# 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

Through what feels like uncountable hours in a hunchback-like state which could only be provoked by the most wholesome of encouragement I managed to produce this thesis(or, the thesis thus far). Remember Hofstadters law: "It always takes longer than you expect, escpecially when you take into account Hofstadters law". Thank you to fellow: **Fellow0, Fellow1, Fellow2, Fellow3, ..., FellowN** where N -> ∞

# Abstract

But not abstract enough.

# Chapter 1

# Introduction

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[1][2]. Gliomas are categorized depending on their glial-cell of origin. 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. However, It is also possible for GBM to develop from other brain cells. This cancer is particularly agressive due to quick reappearance in the brain only a short period after surgery[1]. The heterogeneity of GBM-cells further complicates the healing process by avoiding certain targeted treatments[3].

At present there are four grade-types independent from the tumor type defined by the World Health Organization(WHO) used to describe their aggressiveness and growth. These are low-grade (grade I and II) and high-grade(grade III and IV) where glioblastoma is cathegorized by grade IV [2][4]; these must be examined to determine an appropriate prognosis and line of treatment. However a study by Vigneswaran et al. suggested these grades could be divided further to better describe the features of the tumor e.g. GBM could be divided further to express versions with poor prognosis. Ceccarelli et al.[5] 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 reinforces by the results produced by Vigneswaran et al. The process of de-

termining prognosis and line of treatment have great promise in improving patient outcome by classifying tumors by these subdivisions.

This thesis explains a project whose purpose is to optimize the categorization process by introducing a deep learning model capable of producing tumor-type prediction in a matter of minutes. Tissue from tumors of 53 patients is extracted 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 change their vibration. This vibration is gathered by the instrument and may then be used as data to determine properties of the material[6][7]. This spectra is the data which the model will use as training and testing data. The choice to utilize Raman spectra in this way is due to the methods success in previous studies presenting Raman spectra applied to machine learning algorithms[8][9]. The use of Raman spectra is further motivated by Liu et al.[10] whose work show promise for deep learning models trained on raw Raman spectra. The advantage the method possess in context of multilabel classification compared to other machine learning methods is shown to be considerable in contrast to Support vector machines, Random forest and K nearest neighbors[10]. In chapter two the preliminary theoretical background for machine learning is covered along with the necessary mathematical definitions by which these methods are defined. Understanding the underlying definitions is necessary to validate and confirm the results with expected and public results in the field. Chapter three explains 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 is motivated. The chapter concludes by presenting the results of the exploration-stage. Chapter four presents the deep learning model alongside the performance it yields and the expected impact it will have on the medical field. The primary focus of this thesis will be placed on chapter three and four as they cover the purpose of the project. Chapter five concludes this thesis by giving suggestions to further study along with arguments for and against the use of machine learning in the medical field.

# Chapter 2

# Theoretical Background

Within this chapter can be found the essential mathematical theory on which this thesis is based. The concepts are considered fundamental for understanding the methods applied in our 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

Machine learning is the practice of computing models for relationships between sets of data. The field has garnered significant interest within akademia and industry alike due to the promising result in a number of applications. Within the field there are mainly 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 collection of vector transformations 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 a fundamental version of such a clustering algorithm called *K-means clustering*. The following is the formal definition of *K-means clustering* as defined by MacQueen[11]. Given a set of $N$-dimensional points $E_N$ and a desired amount of partitions $k$ in said population, partition the elements of $E_N$ into a partitioned set $S = \{S_1, S_2, ...S_k\}$. The partitioning of $E_N$ is performed by 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.

$$T_i = \{x : x \in E_N | (\forall_{j \in [0,k]/i} |x - v_i| \leq |x - v_j|)\} \tag{2.1}$$

For centers who share equal distance to any given $x$ the cluster with the smallest index is chosen as the containing set. The partitions $S_i \in S$ are defined by formula 2.2

$$S_i = T_i \cap \bigcap_{j=0}^{(i-1)} S_j^c \tag{2.2}$$

A consequence to this definition is that outliers have a potential to drastically change the quality of the cluster outcomes[12]. 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[12][13] but the algorithm itself has a time complexity of $O(n^2)$[14].

### 2.2.2 Hierarchical clustering

Hierarchical clustering is a clustering method which is less suceptible to outliers compared to K-means. The method produces clusters by iteratively combining the closest clusters according to the linkage criterion. 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 process in counter to the agglomerative strategies by initializing one universal cluster for all data points and then 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[15].

#### 2.2.2.1 Distance metrics

Let $u$ and $v$ be vectors of the same dimension $n$. The euclidean distance(alternatively L2) measure can be used to measure distance between the vectors in euclidean space. It is calculated by taking the square root of the summed squared distance between each vector element in each vector. Formally expressed in equation 2.3.

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

The manhattan distance(alternatively L1) 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 by summing the absolute distance of every element between every vector. Formally 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 by dividing the sum of each element product in each vector by the product of the square roots of each summed and squared vector element. Formally 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[16]: 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 cluster 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[17].

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

### 2.2.3 Feature Selection

In many cases the data available contains numerous features. This is necessary, as gathering enough features is essential to building a sufficient classifier. However as the number of features increases so does the requirement for training data. To avoid this expanding dependence on data, it is necessary to strip the data of certain features which possess minimal description of the data in question or which lack that description entirely [18]. These features which possess the expressive information are not always trivial, which further motivates the use for machine learning to find these features.

## 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 ran-

domly 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

Before a deep learning model is created it is necessary to have extensive knowledge about the data in the project. As certain tumors may be heterogeneous[19], their biological information may make them hard to group which would make model generalization difficult. The analysis explained here is to determine whether Raman spectrum can sufficiently sort the different samples into the sample classes (LGm1 - 6). In this chapter the data available to the project is examined in greater detail; details for how the Raman spectrum were prepared is given to document the preprocessing of spectrum for future use. The chapter begins by describing the mathematical representation of the samples. The number of samples is deemed too small for use in a deep learning model, an explanation is given for how each sample may be separated into individual spectras after which an argument is made in favor of this representation. This separation yields a drastic increase in the number of training examples available which reinforces the plausibility of deep learning models in this context. The chapter then proceeds to explain how the data is to be balanced. An unbalanced dataset would likely invoke bias into the deep learning model which would introduce uncertainty in it's desired predictive capabilities. This is done by duplicating samples in sample-classes which are underrepresented in the dataset. Furthermore, this balancing is performed to maintain majority and minority classes which retains some distributional information from the original dataset. Certain problematic samples are also discarded in this section to alleviate memory issues which would otherwise be present on a modern computer. The chapter continues with feature selection which is also performed to extract the features best descriptive of the different classes with respect to the other classes. This step further aids in alleviating memory issues and run time, the necessity is further motivated by comparison of the results from the extraction. The chapter ends with an analysis of each sample using hierarchical clustering with complete linkage in an attempt to detect and remove eventual outliers within the samples themselves.

## 3.1 Data Representation

The data consists of the Raman-spectrum extracted from the tissue of glioma tumors from 45 patients. Multiple samples of tissue was 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 the patient to whom they belong. This is necessary at the exploration stage since the data at one point will be divided into training and testing examples. The testing data will consist of unique patients to avoid scenarios in which the model is familiar with a patients tumor sample. This structure also allow for ease in handling the amount of patients in the sample-classes. 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, where the third dimension(denoted above by 1738) could serve as a color value. The number 1738 is constant through all samples and represent the length of the Raman-spectra, each element a unique frequency. Furthermore each element inside these arrays is a real number without clear bounds. The largest absolute element found within 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 is motivated by Liu et al.[10] who managed to get satisfactory performance by training a model on raw spectra. To further alleviate memory issues, feature extraction is performed to reduce the sice of the spectra to $(1, 70)$ by extracting the 70 features most capable of characterizing the data. This process is explained in detail in a later section. The provider further states that the spectra may be separated to individual spectrum as their alignment do not add any predictive power. This separation increase the dataset from 45 patients to **(NUMBERHERE!!)** spectra, which is a sufficient number of datapoints for a deep learning model. Displaying each sample is possible by visualizing each spectra as a line, each patient has $w * h$ spectra. Before the lines are drawn the maximum absolute value of each frequency within the entire dataset is found. Using these, the frequencies of each spectra may be normalized to have a maximum value of 1 and a minimum value of -1. By normalization the plots may be compared to identify outliers and determine necessary preprocessing measures. Appendix A contains a comparison by the common pattern these samples portray

compared to one of the samples which will be removed from the dataset. This comparison is only one method of detecting outliers. The reason for this samples removal, along with the removal of other problematic samples is further motivated in the section below.

## 3.2 Data preparation

The preparation of the data is vital to form non-biased models, it is equally important to guarantee that the model is able to generalize to all possible tumor samples. This means the model is expected to perform well on previously unseen data. Should this be impossible the resulting predictive model will fail to grasp the necessary features relevant to each sample-class. The model would in this respect overfit to the data used to train it. To avoid this we preform qualitative analysis on the data to determine the plausibility of such a model. Each sample is categorized according to their subdivision, there are six distinct subdivisions defined by Ceccarelli et al.[5], the labels are denoted by LGm1 - 6.

### 3.2.1 Balancing

The balancing stage is of tremendous importance, should the model be made from the data without balancing, there is great chance the model would tend towards predictions for the majority class(LGm2). The model will as a consequence of it's learning-algorithm become biased towards certain predictions **PROVE THIS!**. Over exposure to examples of a certain class will force the model to associate features with that class which in turn redirects focus from the classes for which that feature could be significant. The data suffers heavily from this problem, a table displaying the distribution of the spectra in each class is shown in Table 3.1.

| Class | LGm1 | LGm2 | LGm3 | LGm4 | LGm5 | LGm6 |
|---|---|---|---|---|---|---|
| # of samples | 5 | 16 | 7 | 13 | 15 | 5 |
| # of spectra | 37319 | 210586 | 39636 | 50660 | 62256 | 20176 |
| percentage | 9% | 50% | 9% | 12% | 15% | 5% |

Table 3.1: Table showing the distribution of data in the initial dataset, the majority class is LGm2 and minority is LGm6. The classes 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 display the total number of spectra across each class, these must be considered when balancing is performed note that there are an equal amount of samples in LGm1 and LGm6 but have a different number of spectra. This is due to the size of each sample drawn from the tumor. 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 5% of the entire dataset. Some samples prove to be problematic as their size make them too big to load into memory. Some even included what appears to be erroneous readings. The plot of one erroneous example is shown in appendix A. For simplicity, these are removed from the analysis, some included a great number of spectra which cause memory errors when they are plotted. The maximum is also present in these plots and as can be seen in appendix A the spectra appear to be erroneous. While there are methods of improving their shape, there is enough data to move to the next stage even after their removal. Following the removal of the problematic samples the data is now represented in Table 3.2

| Class | LGm1 | LGm2 | LGm3 | LGm4 | LGm5 | LGm6 |
|---|---|---|---|---|---|---|
| # of samples | 5 | 11 | 4 | 13 | 15 | 5 |
| # of spectra | 37319 | 71846 | 14896 | 50660 | 62256 | 20176 |
| percentage | 15% | 28% | 6% | 20% | 24% | 8% |

Table 3.2: The data distribution after removing problematic samples

Before the data is balanced, the testing data is selected and separated from the rest. 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 overfit to the patients in training and if the patient samples are heterogeneous with respect to the other samples of the same class. 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, quadrupling the samples in LGm3 and tripling the samples in LGm6. The distributions of the final training dataset is shown in Table 3.3.

| Class | LGm1 | LGm2 | LGm3 | LGm4 | LGm5 | LGm6 |
|---|---|---|---|---|---|---|
| # of spectra | 22374 | 51698 | 11296 | 34276 | 43892 | 12976 |
| Factor | 2 | 1 | 4 | 1 | 1 | 3 |
| # of spectra | 44748 | 51698 | 45184 | 34276 | 43892 | 38928 |
| percentage | 17% | 20% | 17% | 13% | 17% | 15% |

Table 3.3: Distribution of the training data

Table 3.3 shows the number of spectra before multiplying on row one and the number of spectra after on row three. The precentages on row four shows the distribution on the modified dataset. The percentages are now closer in scope, meaning each class is now in relative balance to the other classes.

### 3.2.2 Clustering

To ensure each sample include relevant information requires intuitive analysis. During extraction it is possible droplets of certain fluids were present, it is also possible the laser hit the tissue at a thin point which would scan the plastic underneath. This data must be removed to avoid model dependency on erroneous spectra. To find and remove the erroneous spectra, a clustering method is used. A suggested method is the k-means clustering algorithm by MacQueen [11] performed on every spectra on a sample by sample analysis. The resulting clustering may be displayed as a 2-dimensional image, some of which are shown in **APPENDIX!!**. K-means do however suffer heavily from outlier influences, as these outliers may be present at this stage, it is necessary to consider other methods. Hierarchical clustering is a clustering method which avoids outlier dependency by separating each individual spectra into its own cluster and then reducing the number of clusters depending on the linkage type. Complete linkage is used to ensure diversity among the clusters, measured by euclidean distance[16]. Each spectra is now labeled by a class number, all labels may be displayed as a two dimensional image due to the shape of the samples. Examples of such images are shown in appendix A. The cluster shapes suggest certain outliers within all patients, these outliers are removed accordingly to retain the tissue information exclusively. **Note: I am still unsure of this step, the clusters seem to also capture what could be perfectly normal tissue. Raise this concern with supervisors, it is crutial to get this part right so as to present rigorous results. I have saved images using alternative distance metrics however they seem to largely suffer from the same problem (some even seem to recreate the tumor image entirely)**

### 3.2.3 Feature selection

**Note: Discuss the following with supervisors: feature selection is now done before clustering. The reason I describe clustering first is that I think it would be best to perform clustering first and then extract the features. This extraction did yield features which made the clustering results "make sense" so to speak. However the outliers (eventual plastic or other non-tumor substance) is present during this extraction, meaning non-tissue is taken into consideration during the process. Would it not be better to extract them after removing non-tumor samples? Then, the yielded features might be different! If this is not the case we can proceed without concern, but what if this fails?**

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. Due to the number of spectra inside each sample memory quickly becomes an issue, it is therefore relevant to examine which frequencies would have the most impact on classification. For this reason the best features are extracted with SelectKBestfeatures [16]. The 70 best features were extracted from each spectra which is shown to be appropriate to recreate the original tumorshape by performing clustering

# Chapter 4

# Results

# Chapter 5

# Conclusion

# Bibliography


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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

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

# Appendix A

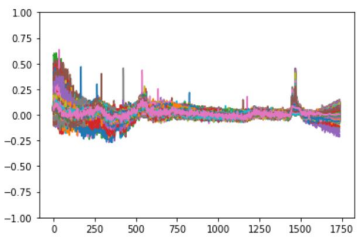
# Appendix

## Appendix B

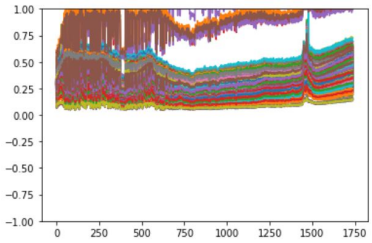
**INSERT TABLE**

## Appendix C

**INSERT Image**




(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