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 where is applied with 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. 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 method also has a considerably time complexity and is prove to be NP-hard [12].

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. 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.

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

Furthermore the algorithm uses a linkage criterion to decide which clusters to merge in the agglomerative method.

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

In this section the data is presented in greater detail. Due to the low number of samples available, it is necessary to examine each sample in detail to determine it's predictability. The data must be sufficiently diverse between the given classes and similar within those classes for the predictive model to work appropriately. Should this not be the case, the model will struggle to reach desired performance by either failing to capture basic features of the data or by overfitting to it. Within this project it is also vital to examine and determine if the samples of one class are different with respect to samples of other classes i.e. The sample-classes are heterogeneous. This criterion is important, since there must be a sufficient difference between the different classes to separate them appropriately. This also requires that the within-class samples are sufficiently similar i.e. All tumors belonging to the same class are homogeneous. Should this criterion not be satisfied the model is expected to overfit to the training data and fail to generalize to the test data. It is also necessary to prepare for eventual worst case scenarios. A possible worst case scenario is that all tumors are heterogeneous, which would mean that any model would be unable to satisfyingly distinguish between the classes. This means that we must evaluate the model on a patient by patient basis as the model would learn to recognize patterns in individual patients. This becomes relevant when selecting which samples to use for training the model and testing it.

3.1   Data Representation

The data of the glioma patients is provided by **PROVIDERS HERE** and consists of the Raman-spectrum of 45 tumors from separate patients. Multiple samples of tissue was extracted from the same tumor in some cases, yielding several samples for the respective tumors. The samples are sorted by the patient to whom they belong; this maintain separation between the patient samples. Moreover, there is

large variation with regard to the sample shape within the data. Each sample is a 3-dimensional array of size $(w, h, 1738)$ 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 number 1738 is constant through all samples and represent the length of the Raman-spectrum. 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 the methylation-types by feeding in one of these samples i.e. one vector of shape $(1, 1738)$

3.2 Data preparation

The preparation of the data is vital to form non-biased models, provided that the data can be restructured and rescaled to sufficiently represent the majority of samples. Should this be impossible the resulting predictive model will fail to grasp the necessary features in the data for forming predictions. The model would in this respect overfit to the data used to train it. To avoid this we preform a qualitative analysis on the data to determine the predictability. Each sample i cathegorized according to their methylation-types, there are six distinct classes of methylation in this data, the labels are denoted by: LGm1, LGm2, LGm3, LGm4, LGm5 and LGm 6.

3.2.1 Balancing

The data used to train a model must in some way be balanced. Should one label-class be over-represented with respect to the others, the model will as a consequence of it's learning-algorithm become biased towards certain predictions **PROVE THIS!**. The data suffers heavily from this problem, a table of the number of spectra in each class may be seen 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

Table 3.1 shows the per class separation in the data, the first row shows the labels of each class. The second row shows the number of samples belonging to each class, these are the tumors which will be analysed. The third row display the total number of spectra across each class, these must be considered when balancing is performed. The last row of table 3.1 shows the percentage each class makes of the entire dataset. Initially LGm2 is the majority class while LGm 6 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. For simplicity they are removed from the analysis. Their plots are shown in appendix **FILL IN APPENDIX!**5. 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. With the test-patients separated, balancing the classes which contain less elements by a factor larger than or equal to two than 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 is be crucial provided the sample pattern is correct. 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 two and the

number of spectra after on row four. The precentages on row five 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[13]. 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 5. 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 [13]. 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] 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.

[13] 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.

Appendix

Appendix A

INSERT TABLE

Appendix B

INSERT TABLE

Appendix C

INSERT Image