Methylation based classification of raman spectra extracted from glioma cells using deep learning (but with a shorter title)

author: Joel Sjöberg 38686
Masters thesis in Computer Science
Supervisor: Luigia Petre
The Faculty Of Science And Engineering
Åbo Akademi University
2020

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 consists primarily of glial cells, 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 suggests the cells importance to the nervous system is greater than once thought, though their actual function is still a matter of speculation [1]. Glioma is a type of brain cancer which manifests within the brains 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]. Gliomas are categorized depending on their glial-cell of origin. There are four types of glial cells (also called neurolia or simply glia) Oligodendrocytes, astocytes, ependymal cells and microglia. Oligendroglioma originates from oligodendrocytes, astocytoma from astocytes and ependymomas originate from ependymal cells. Furthermore, the aforementioned astocytoma-types may develop into glioblastoma multiforme (GBM), the most aggressive form of brain cancer which is may communicate with microglia to increase tumor growth [4]. However, It is also possible for GBM to develop from other brain cells. This cancer is particularly aggressive due to its quick reappearance in the brain only a short period after surgery [2]. The heterogeneity of GBM-cells further complicates the healing process, due to poor response to targeted treatments [5].

At present, the World Health Organization (WHO) has defined four levels of cancer severity used to describe their aggressiveness and tumor growth (these levels are also called grades). These are low-grade (grade I and II) and high-grade (grade III and IV) glioblastoma is cathegorized by grade IV [3, 6]; and these must be examined to determine an appropriate prognosis and line of treatment. However a study by Vigneswaran et al. [6] 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. [7]. Ceccarelli et al. [8]

introduce alternative subdivisions of these classes which show promise in expanding knowledge about glioma tumors and aid in treatment selection. Such evaluation 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 6 distinct classes, labeled LGm1-6. Their analysis showed IDH-mutations in LGm1-3; furthermore LGm4-6 were IDH-wild-type where a majority of tumors could be labeled as glioblastoma. 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 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 producing tumor-type prediction in a matter of minutes. The project relies on tissue from tumors extracted from 53 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 [9, 10]. 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 [11, 12]. The use of Raman spectra is further motivated by Liu et al. [13], whose work show promise for deep learning models trained on raw Raman spectra. The advantage of this method in context of multilabel classification is considerable, when compared to other machine learning methods such as Support Vector Machines, Random Forest and K-nearest neighbor [13].

The thesis is structured as follows, Chapter 2 presents the preliminary theoretical background for machine learning, along with the necessary mathematical definitions by which these methods are defined. Understanding the underlying definitions is necessary to validate and confirm the results. 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 and proceeds to display the methods in the order they were utilized. The use of unsupervised learning and feature selection are also explained and motivated. In Chapter 4, we present the deep learning model, the performance it yields. In Chapter 5 concludes this thesis by giving suggestions to further study along with

arguments for and against the use of machine learning in the medical field. The thesis is concluded in Chapter 6 with a discussion of the impact the deep learning model will have on the medical field. The primary focus of this thesis will be placed on Chapter 3 and 4 as they cover the purpose of the project.

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 [14]. 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 [15]. 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 [15, 16] but the algorithm itself has a time complexity of $O(n^2)$ [17].

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 [18].

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 [19]: 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) [20].

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

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 [21]. 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, certain tumors may be heterogeneous [22], 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 the deep learning model, rendering its desired predictive capabilities uncertain. We achieve this by duplicating underrepresented samples-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 k-means clustering and hierarchical clustering for detecting non-tumor spectra or otherwise erroneous spectra. These methods are tested in contrast to other outlier detection techniques such as the standard deviation test and the interquartile range method. Adrian's criterion is presented and compared to the the spectral images extracted from the samples. 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. For this feature selection is used, representing each frequency within the spectra

as distinct features. This task is simplified by the new representation of the data separated into lists of spectra rather than a collection of spectra represented by the tumor. However in order to extract such features 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 faulty spectra e.g. spectra of blood drops on the sample or plastic which may be reflected form underneath thin tissue. Furthermore some tissue may be necrotic which will affect the spectral signal. Using the extracted features a model can potentially form around the data faster which can be essential as deep learning training require significant computing resources. The features are extracted before and after the removal of problematic spectra for comparison.

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, since there is uncertainty regarding the homogeneity among patient samples. The data will be separated into three separate datasets. These sets are called training set, validation set and testing set. All datasets will consist of unique patients to avoid scenarios in which the model overfits to a patients tumor sample and as a result of heterogeneity. This structure also allows for easier handling of the number of patients in the sample-classes, allowing for analysis on each class exclusively.

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 instrument 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) 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 which is performed by the provider, each element a unique frequency. 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 predict which subdivision the spectra belong to by feeding in one of these samples, i.e., one vector of shape $(1, 1738)$. This strategy is inspired by Liu et al.[13], who managed to get satisfactory performance by training a model on raw spectra. This representation is of great interest, since the value of each spectrum is independent from the surrounding spectra. Separation at this level yields a dataset with more than $300,000$ datapoints, which better suits deep learning tasks.

To prepare for the project each samples spectra is collected and plotted in one single plot to compare spectral information from the sample itself. An example of such plots can be seen in Appendix **??**. Patient HF-1887 is removed from the project completely due to the skewed baseline in the spectra.

3.2 Data preparation

In this section we explain our qualitative analysis on the data, done to determine the plausibility of our model. Each sample is categorized according to their subdivision; there are six distinct subdivisions as defined by Ceccarelli et al. [8], denoted LGm1 - 6. As an initial step each samples spectra is analyzed by visual inspection. The spectra are plotted on a two dimensional surface as lines, each line a unique spectrum. Through this analysis one sample is discarded due to the tilted baseline of the spectra, none of the other samples share this problem. A sample which shares a general shape with the other samples is shown in contrast to the sample selected for removal in **??**. Another sample shows a concerning number of spikes in contrast to the other samples. The number of spectra is also considerably larger compared to the others, which is limiting for some of the methods selected for the analysis. For this reason the sample is also discarded.

3.2.1 Organization and Balance

The model will as a consequence of it's learning-algorithm become biased towards certain predictions. This because 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. The initial data suffers heavily from this problem, a 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 is develops bias towards the patients in training and if the patient samples are heterogeneous with respect to the other samples of the same class. The test-samples are chosen manually, 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. Following this method the majority class will stay the majority which can be crucial provided the sample pattern is similar to the set of all other unseen samples. The resulting dataset is gained by doubling the samples in LGm1, tripling the samples in LGm3 and quadroupling the samples in LGm6. The distributions of the training dataset is shown in Table 3.2.

Table 3.3 shows the distribution of the training data and testing-data. The training data is then balanced exclusively. This is not required in the testing data, since

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# train	21289	51698	20635	34276	37492	12976
# test	14945	20140	11296	16384	24764	7200

Table 3.2: Distribution of the training data and the testing data

it will have no effect on how the model is developed through training. The training data is balanced by replicating each spectra in every patient of the classes which are under-represented. The resulting training-set is gained by doubling the spectra in LGm 1 and LGM 3 and the spectra in LGm 6. The final distribution of the training data following this procedure is shown in table 3.3

Class	LGm1	LGm2	LGm3	LGm4	LGm5	LGm6
# train	42578	51698	41270	34276	37492	38928

Table 3.3: Distribution of the testing data following balancing

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 [19] which is given the f-classif method for ranking the features which yields features deemed significant by ANOVA. The 70 best features were extracted from the training-data in which there are spectra which originate from non-tumor tissue. The features are displayed in Appendix B. Before the features are extracted, each spectra is standardized using z-score standardization, to give each spectrum a mean of zero and a standard deviation of one. This is done for ease of comparison among the spectra. The extracted features show that regions of interest do exist on the spectra. This can be seen by the integers which have a 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 non-tumor spectra 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.

To avoid bias the analysis is performed on the training data exclusively. 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 standard deviation test and interquartile range method, these are deterministic methods that rely solely on the values found within the data. K-means clustering and agglomerative clustering are then performed on the data in attempt to capture potentially complex patterns within the data. The section ends by presenting Adrians criterion, which is a criterion for finding problematic spectra defined by the data provider.

3.3.1 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 3 standard deviations or more. We measure the mean and standard deviation on each frequency from the unbalanced training-set; 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 3 or more standard deviations from the mean. In any given sample, each spectrum includes on average 111 frequencies which are separated by 3 standard deviations or more from the mean. We specify that a spectrum whose number of frequencies which fail the test is a spectrum possessing more than 111 outlier-freuencies, is deemed an outlier. A display of which spectra would be removed from the samples of type LGm1 is shown in Appendix **??**. Using this test, many samples outline small spots on 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 include them or not. Some samples do show clear areas of outliers but areas within those regions are falsely classified as non-outliers. There is also one sample for which the test classifies approximately 97% as outliers which is not confirmed by the provider to be problematic. Despite promising results for some samples, we decide to turn to other methods for outlier detection.

3.3.2 The Interquartile range method

Similar to the standard deviation test the interquartile range method is a purely statistical analysis method which detect outliers in terms of which percentile the points fall into. The 25th and 75th percentiles are calculated on each frequency for

the entire training set. Unlike the previous test, 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. The resulting outliers are shown in Appendix **??**. Compared to the standard deviation test, this method produces similar results. However, many regions are better represented by the interquartile range, showing well defined areas where unknown material is 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.

3.3.3 Hierarchical clustering

The next method of analysis for outliers 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 these issues. This means there is little to gain in comparison among separate samples however the deterministic nature of the algorithm shows promise for rigorous results in the clusters. 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 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 compute between 2 and 7 clusters.

Single linkage merges clusters by way of merging those which posses points with minimal distance. One flaw in this criterion is that clusters may easily be merged from one parent cluster until all points have been connected in a chain. 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 especially apparent if we allow the method to use more than two clusters. All different cluster amounts fail to separate the outliers and instead separate a small number of points. Examples of these cluster results are displayed in **Appendix**.

Average linkage yield similar results as single linkage with the exception that 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 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, but this would not remove all outliers and some problematic samples in LGm3 would have the majority of outliers preserved.

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.

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**. 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 choise to be 3 clusters since many outliers are present in this choice, but the problem is still present in this choice.

3.3.4 K-means clustering

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

3.3.5 Adrians criterion

Adrians criterion is a criterion specified by Adrian Lita for separating outlier spectra from the rest. It states that should any value between frequency 1463 and 1473 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 consistency in the other methods explained in this section, we choose to discard outliers found by adrians criterion despite the criterion missing certain areas such as the upper part of sample HF-1293 which does satisfy the criterion, but is also confirmed to be plastic. Our hope is that the model will be able to ignore these and other potential outliers which escape Adrians criterion.

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 criterion. 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] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

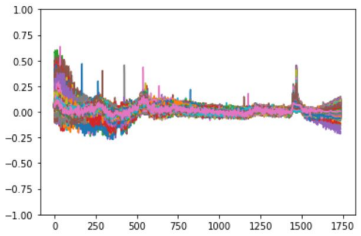
[21] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

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

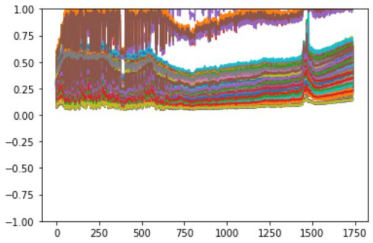
Appendices

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

begin 250 309 522 523 524 525 526 527 528 529 553 563 645 646 647 648 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500 1501 1502 1503 1511 1512 1513 end

Appendix C

Spectral Images For Outlier Detection