for which that feature could be significant. This is why it is important to train models on balanced data. Having a balanced dataset means the number of examples in each class are uniform in the entire training set. Training on a balanced dataset also avoids bias towards data distribution, where certain classes may be predicted more frequently than others simply because the distribution was used in training. The initial data suffers heavily from this problem. Training on the data would result in a model which produces frequent predictions for the majority class (the class which has the most spectra) and perform inconsistently for the other classes. The data distribution is shown in Table 3.1.

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# of samples	5	11	4	10	11	4
# of spectra	37319	71846	31931	50660	62256	20176
percentage	14%	26%	12%	18%	23%	7%

Table 3.1: Table showing the distribution of data in the initial dataset after removing the problematic samples. The number of samples are displayed on the first row, the number of spectra in each class is shown on the second row. The percentage of the entire dataset is shown on the third row. The majority class is LGm 2 and minority is LGm 6. Classes LGm 1, 3 and 6 must be expanded to balance the data.

Table 3.1 shows the per class separation in the data, the header row shows the labels of each class. The first row shows the number of samples belonging to each class, these are the tumors which will be analyzed. The second row displays the total number of spectra across each class; these must be considered for balancing. Note the equal amount of samples in LGm3 and LGm6, but the difference in number of spectra within them. This is due to the varying size of all samples drawn from the tumors. Some samples share the same size, however the important fact is that the samples lack a uniform shape, which must be considered during the analysis. The last row shows the percentage each class makes of the entire dataset. Initially LGm2 is the majority class while LGm6 is the minority, consisting of only 7% of the entire dataset.

Before the data is balanced, the testing data is selected and separated from the rest manually. This is done by separating at least one patient and all their samples from the rest of the data. This way it will be possible to test if the model develops bias towards the patients in training and if the patient samples are heterogeneous with respect to the other samples of the same class. Samples are chosen with the criterion that approximately 30% of each class is represented in the test set. Balancing the classes which contain less elements by a factor larger than or equal to two compared to the majority class (LGm2) is done by repeating the spectrum in each sample by that factor. This methods does not perfectly balance the data to have a uniform distribution of classes, but it does make the underrepresented classes frequent enough to circumvent the issue of unbalanced datasets. The multipliers for each class are chosen so as to retain the majority class i.e. LGm2 will remain the majority class after balancing is done. There is also a need for the validation set, the validation will be used to monitor the models performance on unseen data during training. The samples for this set are chosen such that one sample from each class is allocated to the set. The separation of the entire dataset into three distinct sets is arbitrary and so can be performed automatically. However, we choose to perform this separation manually to maintain consistency in this analysis. The resulting sep-

aration is shown here as it is the training set from which the features are extracted in this project. We discover that the results will differ slightly depending on which samples are chosen for which set. The distributions of all datasets are shown in Table 3.2.

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# training	17689	47602	37557	30180	33396	9376
# validation	3600	4096	3600	4096	4096	3600
# test	14945	20140	11296	16384	24764	7200

[htb]

Table 3.2: Distribution of the three datasets

Table 3.2 shows the distribution of the different datasets used in this project. The training set is then balanced exclusively. This is not required in the test set or the validation set, since they will have no direct effect on how the model is developed through training. It is also important that the validation and test sets are not uniform, since it will prove whether the model can generalize to different prediction distributions. The training set is balanced by replicating each spectra in every patient of the classes which are under-represented. The number of sample replications per class can be computed by the following method. Let m be the LGm class which contains the majority of spectra in the set and $|LGm_n|$ be the number of spectra in LGm class n. The number of replications for each class can then be computed by $\lfloor \frac{|LGm_m|}{|LGm_n|} \rfloor$. The distribution of the training set after applying the replication method is shown in Table 3.3.

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# training	35378	47602	37557	30180	33396	46880

Table 3.3: Distribution of the testing data following balancing

As is shown, the data does not have a uniform distribution, LGm2 is still significantly larger than LGm4. The important thing is that distribution is in better balance relative to the original distribution. We believe this approach is good enough and proceed to the next stage.

3.3 Analysis

Following the balancing, the first step in the analysis is to find the frequencies which best describe the data with respect to the methylation-types. Each number in the

spectra is a frequency at which the scattered light is gathered. This light is expected to be sufficient for predicting the methylation-type of the tumor-tissue. It is speculated that, to sufficiently categorize the spectra into the methylation-types, only certain frequencies are required. For this reason, the best features are extracted with SelectKBestfeatures [21] which is given the f-classif method for ranking the features in order of their significance using ANOVA. Following this step, we may pick any arbitrary number of features by extracting them in the order given by f-classif. The 70 best features are extracted from the training set in which there are outliers present still. The features are displayed in Appendix B. The extracted features show that regions of interest do exist on the spectra. This can be seen by the integers which have a successive difference of one, suggesting that the region of interest exist somewhere in specific parts of the spectra. It is worth noting here that the features selected might be correct provided the amount of outliers is sufficiently small to be ignored by the feature selection method. Due to this uncertainty, the data will be separated from the outliers and feature selection will be performed a second time at the end of this section.

All data is subject to this analysis as all samples must be curated before the model can be trained. In the methods where statistical constants must be calculated, the training data is used, this avoids bias towards the other datasets whose primary purpose is to evaluate the model. The goal of the analysis is to find a uniform criterion which each spectrum must fulfill to be considered *clean*. Spectra which fail to satisfy this criterion will be discarded form the project entirely. In this section the methods of analysis used are described and their results examined. The section begins by examining the frequency criterion, which is a confirmed criterion all spectra must satisfy to be considered "clean". This criterion was provided by the data provider. SDT and IRM are compared to the frequency criterion for validation, these are deterministic methods that rely solely on the values found within the data and are commonly used to detect outliers in data. *K-means* clustering and Hierarchical clustering are then performed on the data in attempt to capture potentially complex patterns within the data. These methods are specifically designed to allow for grouping of data based on the similarities between data points within the dataset. The section ends by selecting the method which best detect outliers, this method is then added to the preprocessing stage i.e. following this project, all samples must be curated using this method before they can be used by the machine learning model.

3.3.1 The frequency criterion

The frequency criterion is a criterion specified by Adrian Lita for separating outlier spectra from tumor spectra. The criterion states: "should any value of frequencies between 1463 and 1473 of any spectrum be below 5000, then that spectrum is defined to be and outlier". Given that this criterion is defined by the provider, and the lack of knowledge regarding where outliers may be positioned on the samples, we choose this criterion as the starting point of this analysis. The extracted features in Appendix B include parts of the range on which the criterion is based, which is promising. The range being present in the extracted features can also mean that the outliers work to influence the relationship between the spectra and the methylation types we want to predict. In which case the removal of the outliers are essential for building an unbiased model. We examine the results of the criterion to gain insight into where these outliers are positioned. The criterion is confirmed to miss some outliers and so it mainly functions as an initial approximation of the areas where outliers are present. An example of this is in sample HF-2849 of LGm3 shown in Figure 3.3.

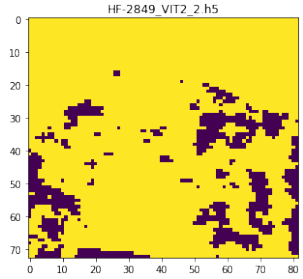


Figure 3.3: Sample HF-2849 from LGm3. The sample is confirmed to have necrotic tissue present in the upper part.

The criterion is satisfactory for capturing areas in the lower parts of the sample. However, the upper part of the sample is guaranteed to have spectra from necrotic tissue which is unreliable for describing the LGm class of the tissue. It appears the criterion is well suited for detecting liquid material on the tissue, since many samples show smaller spots of interest and fail to completely capture larger areas of outliers. This is evident in Figure 3.3, as multiple small spots are detected. This is one of several examples for how the criterion fails to detect all outliers reliably, but in the majority of the samples, the outliers are detected sufficiently well. One such case can be seen in sample HF-868 displayed in Figure 3.4.

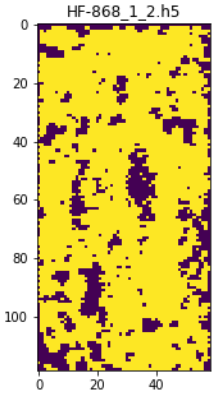


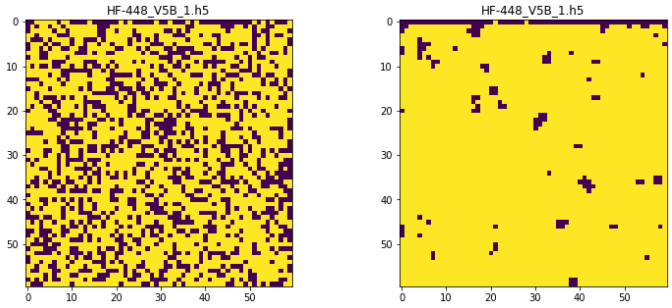
Figure 3.4: Sample HF-868 from LGm1. The areas detected are confirmed outliers.

According to the data provider, the outliers present in HF-868 are sufficiently captured by the frequency criterion. The outliers form as small, sporadic areas on the surface of the tissue. This shape is what the outliers are expected to have in the samples which are known to include them. However, since the frequency criterion fails to capture outliers in certain samples (possibly due to the material information of those outliers), we explore changes made to the criterion e.g. making the interval greater (checking frequencies ranging from 1458 to 1478) and checking for frequencies which are below 20000 within the interval the criterion concerns. The results vary greatly from the original criterion. Though some of the spectra originally ignored by the criterion now become visible, not all regions are sufficiently captured. This indicates the necessity to use other parts of the spectra which appear to be relevant for outliers. The spectra are too big for further manual inspection, which greatly motivates the application of machine learning methods. The frequency criterion will, however, serve as a ground truth in this analysis. Going forward in this section we aim to evaluate all following methods by comparing them to the results of the frequency criterion. Examples which best reflect the methods performance in outlier detection will be displayed for comparison with this criterion. This way, we have some knowledge about where outliers are detected. While the criterion is insufficient in detecting all outliers (as demonstrated in this subsection), the general patterns discovered here must be present in the results yielded by the other methods. If the method under analysis fails to produce results corresponding to the frequency criterion, we opt to disregard that method. The desired method optimally produces similar results as the frequency criterion and finds more areas on the samples where we know outliers are present. Ideally, the methods also aid in discovering new outliers.

3.3.2 The standard deviation test

The standard deviation test is a test by which the data is centered around the mean and given a standard deviation of one. With this setup, outliers are defined as points which are separated from the mean by three standard deviations or more. We measure the mean and standard deviation on each frequency from the unbalanced training-set; we ignore the balancing to avoid changes to the mean and standard deviation which the balancing helps produce. The values are then used to standardize the spectra belonging to each tumor. A spectrum is deemed to be an outlier if the number of frequencies in that spectrum exceed an arbitrary value. We approximated the value by performing the test once while monitoring the average number of frequencies which lie three or more standard deviations from the mean. In any given sample, each spectrum includes on average 111 frequencies which fail the test. We specify that a spectrum which fails the test on more than 111 frequencies is deemed an outlier and removed.

Using this test, many small areas are detected in each sample, it suggests there is something present in those places, but they do not possess a clear shape by which we can decide whether to discard them or not. An example of such a sample is shown in Figure 3.5.



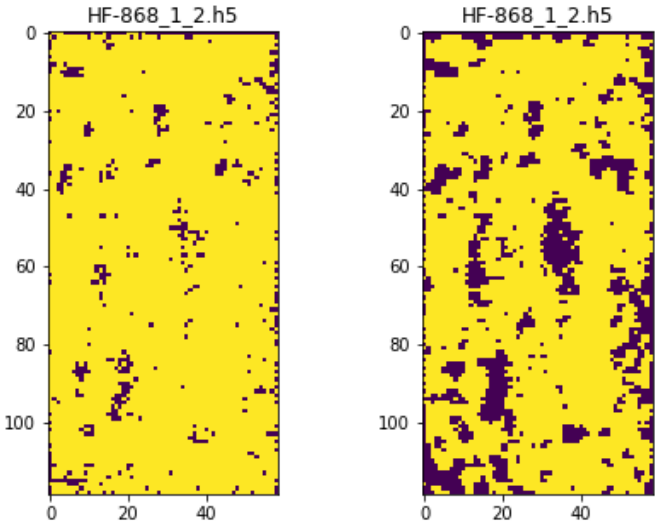
(a) The result of SDT for sample HF-448, Several points are labeled as outliers

(b) The frequency criterion outliers for HF-448

Figure 3.5: A comparison between the SDT and the frequency criterion for sample HF-448.

SDT does not produce results similar to the frequency criterion, some spectra do correspond among the results, but the amount of outliers falsely labeled as outliers by SDT is troublesome. Roughly 50% of the samples in the entire dataset yield similar results with SDT. The method would as a result of this, discard too much from most samples, depraving the dataset of a great number of tumor spectra. Some areas are formed around the individual spectra, suggesting the presence of an unknown material, but the sporadic points in the surrounding area make it unclear

where that material begins and ends. Furthermore, we must develop a criterion for which points to discard. Such a criterion would need to distinguish between sporadic points which are miss-classified and areas of real outliers. However, within some samples there are areas of outliers correctly labeled, but those areas are not sufficiently defined. One sample with this result is shown in Figure 3.6.

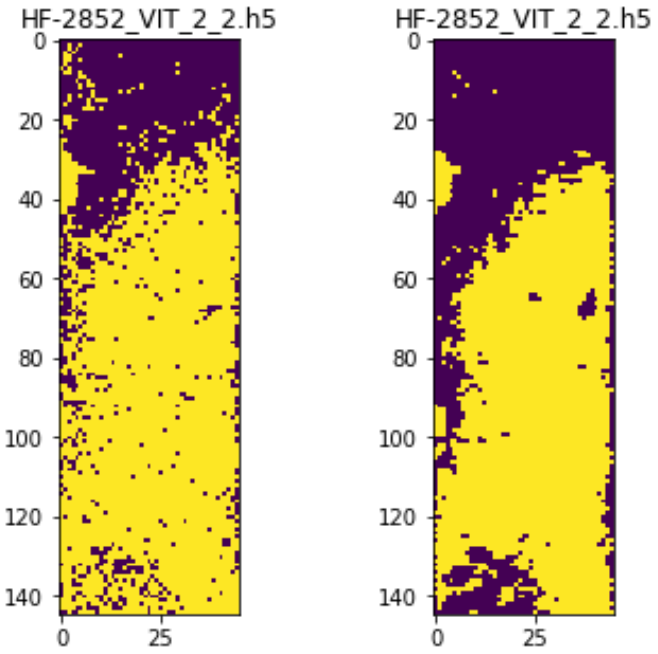


(a) The result of SDT for sample HF-868

(b) The frequency criterion outliers for HF-868

Figure 3.6: A comparison between the SDT and the frequency criterion for sample HF-868. Many spectra correspond to the frequency criterion.

Sample HF-868 is less sporadic compared to HF-448, the outliers are instead formed around common points of interest which strongly suggests outliers are present in that area. The lack of definition in each area however, is not sufficient. The outliers must form concrete shapes with clear definition. The sample suggests outliers are present in the different areas, but all points are not present to make the shapes as defined as they are by the frequency criterion. The sporadic results show that this method is insufficient to detect the majority of outliers in all samples. It must be noted that despite this underwhelming result, some outliers are detected, suggesting further changes to the method could yield better results, though greedy application is not going to work for all samples. Throughout all the six LGm classes, the method yields the best results for LGm6 where it shares many patterns with the patterns produced by the frequency criterion. Another example of the methods promise is in sample HF-2852 of LGm3 shown in Figure 3.7.



(a) The result of SDT for sample HF-2852

(b) The frequency criterion outliers for HF-2852

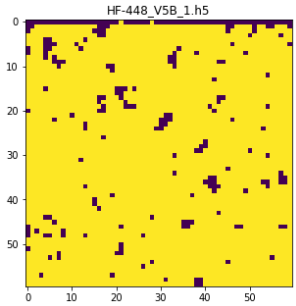
Figure 3.7: A comparison between the SDT and the frequency criterion for sample HF-2852.

Sample HF-2852 has a large area in the upper part which consists of necrotic tissue. The spectra of that tissue differs from other spectra sufficiently well, allowing the method to distinguish between tumor and non-tumor with considerable accuracy. This is due to the amount of otherwise healthy tissue present in the sample. The methods inability to capture all outliers is also present despite this, as it allows several smaller spots in the area of the necrotic tissue to be classified as healthy tumor tissue. The area at the bottom of the sample is filled with outliers according to the frequency criterion. These results suggests that the method is best used in detecting necrotic tissue, but not in context of detecting other kinds of outliers.

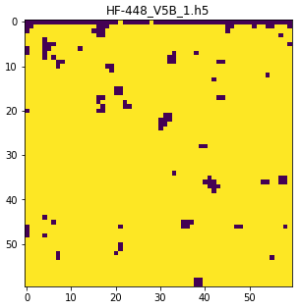
3.3.3 The Interquartile range method

Similar to SDT, IRM is a purely statistical analysis method which detects outliers in terms of which percentile the spectra belong to. The 25th and 75th percentiles are calculated on each frequency for the entire training set. Like SDT, this method yields a varying amount of outliers for each sample. We instead define the allowed number of outlier frequencies within one spectra to be equal to the average number of outlier frequencies within the analyzed sample. Many regions are better represented by IRM, showing well defined areas where outliers are clearly present. The

amount of individual spots are less frequent which shows promise in the method, as e.g. blood is expected to cover a larger area if present. The improvement from the standard deviation test is seen in Figure 3.8.



(a) The result of the interquartile range for HF-448, Several well defined areas are separated form the majority of spectra

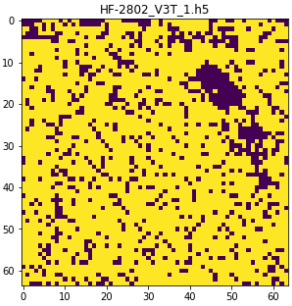


(b) The standard deviation test for sample HF-448

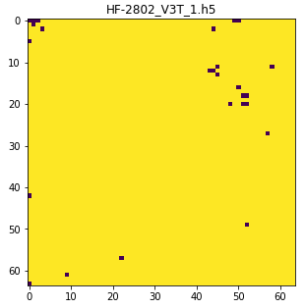
Figure 3.8: A comparison between IRM and the frequency criterion for sample HF-448.

SDT failed to separate outliers adequately in sample HF-448 while IRM is better suited to detect the outliers in this sample. The upper part of the sample has a well defined line where the sample is presumably cut, meaning the plastic underneath the tissue might be visible. We also note that many outliers detected by IRM also appear to be captured by SDT. While the areas are hard to distinguish among all sporadic spots, the larger spots in Figure 3.8 (a) seem to have some trace in Figure 3.5. The correspondence with the frequency criterion in Figure 3.8 (b) further shows the methods capability in contrast to SDT as the method seem to possess greater ability in defining the areas where outliers are present. There are still some spots appearing randomly around the sample surface which would ideally be ignored, but while we have some confirmation on where outliers are present in certain samples, we do not know exactly where the outliers are. The individual points being labeled as outliers suggests the method struggles with the same issue SDT suffers from. Though it appears to be less severe in all samples belonging to LGm1, there are still a considerable amount of spots through all samples within the class. Especially LGm4 have samples for which both methods perform poorly, with a considerable amount of sporadic spots appearing in some samples when applying IRM and a significant lack of spots when applying SDT i.e. None of the methods works sufficiently well to detect the outliers. It should be noted that both methods detect outliers in areas which the frequency criterion also produces. But the areas are not corresponding well in either method. This issue is apparent in sample HF-2802 of LGm4, shown

in Figure 3.9.



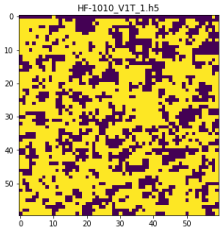
(a) The result of the interquartile range for HF-2802, some larger areas are detected with sporadic spots around the entire spectra.



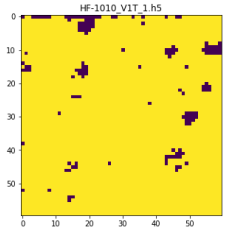
(b) The result of the standard deviation test for HF-2802. Few outliers are detected

Figure 3.9: A comparison between the interquartile range and the standard deviation test for sample HF-2802. In comparison, the interquartile range locates better defined areas than the standard deviation test, but many sporadic spots are present.

In Figure 3.9, the results of IRM and SDT are compared on sample HF-2802. Both methods yield poor separation between the outliers and tumor spectra. IRM is capable of detecting a large mass of some material on the sample, however, many sporadic spots appear around it, the majority of these spots are tumor spectra which would be discarded by the method. In stark contrast, SDT struggles to detect anything in the sample, only yielding small spots in the larger areas found with IRM. Fewer samples suffer from the stochastic results present in SDT, however all samples are not strictly improvements from SDT. One such example is sample HF-1010 belonging to LGm2, shown in Figure 3.10.



(a) The result of the interquartile range for HF-1010



(b) The result of the standard deviation test for HF-1010.

Figure 3.10: A comparison between the interquartile range and the standard deviation test for sample HF-1010, LGm2.

The comparison in Figure 3.10 show one of few samples where IRM fails to sufficiently separate outliers from tumor spectra. A considerable number of spectra

are falsely labeled as outliers, and those spectra seem to form many smaller areas without sporadic points in the surrounding area. In this rare case, SDT exhibits sufficient separation of the outliers, as the detected areas correspond well with the the position of known outliers found by the frequency criterion.

Few samples follow the same conclusion however, these varying results show clear signs that neither of the methods are suitable to perfectly rid each sample of outliers. We therefore dismiss them from the analysis, while keeping the results for comparison with the other methods and continue the analysis by analyzing unsupervised machine learning methods for outlier detection.

3.3.4 Hierarchical clustering

The next method for outlier detection is hierarchical clustering, we choose to utilize agglomerative clustering as an arbitrary choice. Due to the algorithms demanding time complexity and memory constraints, we choose to analyze each sample separately to avoid memory errors. Each sample must be given one unique model which then divides the data into the number of clusters specified. This means there will not be a universal model designed to separate all samples. We compare the results of different distance metrics and choose to utilize Euclidean distance to measure distance among the clusters. This conclusion is due to the uncertainty in vector shape and angles on which Cosine similarity is dependent. The Manhattan distance would be a better choice compared to Cosine similarity. However, Euclidean distance magnifies long distances which should aid the algorithm in selecting clusters for agglomerative merging. Following the choice of distance metric we examine the linkage criteria available. In each case, the algorithm is set to run multiple tests where it separates the data into different amounts of clusters. We choose to initialize five different models to compute between 2 and 7 clusters respectively. Due to the deterministic nature of the algorithm, the results in one cluster will be present in the other clusters as the the number of clusters increases.

Single linkage merges clusters by way of merging those clusters which posses points with minimal distance. Should the spectra within the samples be significantly "far apart" in the data, the criterion might start producing many unique clusters in a "chain". If the number of clusters are sufficiently small, the number of cluster separations will be few, and few spectra will be allocated to those clusters. This can result in few separations for some samples, which will render the method unusable for our purposes.

The criterion yields clusters which appear as individual points in the spectra and fail to detect areas where known outliers are present in the samples, suggesting the

aforementioned flaw is present in this methodology. This is apparent as we see few clsuter areas form in any of the initialized models, even if we allow the method to use more than two clusters. All different cluster models fail to separate the outliers and instead separate a small number of points. An example of this phenomenon is displayed in Figure C.2.

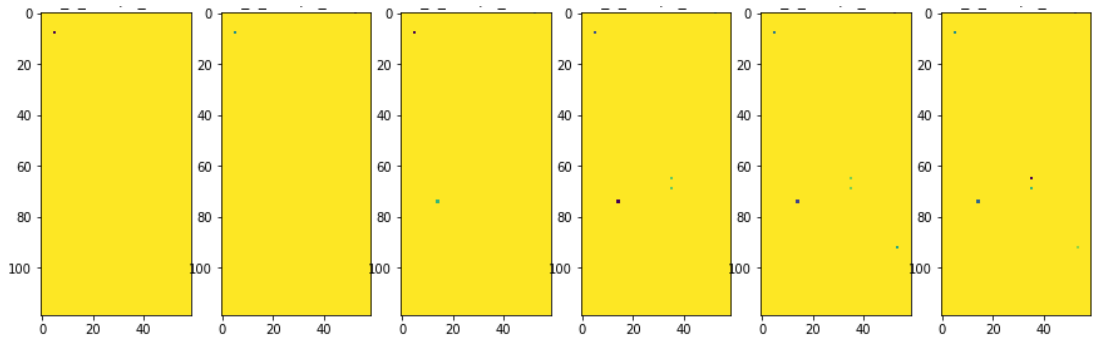


Figure 3.11: Single linkage on sample HF-868 from LGm1, The test fails to detect any outlier areas. Leftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

The models fail to detect any areas where outliers are present, only yielding an insignificant number of outliers in the entire sample. Models computing more clusters also fail to find significant areas and smaller spots appear as the number of clusters increase. As the number of clusters increase, several of the outliers seem to belong to their own clusters, which is a sign of the clusters being computed in a "chain" as previously stated. Due to this unsatisfactory result, Single linkage will not be used as linkage criterion in this project.

Average linkage yield superior results compared to single linkage since some areas become apparent as we increase the number of clusters. Moreover, the criterion does not have a set number of clusters which is guaranteed to include all outliers for all samples. For certain samples the outliers are visible when forcing the algorithm to agglomerate to two clusters and others only show them once five or more clusters are allowed. It is possible to use the criterion for discarding the outliers if the majority cluster is preserved when computing 7 clusters, while the rest of the spectra are discarded. However, this would not remove all outliers and some problematic samples in LGm3 would have the majority of outliers preserved. The improvement from Single linkage is made apparent in Figure 3.12

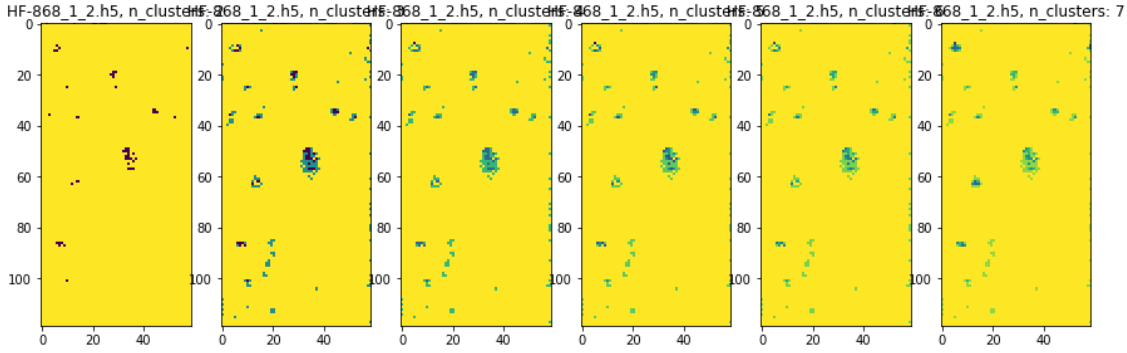


Figure 3.12: Average linkage on sample HF-868 from LGm1, Leftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

Complete linkage merges the clusters which posses elements with the smallest possible maximal distance between them. This avoids the setback of single linkage as a majority cluster is harder to form early under the criterion. The method produces similar results as average linkage, few areas with outliers are detected when fewer clusters are permitted. However, some outliers are present when computing 2 or 3 clusters. Allowing 7 clusters to form allows the algorithm to capture many of the areas of interest, however, this captures too much in some samples. We recommend the majority cluster is maintained when allowing 4 clusters with the method. But this suffers from the same setback as our conclusion on average linkage.

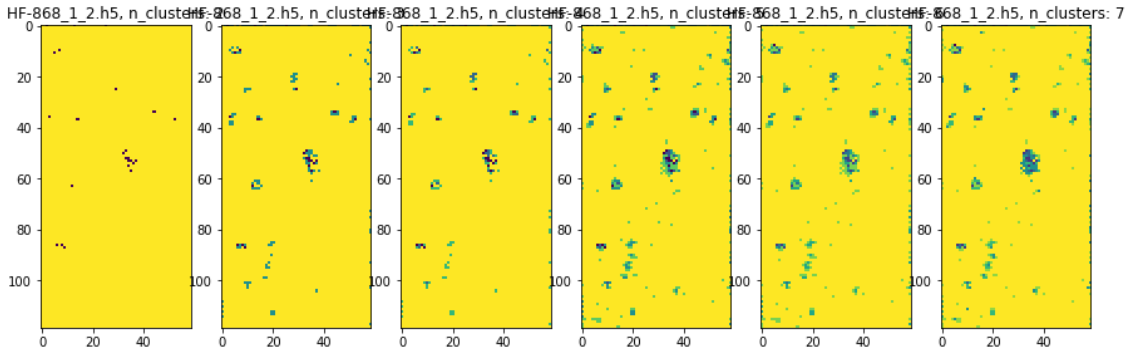


Figure 3.13: Complete linkage on sample HF-868 from LGm1. Leftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

Ward linkage merges clusters which posses minimal variance between their respective elements and as such works well with Euclidean distance. We do stress that the clusters are merged by measuring variance among cluster elements and there is no guarantee the outlier spectra should share in characteristics which would result

in low inter-cluster variance. Despite this lack of guarantee the clusters form at the precise location of outliers. Allowing 7 clusters produce a near picture perfect image of the biological tissue from which the spectra were measured. These results show that the algorithm is capable of organizing the spectra according to their visual information which aid us in understanding shape and state of the samples. Some of the spectral images fromed by the clusters are shown in **Appendix**.

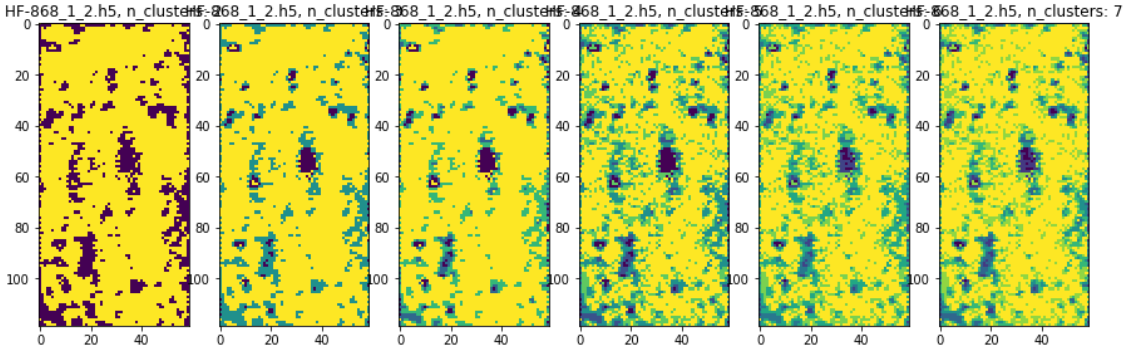


Figure 3.14: Ward linkage on sample HF-868 from LGm1. Leftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

The issue is finding a uniform criterion on which we can discard the outliers. Removing every cluster except for the majority cluster in the case where 7 clusters have been formed would remove legitimate spectra which are suitable for training a model. In fact, there is no optimal choice in this case, as certain samples have their outliers sufficiently captured in a setup allowing for 2 clusters. While others show their outliers in arbitrary numbers of clusters. Selecting too many clusters will result in some samples loosing legitimate spectra which is undesirable in context of maintaining a sizable dataset. By visual inspection, we deem the optimal choice to be 3 clusters since many outliers are present in this choice, though the problem is still present in this alternative. In particular, sample HF-2485 suffers from this problem, the cluster results are shown in Figure 3.14.

The model separates much of the healthy tissue in models computing 3 clusters or fewer. In the model computing 4 clusters, the utliers are visible as dark spots which perfectly correlates with the image of the tissue. The surface of the tissue is also well represented in the model computing 7 clusters.

3.3.5 K-means clustering

We continue this analysis by performing K-means clustering in attempt to separate the outliers. We flatten the training data and reorganize it in random order to avoid

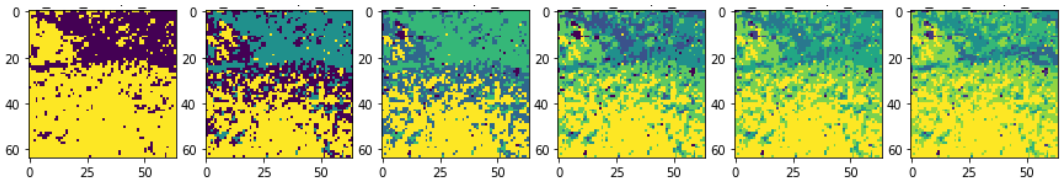


Figure 3.15: Sample HF-2485 from LGm5, clustered by Hierarchical clustering using Ward linkage. From left to right, is shown the results of computing from 2 to 7 clusters respectively. The oultliers are present from image 4 and up.

bias towards recurrent LGm-classes in the dataset. The examples are then used to fit 5 K-means models to compute 2 to 7 clusters respectively. This method has the advantage of being trained on the entire training set whereas Hierarchical clustering possesses too great a time and space complexity which makes similar experiments impossible with our current hardware. Under this structure we may now compare the results between sample predictions. We observe that the stochastic nature of the algorithm produce results of varying quality. In contrast to Hierarchical clustering, K-means do not produce clusters as subdivisions of previously seen clusters. This is due to the algorithm being computed several times with different initialization settings. The final result which the algorithm yields is the cluster state possessing minimal inertia compared to the other computations. Compared to all other methods of analysis, this method is able to capture small segments of the upper part of the sample HF-1293, shown in Appendix **??**. While the comparison among the different models created loses some credibility in this setting, the comparison between the model results among the different samples are promising. We find that the model in which two clusters are computed, the samples are divided in ways which corresponds with Adrians criterion (explained in the next section) whereas samples where minimal amounts outliers are present the clusters seem to form around healthy tissue, which in turn create an image resembling the surface of the sample-tissue. In certain cases the clusters fail to capture known outliers but other models allowing for more clusters capture them sufficiently well. One example of this is sample HF-868 where the two cluster model fails to detect the relevant outliers but the model computing 3 clusters capture the outliers in near perfect resemblance to Adrians criterion. The other models however then loose some of the outliers around the shapes present while still capturing the center of the shapes. The likeness to Adrians criterion returns in the model computing 6 clusters but this is again lost in the model computing 7 clusters. And like Ward linkage for Hierarchical clustering, there are some samples where the majority cluster is hard to evaluate, one such sample is HF-2544 which shifts the majority cluster between the model computing

2 clusters and the rest of the models. Due to the variety among the different models finding a uniform criterion for detecting outliers is problematic. Many samples are such that the models steadily increase the number of outliers clustered, but this relationship is not constant through all samples, making it insufficient for use in the hypothetical criterion. The shifting of majority cluster in the aforementioned sample further complicates matters, since the majority cluster may not be capturing healthy tissue in some samples. For this reason we deem the method insufficient for separating outliers though we add we note its promise in capturing information about the tissues visual aspects, which makes the method comparable to Ward linkage for Hierarchical clustering.

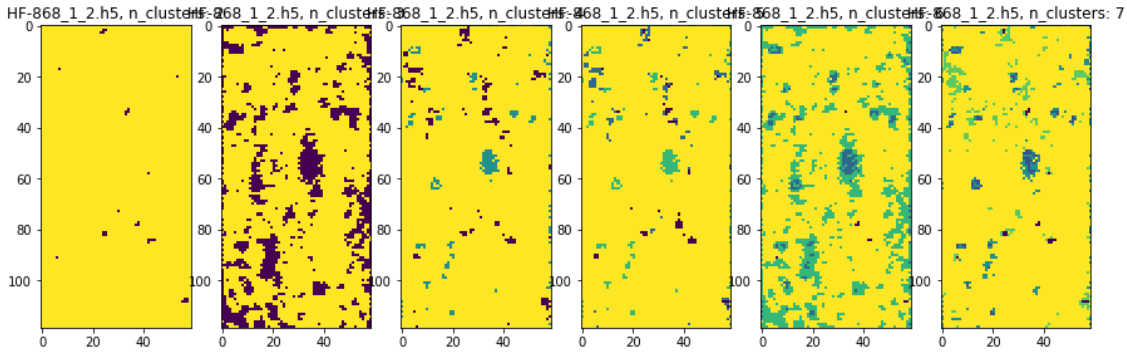


Figure 3.16: Single linkage on sample HF-868 from LGm1, The test fails to detect any outlier areas. Leftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

Concluding this section we also perform clustering with the selected features gained from the beginning of this section **Reffer to appendix for this!**. Limiting the spectra to the features which best separate them into the 6 LGm-classes has advantage for this algorithm, many areas of outliers become easily spotted by allowing 2 clusters in many cases. Complete linkage gains the biggest advantage out of this method. None of the above mentioned configurations manage to capture the upper part of HF-1293, however, when the data is limited to the features selected, the area becomes visible to some extent. This suggests there are frequencies in the spectra which complicate the separation of outliers. While we see that the method clearly works in many cases, using all features proves to be troublesome for the majority of linkage criteria. Furthermore, the features are computed by means of separating the spectra into the 6 LGm-classes, they are not computed due to their efficiency for detecting legitimate spectra and outlier spectra. Following the removal of outlier spectra we decide to label each spectrum by a binary value, a spectrum is assigned true if it is legitimate and false if it is labeled an outlier by Adrians cri-

terion. We then run feature selection once more, extracting the features which bets separate legitimate spectra from outliers. The features are displayed in Appendix **appendix:Features**.

Chapter 4

Results

Chapter 5

Conclusion

Bibliography

[1] C. S. von Bartheld, J. Bahney, and S. Herculano-Houzel, "The search for true numbers of neurons and glial cells in the human brain: a review of 150 years of cell counting," *Journal of Comparative Neurology*, vol. 524, no. 18, pp. 3865–3895, 2016.

[2] S. Jäkel and L. Dimou, "Glial cells and their function in the adult brain: a journey through the history of their ablation," *Frontiers in cellular neuroscience*, vol. 11, p. 24, 2017.

[3] O. Gallego, "Nonsurgical treatment of recurrent glioblastoma," *Current oncology*, vol. 22, no. 4, p. e273, 2015.

[4] F. E. Bleeker, R. J. Molenaar, and S. Leenstra, "Recent advances in the molecular understanding of glioblastoma," *Journal of neuro-oncology*, vol. 108, no. 1, pp. 11–27, 2012.

[5] S. L. Maas, E. R. Abels, L. L. Van De Haar, X. Zhang, L. Morsett, S. Sil, J. Guedes, P. Sen, S. Prabhakar, S. E. Hickman, *et al.*, "Glioblastoma hijacks microglial gene expression to support tumor growth," *Journal of neuroinflammation*, vol. 17, pp. 1–18, 2020.

[6] A. Dirkse, A. Golebiewska, T. Buder, P. V. Nazarov, A. Muller, S. Poovathingal, N. H. Brons, S. Leite, N. Sauvageot, D. Sarkisjan, *et al.*, "Stem cell-associated heterogeneity in glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment," *Nature communications*, vol. 10, no. 1, pp. 1–16, 2019.

[7] K. Vigneswaran, S. Neill, and C. G. Hadjipanayis, "Beyond the world health organization grading of infiltrating gliomas: advances in the molecular genetics of glioma classification," *Annals of translational medicine*, vol. 3, no. 7, 2015.

[8] Y. Hirose, H. Sasaki, M. Abe, N. Hattori, K. Adachi, Y. Nishiyama, S. Nagahisa, T. Hayashi, M. Hasegawa, and K. Yoshida, "Subgrouping of gliomas

on the basis of genetic profiles," *Brain tumor pathology*, vol. 30, no. 4, pp. 203–208, 2013.

[9] M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, *et al.*, "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma," *Cell*, vol. 164, no. 3, pp. 550–563, 2016.

[10] L. Dang, K. Yen, and E. Attar, "Idh mutations in cancer and progress toward development of targeted therapeutics," *Annals of Oncology*, vol. 27, no. 4, pp. 599–608, 2016.

[11] D. A. Long, "Raman spectroscopy," *New York*, pp. 1–12, 1977.

[12] P. Graves and D. Gardiner, "Practical raman spectroscopy," *Springer*, 1989.

[13] M. K. Maruthamuthu, A. H. Raffiee, D. M. De Oliveira, A. M. Ardekani, and M. S. Verma, "Raman spectra-based deep learning: A tool to identify microbial contamination," *MicrobiologyOpen*, vol. 9, no. 11, p. e1122, 2020.

[14] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[15] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, "Deep convolutional neural networks for raman spectrum recognition: a unified solution," *Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017.

[16] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[17] S. Chawla and A. Gionis, "k-means–: A unified approach to clustering and outlier detection," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 189–197, SIAM, 2013.

[18] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is np-hard," in *International Workshop on Algorithms and Computation*, pp. 274–285, Springer, 2009.

[19] M. K. Pakhira, "A linear time-complexity k-means algorithm using cluster shifting," in *2014 International Conference on Computational Intelligence and Communication Networks*, pp. 1047–1051, IEEE, 2014.

[20] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The computer journal*, vol. 26, no. 4, pp. 354–359, 1983.

[21] 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*, vol. 12, pp. 2825–2830, 2011.

[22] C. Shalizi, "Distances between clustering, hierarchical clustering," *Lectures notes*, 2009.

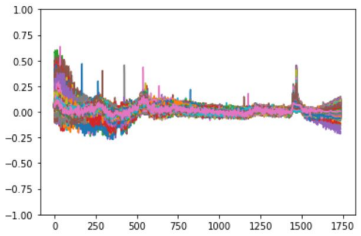
[23] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[24] D. Friedmann-Morvinski, "Glioblastoma heterogeneity and cancer cell plasticity," *Critical Reviews™ in Oncogenesis*, vol. 19, no. 5, 2014.

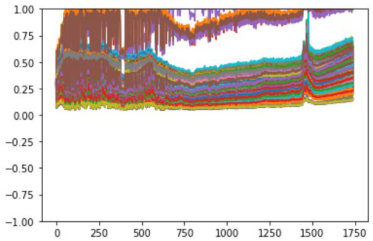
Appendices

Appendix A

Spectral Plots



(a) Spectra from patient HF-1293. The rest of the samples available share in this pattern with some deviations



(b) Spectra from patient HF-1887. The frequencies tilt towards the upper part of the plot. The example is decidedly removed from the analysis.

Figure A.1: Examples of samples drawn from the data, HF-1293 display a common pattern across all samples, HF-1887 is removed due to problematic handling

Appendix B

Feature Selection

509, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 532, 533, 538, 539, 540, 541, 545, 546, 547, 548, 549, 550, 551, 552, 553, 562, 563, 647, 1449, 1450, 1451, 1452, 1453, 1454, 1455, 1456, 1457, 1458, 1459, 1460, 1461, 1462, 1463, 1464, 1465, 1468, 1469, 1470, 1471, 1472, 1473, 1474, 1475, 1476, 1477, 1478, 1479, 1480, 1481, 1482, 1483, 1484, 1485, 1487, 1492, 1494, 1495, 1496, 1497

Appendix C

Spectral Images For Outlier Detection

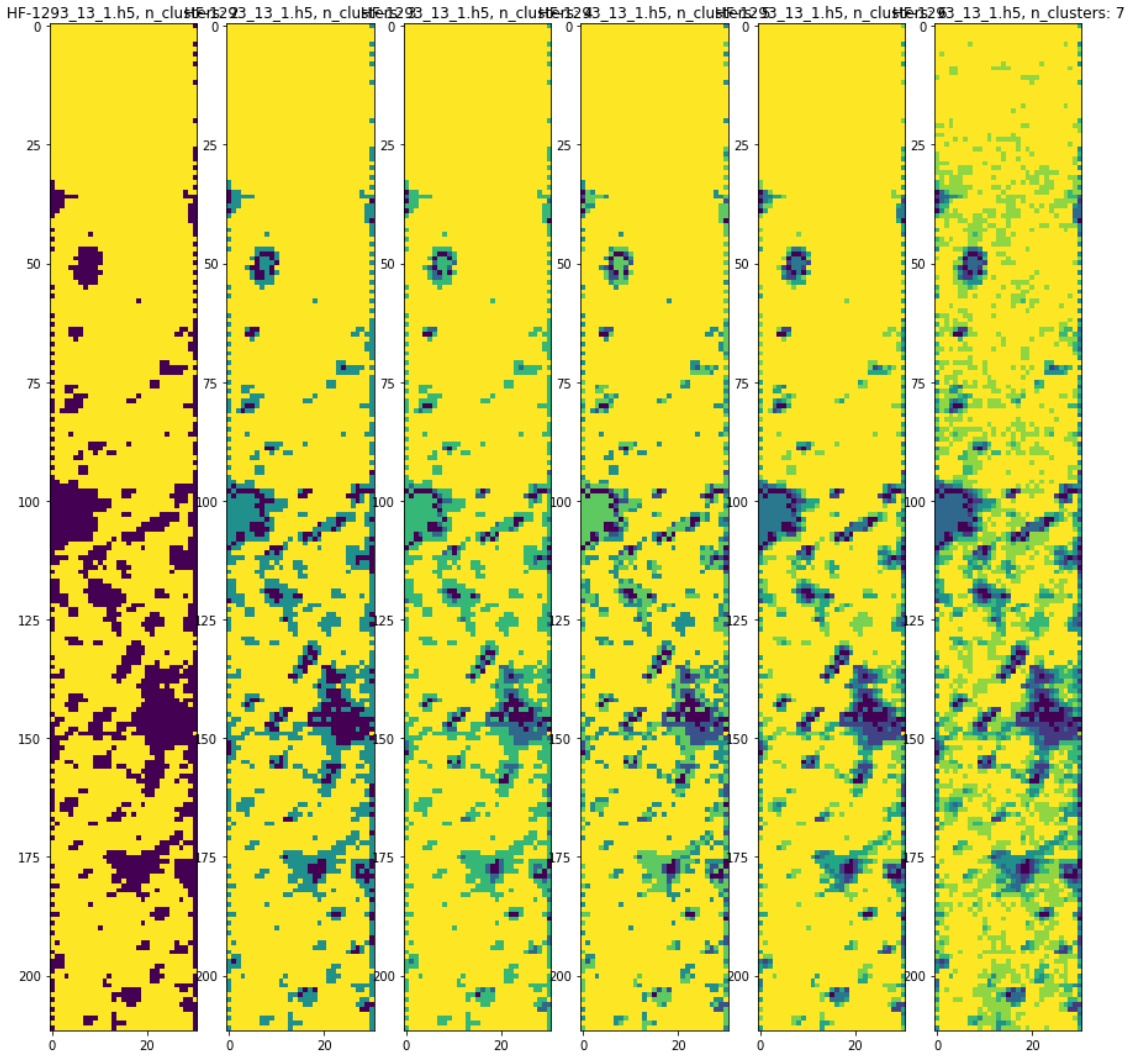


Figure C.1: Single linkage on sample HF-868 from LGm1, The test fails to detect any outlier areas. LEftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.

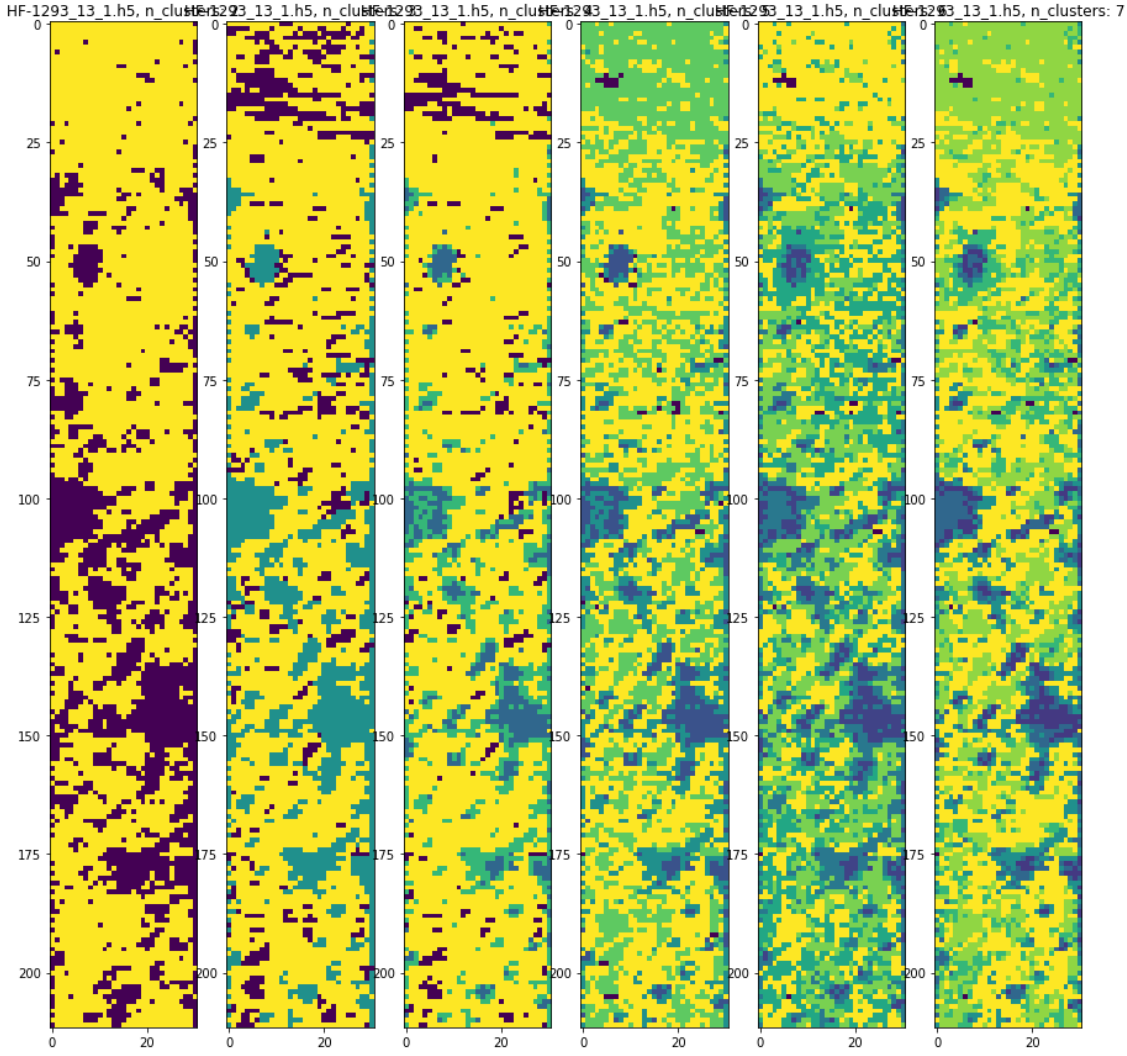


Figure C.2: Single linkage on sample HF-868 from LGm1, The test fails to detect any outlier areas. LEftmost image is the result of the model computing 2 clusters. The number of clusters increase towards the rightmost image.