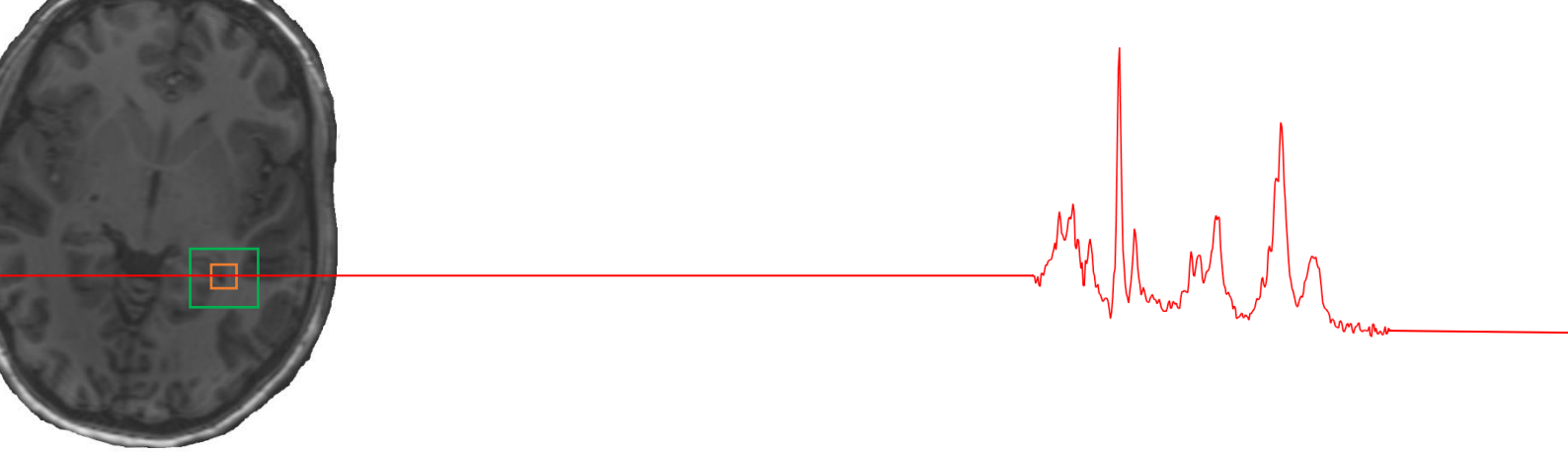

Differentiating Pediatric Brain Tumors based on MR Spectroscopy with Machine Learning

Clinical internship – M3

M.F. Brasz, BSc - S1478958
Neuro-Oncology – Princess Máxima Center
University of Twente

UNIVERSITY OF TWENTE.

prinses máxima centrum kinderoncologie

Graduation commission

Chairman &
Technical Supervisor UT:
Dr. ir. F. van der Heijden

Medical Supervisors:
Prof. dr. E. Hoving
Dr. M. Lequin

Technical Supervisors UMCU:
Dr. ir. J. Wijnen
Dr. ir. E.C. Wiegers

Process Supervisor:
R.J. Haarman

Extra Supervisor:
Dr. ir. A.M. Leferink

ACKNOWLEDGEMENTS

Almost twelve months ago I started my final internship and graduation project at the Princess Máxima Center, at the department of neuro-oncology/surgery. This was a place where I was able to return after a short internship a year before, and a place where my interest for research within pediatric oncology started. I want to express my deep gratitude to Eelco Hoving, who gave me the opportunity to come back to the PMC and work on a topic which allowed me to be involved in technical development to improve patient care. In my belief, this is exactly what technical medicine is meant for. Next to this, thank you for the encouragement and interest in my project.

Thank you Maarten Lequin for involving me in your daily clinical work, but also motivating me and setting high expectations for the research project. Seeing your drive and passion for this topic was an amazing motivation and it resulted in collaborations with many other centers. Hopefully, there will be more to come.

Ferdi van der Heijden, I am more than grateful for your time, dedication and honest feedback. I could not have performed this research without your help. Thank you for sharing your expertise, providing machine learning lectures, visiting the PMC and answering many of my questions.

During this year, I got the opportunity to work with the Imaging Division department of the UMC Utrecht. Here I was able to further develop my soft skills and obtain technical background knowledge from such resourceful and kind people. Thank you all for making me feel welcome. A special thanks goes out to Jannie Wijnen and Evita Wiegers. Jannie, it was a great experience working with you. You educated me on, what was for me, a whole new topic. Our meetings and their results were creative and fun, and your feedback was much appreciated. Both you and Evita helped me to obtain a clearer vision on the process of research and how to set goals. Your questions challenged and stimulated me to come up with new ideas. Evita, you were always ready to help and I could share my achievements, questions and insecurities with you. You have helped build my self-confidence as a technical physician. You were able to make me feel equal by your compassion, involvement and many other qualities.

During my internship I had the pleasure to be able to share MR spectroscopy data with other centers and discuss the topic with other brilliant minds. These people showed great interest in the project and with their help, there are so many possibilities to expand the research. Thank you for the collaboration and feedback.

The last few years Rian Haarman has guided me during my growth on professional soft skills, but also on a much more personal level. Your enthusiasm and curiosity to my development, thoughts and ideas have helped me to answer some difficult questions. I started my internships more than two years ago, and since then I have had the pleasure to share such an immense growth with you. Thank you for guiding me during my search for

my identity as a technical physician, but also becoming a more grounded human being, capable of much more acceptance and self-love.

As my master's draws to a close I am able to reflect on all three years and there are more people I would like to thank. Thank you Anne and Frank for being such good teachers, providing a safe environment and teaching me to be more analytical. Both your critical way of thinking and honesty motivated me to try harder. Therefore, you have provided me with the opportunities to grow academically and personally. Thank you Tanja, for being such a good friend and being a safe haven at the university, you are always there to listen and you are so appreciated. Then there are Beau and Michelle, who I studied and travelled with, and who I shared many coffee breaks during our time at the UMC Utrecht. Thank you for being there during the relaxed and stressful times. I would like to give a shout out to Femke and Martine, I am beyond lucky to have you as close friends and roommates. We have our own version of Friends. There are so many more people who play such an important role in my life and who I was able to talk to, to keep me motivated, to share a moment of self-reflection, or discuss my research project with. I count myself very lucky for that.

Finally, I would like to thank my mom, dad, brother and sister. You are always there for me and I am so grateful for your trust and love.

Mechteld

May 2020

ABSTRACT

The attainment of an accurate diagnosis is key in treatment planning in pediatric neuro-oncology, because treatment and prognosis are based on tumor location, histopathology and staging. Magnetic resonance (MR) spectroscopy during routine scanning could improve diagnostic value of imaging and aid in differentiation of pediatric brain tumors earlier in the diagnostic stage. Imaging at cellular level remains impossible, but MR spectroscopy provides direct insight into biochemical processes in cells. Metabolism in cancer cells differs for different tumor types. Several studies showed the added value of MR spectroscopy in differentiating tumors types. Therefore we aim at developing a tool based on machine learning (ML) that can help radiologists with assessing brain tumor type.

Since raw data is saved since June 2019 in the Princess Máxima Center (PMC), a lot of raw data from MR spectra before June 2019 are missing. Therefore, the value of the use of screenshots is being evaluated. From these two data sources, in total four datasets were obtained to cluster the data; 1) vectors of the original raw data, 2) concentrations determined by automatic quantification of the raw data, 3) a line segmented from screenshots, and 4) peak/height ratios determined from screenshots.

After performing quality control, three different classes were formed with a total of 32 screenshots and 17 raw datasets. Three tumor types (medulloblastomas (10 screenshots; 5 raw datasets), DIPGs (11 screenshots; 8 raw datasets) and pilocytic astrocytomas (11 screenshots; 4 raw datasets)) were classified with different classifiers.

Feature reduction was performed with feature selection and feature extraction methods (principal component analysis (PCA)/linear discriminant analysis (LDA)). Five different machine learning algorithms were applied (Linear Discriminant Classifier (LDC), Quadratic Discriminant Classifier (QDC), K-nearest neighbor (KNNC), support vector machine (SVM), feed forward neural networks (FFNN)). Error rates were determined with help of a cross validation method. The lowest error rates after a PCA were as followed: raw data vector (0.2000); raw data concentration quantification (0.1294); segmented line from screenshots (0.2000); peak/height ratios from screenshots (0.2562). The lowest error rates after an LDA were: segmented line from screenshots (0.0000); peak/height ratios from screenshots (0.2188).

Because of a high suspicion for overfitting, the dataset with the segmented line was trained and tested with different training and test sets. The error rates varied from 33% to 66% for the test set. KNNC and FFNN provided the lowest error rates. The simpler, more robust, classifiers (LDC and QDC) were less accurate in the classification of the test set. More freedom from the KNNC and FFNN provided better results, but overfitting should be avoided by creating different test sets, or by applying cross validation.

For now, the datasets were based on the amount of features. But medulloblastomas, DIPGs and pilocytic astrocytomas are already easier to distinguish than other tumors by neurological examination and MR imaging. More data should be collected to create datasets with a higher clinical impact.

The results demonstrate the possibility of using screenshots as a replacement for raw data. This could give many centers the possibility of using their screenshots and broaden their dataset. The amount of data should be increased to further examine the potential of differentiating pediatric brain tumors based on MR spectroscopy with machine learning to become a clinical relevant classification tool.

CONTENTS

ABBREVIATIONS

Abbreviation	Phrase	Pag. Nr.
Ala	Alanine	5
ANN	Artificial neural networks	7
Cho	Choline	5
Cit	Citric acids	5
CNS	Central nervous system	1
ConcD	Concentration dataset	11
Cre	Creatine	5
CRLB	Cramér-Rao lower bound	12
CSF	Cerebrospinal fluid	2
CSI	Chemical shift imaging	4
DIPG	Diffuse pons intrinsic glioma	5
FFNN	Feed-forward neural networks	8
FID	Free induction decay signal	12
GCT	Germ cell tumor	10
Gln	Glutamine	5
Glu	Glutamate	5
Glx	Sum of glutamine/glutamate	5
ICC	Intra/Interclass correlation coefficient	15
KNNC	K-nearest neighbor classifier	7
LDA	Linear discriminant analysis	7
LDC	Linear discriminant classifier	7
Lac	Lactate	5
Lip	Lipids	6
LineSD	Line segmentation dataset	12
mI	Myo-Inositol	6
ML	Machine learning	1
MM	Macromolecules	6
MR	Magnetic resonance	1
MRS	Magnetic resonance spectroscopy	1
NAA	N-Acetylaspartate	6
NMR	Nuclear magnetic resonance	4
PCA	Principal component analysis	7
PCr	Phosphocreatine	5
Peak/HeightD	Peak/Height dataset	12
PMC	Princess Máxima Center	2
PRESS	Point resolved spectroscopy	4
Ppm	Parts per million	4
QDC	Quadratic discriminant classifier	7
RawD	Raw data dataset	12

RF	Radio Frequency	3
SVM	Support vector machine	7
SVS	Single voxel spectroscopy	4
Tau	Taurine	6
TE	Echo time	4

1. INTRODUCTION

The most common solid tumors in children are located in the central nervous system (CNS), accounting for nearly 25% of all childhood cancer. [1, 2] Brain tumors comprise many different diagnoses, varying from benign to malignant. High grade tumors have a worse mortality rate compared to low grade tumors, but the morbidity rate is very high for both these types. [2] The mortality rate for all CNS tumors remains worse compared to other pediatric solid tumors. The 5-year survival rate for all brain tumor types combined has been increased to 70% over the past decade, which is still worse compared to the survival rate for all pediatric solid tumors, which is 80%. [3-5]

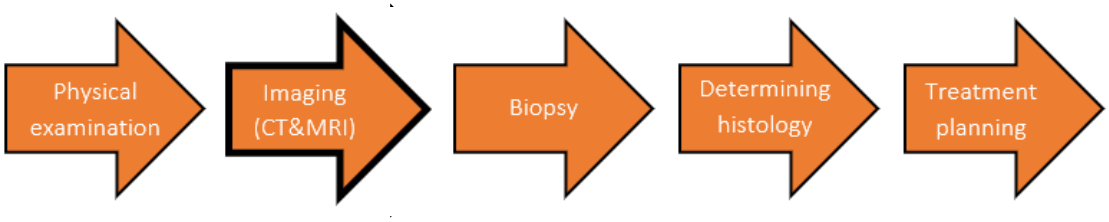

Figure 1.1: The clinical pathway, undertaken by pediatric patients is displayed in this figure. At time of the second arrow, MR spectroscopy could aid in differentiation of tumor types.

An accurate diagnosis is key in treatment planning for children with brain tumors, because treatment and prognosis are based on tumor location, histology and staging. [6] A non-invasive, cost effective tool would be valuable for brain tumor characterization. [7] Characterizing pediatric brain tumors based on solely magnetic resonance (MR) imaging is difficult. Histopathology remains necessary to derive a complete diagnosis based on the tumor's histologic and biologic behavior and to determine the treatment strategy. Therefore, histopathology is the gold standard to classify tumor types. However, research has been performed to investigate new techniques to capture the physiological processes within the tumor to accelerate the diagnosis and treatment planning.

Proton magnetic resonance spectroscopy (MRS) is a non-invasive technique, able to provide biochemical information about tissue. [7] Multiple studies show the possibility to differentiate brain tumors with the help of MR spectroscopy. [7-10] High and low grade gliomas can be distinguished, based on MR spectroscopy data, but also different tumor types can be differentiated as well. These researches have shown the added value of MR spectroscopy in the non-invasive evaluation of tumor tissue. The diagnostic value of imaging might be higher when MR spectroscopy is included in routine scanning. (Figure 1.1)

Unfortunately, interpretation of the spectra can be difficult. [11] Therefore, an algorithm to differentiate tumor types as performed in other studies may be useful. [8, 11, 12] Machine learning (ML) makes it possible to classify medical images. [13] This can also be applied to the MR spectra. [14, 15] Next to this, ML is able to recognize patterns that have not been discovered yet. Patterns that seem to be meaningless for radiologists, could play

an important role in the differentiation of tumors. This could enable accurate non-invasive brain tumor diagnoses.

Datasets, containing raw data and automatically generated bitmap images from processed data, from here on called screenshots, are obtained from the Princess Máxima Center (PMC). These images are generated by the MRI host and saved in .jpeg format. Using the screenshots an effort is made to recreate raw data if it is lost, since raw data has not been saved in previous years. Histology of the brain tumors has been verified by biopsies or cerebrospinal fluid (CSF) analysis. These datasets can be used in a ML algorithm. [13] With histology-based diagnosis, the ML algorithm can be based on labels (i.e. the diagnosis), and not on given features by radiologists. This makes it possible for the ML algorithm to search for new features, recognizable by a computer.

Pattern recognition can result in rapid differentiation of pediatric brain tumors and might be able to filter out human diagnostic misinterpretations by radiologists. This research aims to obtain MRS data from screenshots and MRS raw data from the Princess Máxima Center and classify pediatric brain tumors based on this MRS data. The proof of concept is to classify three pediatric brain tumor types, and a long term goal would be to develop a tool with diagnostic value to help differentiate pediatric brain tumors during routine scanning.

2. BACKGROUND INFORMATION

2.1 Magnetic Resonance Spectroscopy

MR spectroscopy can be used to analyze the chemical composition of tissue by obtaining signal from nuclei which possess magnetic properties. These nuclei have a magnetic moment (spin) and when they are positioned in an external magnetic field, they will start to spin around their axis with a MR frequency (known as the Larmor frequency). By applying a radiofrequency (RF) pulse (with the same MR frequency as the Larmor frequency) to the tissue, the protons flip and they send out an echo when they are returning to their steady-state position (known as T1/longitudinal relaxation). After an RF pulse has been applied, the protons start dephasing in the transversal direction as well (known as T2/transversal relaxation). A refocusing pulse is applied to help rephasing protons and recover the transverse magnetization. With the obtained echo, MR spectroscopy is able to detect molecules with relatively low molecular weight and the ability to move within intra- and extra-cellular spaces. Since many of these molecules are part of a metabolic pathway, MR spectroscopy is sensitive to detect aspects of metabolism within tissue. [16, 17] Figure 2.1 shows a healthy example of an MR spectroscopy spectrum from the brain and its metabolic aspects. The area under the curve of a certain peak is related to the concentration of that metabolite.

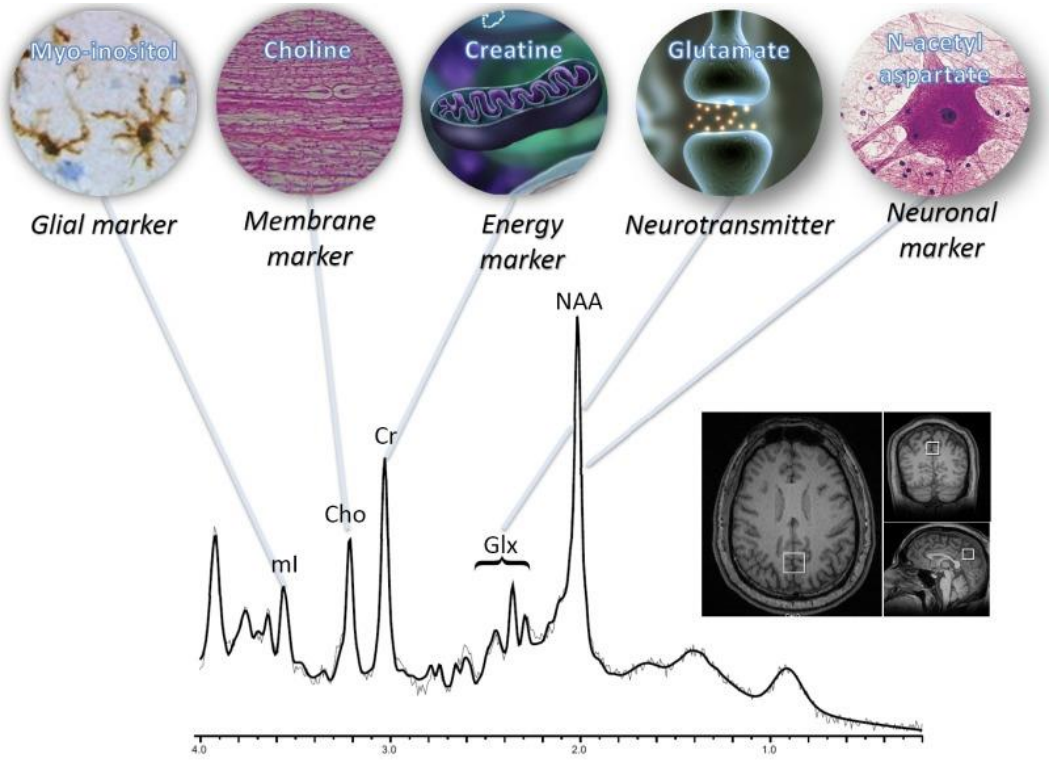

Figure 2.1: An example of an MR spectrum, obtained from brain tissue. Different metabolites and their function are shown with the corresponding peaks in the spectrum. [18]

^{1}H, ^{31}P, ^{19}F, ^{13}C or ^{23}Na are various nuclei which can be used to obtain in vivo MR spectra of metabolites. ^{1}H MRS is widely used, because ^{1}H nuclei are present in many metabolites and this technique can be performed on clinical MR systems. [16]

The frequencies most common used in clinical MR spectroscopy are 64 (1.5T), 128 (3T) and 298(7T) MHz. Metabolites in an MR spectrum can be recognized because of the chemical shift. The chemical shift effects are related to the electrons which surround the proton and affects the magnetic field that is sensed by the nucleus, and thus its Larmor frequency. Therefore, different metabolites resonate at a different frequency and this information can indicate which molecules are present. The chemical shift is expressed in parts per million (ppm). A proton with a dense cloud of electrons will resonate more upfield in the spectrum (lower ppm), while the MR frequency of a less shielded proton will shift downfield in the spectrum (higher ppm). [19] J-coupling results in peak splitting, due to neighboring protons within the same molecule. Due to bad resolution, these peaks cannot be resolved from each other. This can also appear as peak broadening. [19-21]

Because of the low absorption of electromagnetic radiation, NMR (nuclear magnetic resonance) is an insensitive technique. But the low energy absorption results in a noninvasive technique, perfect for in vivo measurements. The properties of ^{1}H nuclei in different tissues have shown different NMR properties. Reports mention different properties (relaxation time) for malignant tissue, when compared to normal tissue. [19] Next to this, the different levels of metabolites help to distinguish brain tumors. [10]

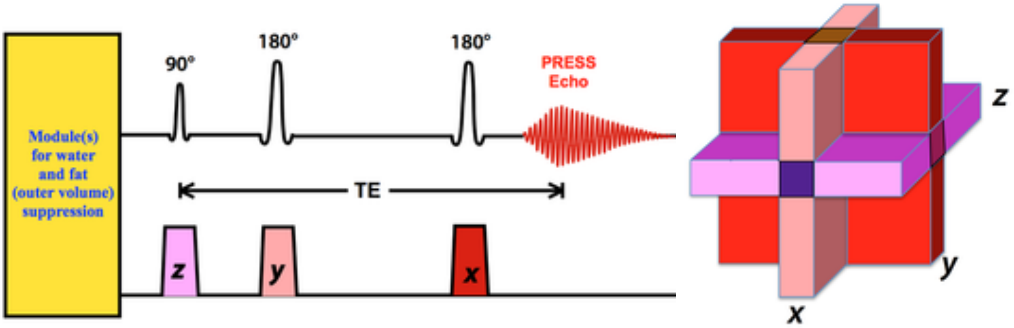

Figure 2.2: A simplified diagram of the PRESS sequence. A 90° RF pulse is given in the Z direction, followed by two 180° RF pulses in de X and Y direction. The protons in the voxel, excited by all three pulses, produce an echo at time TE. [22]

A Point RESolved Spectroscopy (PRESS) sequence is used for clinical MR spectroscopy. Figure 2.2 shows one excitation pulse and two refocusing pulses to select a voxel. An RF pulse (90° – 180° – 180°) is applied in all three orthogonal gradients (x, y and z direction) and only the protons which have experienced all three RF pulses, within the cubic-shaped voxel, produce a signal at the echo time (TE). This technique is used for single voxel spectroscopy (SVS), and can be used for chemical shift imaging (CSI) as well. This technique derives a signal from multiple voxels and can be performed 2D and 3D.

SVS is most widely used in clinical setting. This setting provides a high signal-to-noise in a relatively short scanning time. This technique is also used in the PMC. The voxel is placed

in tumorous brain tissue, in the solid part. The cystic components of a tumor as well as surrounding bone, epidural fat or liquor are avoided as much as possible. [23, 24]

2.2 Metabolites

The following metabolites have shown to differ between various pathologies and brain tumors:

Alanine (Ala, 1.47 ppm)
Ala is a nonessential amino acid, present in the brain of mammals. An increase of this metabolite has been observed in meningiomas. The peak can be overshadowed by resonance of lactate and lipids. Longer echo times, or spectral editing is needed when differentiating the Ala peak from lipids. [19, 24]

Choline (Cho, 3.21 ppm)
An increase of the choline signal is seen in cancer. Cho is involved in the synthesis and degradation of the phospholipids pathway. Therefore, it is involved in the cell membrane turnover. At short echo times the peak overlaps mI, glucose and Tau. [19]

Citrate (Cit, 2.57 ppm - 2.72 ppm)
Cit helps generating energy and biosynthesizes lipids. Due to high cell proliferation in certain brain tumors, the metabolism in the cell changes. A decrease of a Cit peak for diffuse intrinsic pontine gliomas (DIPGs) during or after treatment, compared to pretreatment MRS data, might indicate worsening of the malignancy grade. [7] Perturbation in citrate levels occurs in pediatric brain tumors and can be imaged with MR spectroscopy. The spectra of these tumors show an increased level of Cit. The resonances can partially overlap with resonances from creatine and glutamine. [7, 19, 25, 26]

Creatine (Cr, 3.03 ppm, 3.93 ppm)
Together with phosphocreatine (PCr), Cr plays an important role in the energy metabolism of tissue. This metabolite is present in neurons and glial cells. The concentration is higher in gray matter, than in white matter. Within tumorous tissue, the Cr levels can be decreased. [19]

Glx (2.0 - 2.46 ppm, 3.6 ppm - 3.8 ppm)
Glutamine (Gln) is synthesized from glutamate (Glu) in astroglia cells and is broken down again to Glu in neural cells. Therefore, it plays a role in regulation of neurotransmitters. Glu is known to be the major excitatory neurotransmitter in the brain and contributes to the mitochondrial metabolism. [24] GABA functions as a neurotransmitter. [27] Glx is the sum of both, and is used to quantify the presence of Gln and Glu separately. Gln, Glu and GABA are difficult to distinguish at low magnetic fields. [19]

Lactate (Lac, 1.3 ppm, 4.1 ppm)
The presence of Lac in the brain is normally around 0.5 mM. The metabolite is the end-product of anaerobic glycolysis, but in cancer cells increased levels of Lac are caused by lactic acid fermentation after high rate aerobic glycolysis. [19] This phenomenon is called

the Warburg effect. [28] An increase in Lac is shown in spectra of low grade tumors, due to this different metabolism in cancer cells and in necrotic tumors, due to ischemia. [7, 28, 29] Lac can overlap resonances of macromolecular structures and lipids. [19]

Lipids (Lip, 1.3 ppm, 0.9 ppm)
The lipids form the myelin sheaths, surrounding healthy neurons. Higher levels of Lip are seen in multiple tumors, among which are meningiomas and glioblastomas. After damage or disruption of tumorous tissue, the macromolecules (MM) transform into mobile lipids. [29] An increase of Lip could represent necrotic areas within tumors. [19, 24]

Myo-Inositol (mI, 3.55 ppm)
Although the exact function of mI is unknown, it is suspected to be a glial marker and may play a role in the osmotic regulation of the cell. This cyclic sugar alcohol has six detectable protons and gives rise to multiple groups of resonances. [19]

N-Acetyl aspartate (NAA, 2.01 ppm, 2.5ppm)
The exact function of NAA remains unknown, but it might play a role in storing acetyl groups for myelin or fatty acids syntheses. It could also be involved in osmoregulation, or be one of the breakdown products of the neurotransmitter NAAG. The peak is found to be a marker of neuronal cells and gives information about the neuronal density or neuronal dysfunction. [7] The concentration NAA is slightly higher in gray matter when compared to white matter. [19] MR spectra from tumorous tissue are related to a decrease in NAA, due to neuronal cell loss. This metabolite is only located in the nervous system and peaks at 2.01 ppm. Smaller doublet-of-doublets occur at different values in the spectrum.

Taurine (Tau, 3.25 ppm, 3.42 ppm)
Taurine is an amino acid in the brain and plays a role in cytoprotection. An increase of this metabolite has been observed in medulloblastomas. [7, 14, 30, 31] The signals from Tau should be taken into account when quantifying the Cho peak, because the taurine peak forms a significant part of this peak. Especially when obtaining spectra with lower magnetic fields, mI, Cho and Tau overlap. [19]

2.3 Machine Learning

Machine learning can be used to classify the spectra of different tumors. [8-10] The algorithm uses datasets provided by the user. The amount of given features (datapoints) determine the dimensions which are used by the algorithm to classify data. Each extra feature, means an extra dimension in which the data is classified. With unsupervised ML, an algorithm determines the labels. It is able to classify a dataset, based on features. When supervised ML is used, data is classified based on the given labels. This means, labels are already given to the data with prior knowledge and these labels indicate which object belongs to which class. Since the ground truth is already determined for all the images by biopsies, the data is labeled (based on the ground truth) and a supervised ML can be applied. [12]

The amount of information has to contain enough variance to classify the data, but too much information can result in overfitting. Chances of overfitting increase when there is a high number of features and a small number of samples. The classification will be based on that specific data, but the algorithm is not capable of learning with new data. When using a small sample size dataset, it is best to use the amount of features necessary to provide enough variance. Since every feature adds an extra dimension, too many features within a dataset results in too many dimensions. To decrease the amount of features (and the dimensionality), feature selection or feature extraction can be used. Feature selection does not alter the original features, but selects the most significant features based on some criterion, such as level of variance. [32] Another manner to reduce dimensionality is feature extraction. Both principal component analysis (PCA) and linear discriminant analysis (LDA) are linear transformation techniques (Figure 2.3). LDA is supervised, whereas PCA is unsupervised. PCA ignores class labels when reducing features, but LDA does not. LDA helps to determine useful features by projecting the data and detect when classes are well separated. Since LDA outperforms PCA, it is preferred to use an LDA. [33]

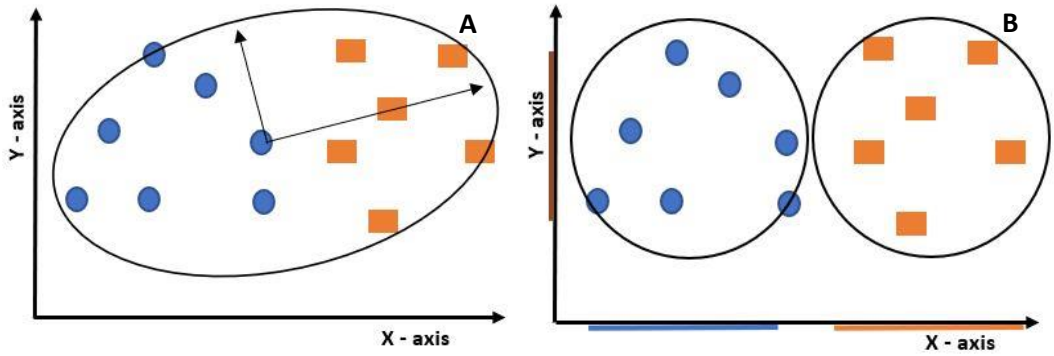

Figure 2.3: A PCA and LDA are visualized. (A) Visualization of a PCA. Classes are not taken into account, the PCA searches for features with the most variance. (B) Visualization of an LDA, a supervised dimensionality reduction method. On the Y-axis a bad projection is seen, while on the X-axis a good projection is seen. With the projection on the X-axis, the two classes can be distinguished.

Different algorithms can be implemented to use pattern recognition on MR spectra. A Linear Discriminant Classifier (LDC) is a classification method which separates the classes by a linear vector. A Quadratic Discriminant Classifier (QDC), relies on the same method, but it divides the classes with the use of a quadratic decision surface. These two methods have assumptions about the dataset and they are simpler than other classifiers. (Figure 2.4) They have less computational time and less internal parameters. There are also distribution free classifiers, that do not depend on prior knowledge about the distribution of the data. Important supervised distribution free classifiers used in different researches are artificial neural networks (ANN), support vector machine (SVM) and k-nearest neighbor classifier (KNNC). [10, 12, 13] These classifiers can be used for labeled data, to generate correct labels for prospective datasets, based on discovered features in the training data.

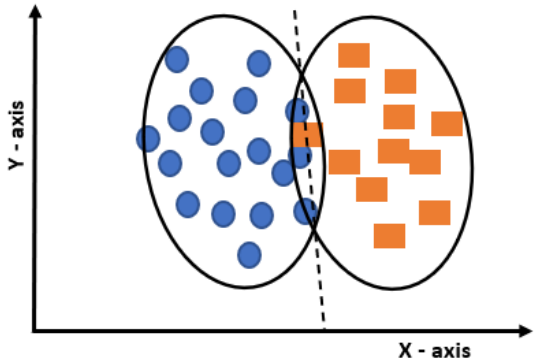

Figure 2.4: A visualization of an LDC and QDC. The straight line dividing the two classes is provided by the LDC. The circles are the decision boundaries provided by the QDC.

A KNNC provides classification on the principle of a plurality vote of its neighbors. K is a hyper parameter, which determines how far away to search for neighbors. Figure 2.5.A shows an example of KNNC. An unclassified datapoint has to be classified. When k is chosen to be three, the three nearest neighbors vote for Class A, but when k is chosen to be seven, the seven nearest neighbors vote for Class B. SVM helps with classification by creating the best line that segregates the classes in an n-dimensional space (Fig 2.5.B). A feed forward neural network (FFNN) is a type of an ANN. FFNN have units, also called neurons. They are arranged in at least three layers, an input layer, hidden layer and output layer. With back propagation this network gains new knowledge and adapts the weight given to the neurons. This ML algorithm is flexible, able to solve difficult tasks and with it, accessible software frameworks have been produced. However, training a FFNN is computationally expensive (Figure 2.6). [12]

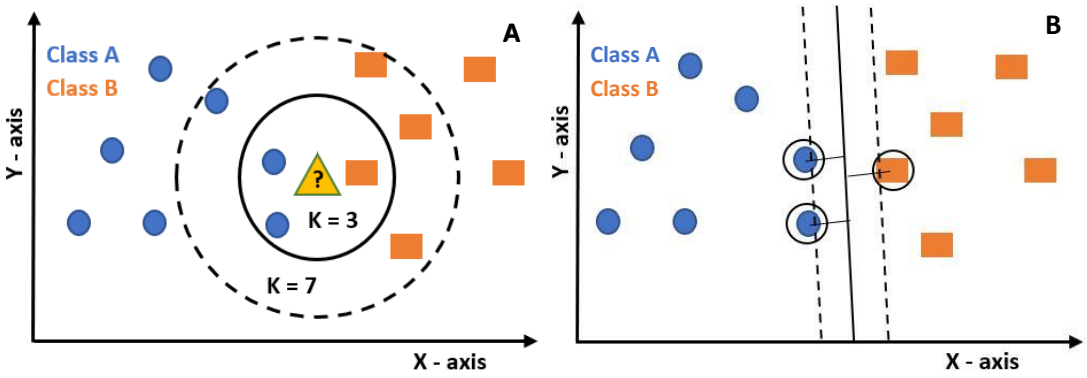

Figure 2.5: A KNNC and SVM are visualized. (A) Visualization of a KNNC, where K = 3 and K = 7 have been proposed as hyperparameters. (B) Visualization of an SVM, where the distance from three datapoints orthogonal to the line is the same.

Cross validation is used to measure the performance of the algorithm. It makes multiple smaller datasets and one of them becomes the test set, while the others become the

training set. Every time the algorithm trains, another sample becomes the test set. The average of the error rate is provided as a result. [30]

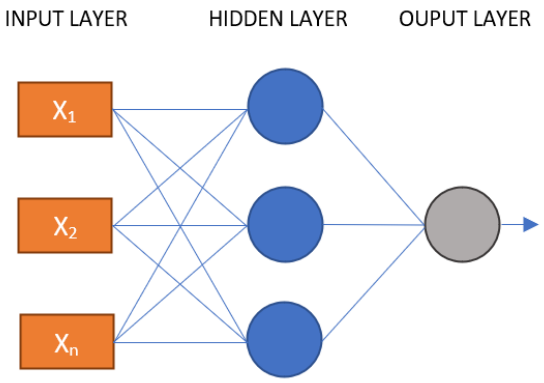

Figure 2.6: A FFNN with back propagation. The layers have different weights, based earlier obtained results and their feedback.

3. DATA ACQUISITION

The output of a machine learning algorithm is determined by the given input. Therefore, the manner of obtaining data reflects on the outcome. Because this was an important part of the research, a separate chapter about data acquisition is included.

Raw data is preferred when collecting MR spectroscopy data. Unfortunately, within the PMC, raw data for MR spectroscopy has not been saved till June 2019. Another way of collecting MR spectroscopy data is through the use of screenshots. Within this chapter an effort was made to recreate the raw data from screenshots. A method to obtain this data will not only be helpful to recreate data, but may also be used for an easy accessible tool for medical personnel. Working with raw data is harder to implement in clinical routine than working with the screenshots seen by radiologists daily.

3.1 Methods

3.1.1 Study design

This retrospective, single-center, study was performed during 2019 and 2020 and included pediatric patients who presented with a (suspicion of a) brain tumor to the institution from 2016 to 2020. Ethical committee approval was waived. Patients were selected through a search of an in-house clinical neuro-oncology database. Patients, 21 years of age or younger, with a clinical diagnosis of a brain tumor based on analysis of a biopsy or CSF sample, and an acquired MR spectrum during MRI scanning, were included.

3.1.2 Clinical Data

The electronic medical record for each patient was reviewed for the following: 1) demographic information, including age and sex, 2) tumor histology (e.g. low grade glioma, high grade glioma, medulloblastoma, ependymoma, pilocytic astrocytoma, germ cell tumors (GCT), DIPG), 3) tumor site (pineal, suprasellar, midline, posterior fossa, purely intraventricular), and 4) treatment history, including prior surgery (date and biopsy versus debulking), chemotherapy (date and type), and radiation planning dose.

3.1.3 Scanner Hardware and sequence parameters

All pediatric brain tumor spectra were acquired with the following MR systems; two 1.5T (Achieva and Ingenia, Philips Healthcare, Best, The Netherlands) and four 3T (Ingenia, Ingenia CX, Achieva and Elition, Philips Healthcare, Best, The Netherlands). Standard imaging protocol consists out of T1 and T2 weighted brain images, and T1-weighted post contrast. Single voxel MR spectra were acquired with Point RESolved Spectroscopy (PRESS)**,** 64 – 128 averages, TE= 35 - 40 ms, TR=2 seconds. Cubic voxels varied from 0.6 cm^3 to 6 cm^3.

3.1.4 Imaging protocol

Data was acquired in accordance with current clinical protocol, this included a pre-gadolinium sagittal 3D T1, axial DTI, ASL, and a SV PRESS sequence. Following an intravenous injection of gadovist at 0.1 mL/kg, a sagittal 3D T1 and T2 was obtained. Scan parameters such as voxel placement and size were optimized per patient.

3.1.5 Datasets

Different methods were used for data acquisition. The data collection started with two different data sources. Raw data was obtained directly from the MRI scanner and screenshots were generated by the MR machine and send to the archive system (PACS). From these two data sources, four different datasets were created:

1. The raw dataset, referred to as RawD.
2. Concentrations of the metabolites obtained from the raw data. This dataset will be called concentration dataset (ConcD).
3. Array data, obtained from the screenshot. This is a segmentation of the line representing the spectrum, referred to as line segmentation dataset (LineSD).
4. Peak/height ratios of different peaks from the screenshots, from here on named peak/height dataset (Peak/HeightD).

An example of the four different datasets is shown in Figure 3.1. This dataset will be denoted as Case I., and corresponds to data of a patient with a medulloblastoma in the posterior fossa.

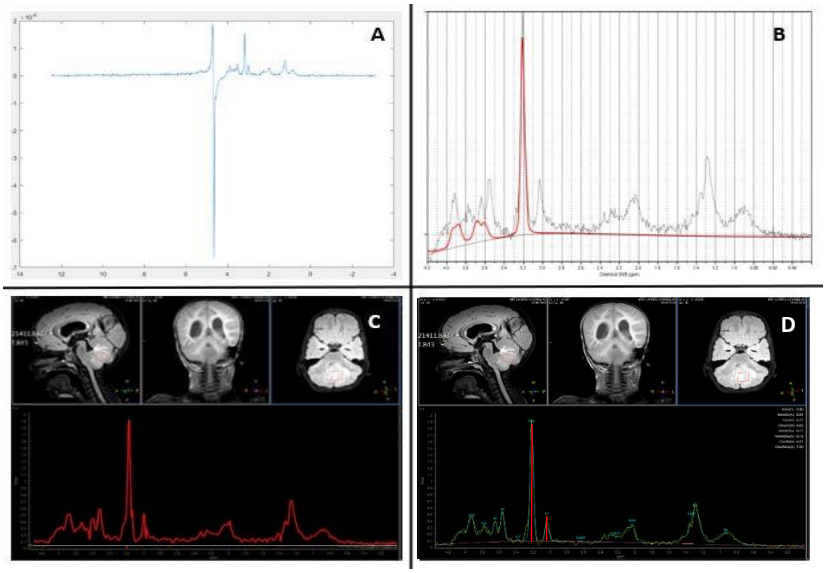

Figure 3.1: Two different data sources (raw data and screenshots) result in four different datasets from Case I. (A) A plot of the vector of raw data (RawD). (B) The fit of Choline is displayed to show the quantification of one metabolite for the concentration dataset (ConcD). (C) The line segmentation to recreate raw data out of screenshots (LineSD). (D) A screenshot where two peaks are highlighted, from which a ratio can be calculated (Peak/HeightD).

3.1.5.1 Data source 1: RawD and ConcD

The raw data consists of complex numbers resembling the free induction decay signals (FIDs). The FIDs consist of multiple overlapping signals and are superimposed compared to the background noise. Fourier transformation of a FID results in a spectrum, where every peak represents a certain metabolite. The spectrum also contains a large (residual) water signal. [34] The RawD has been collected from the MRI scanner. Since protocols are different on MRI scanners with different field strength (1.5T; 512 datapoints, 3T; 1024 datapoints), the same array size was obtained with zero-filling. [35] [36]

From the raw data, the concentrations of the metabolites were determined with software for automatic quantification (LCModel). This automatic quantification determines the concentrations for different metabolites present in the brain, based on a linear combination of simulated signals. These simulated signals are basissets, pre-installed in LCModel, to provide prior knowledge. A short TE = 38 ms simulated basisset is pre-installed for 1.5T acquired MR spectra and a short TE = 35 ms simulated basisset is pre-installed for 3T acquired MR spectra. The method fits individual metabolite signals in the spectra and calculates the area under the curve. The area under the curve in relation to the unsuppressed water signal, quantifies the concentration. Multiple metabolites which peak at almost identical frequencies can be distinguished and quantified by their signals from other parts in the spectrum. [37] If one metabolite is presented as multiple peaks at different frequencies, the quantification takes all peaks from that metabolite (the entire simulated spectral pattern per molecule) into account. A selection of the metabolites will be determined for the ConcD, to evaluate which dataset has the best ability to differentiate tumor types (Table 3.1). The following options are possible:

1. Include all twenty-two metabolites quantified with LCModel.
2. Include a selection of metabolites based on prior knowledge.
3. Include metabolites when over 50% of the cases have a Cramer Rao Lower Bound under 50%.

Based on prior knowledge is meant for metabolites which are already correlated to pediatric brain tumors. [7-10, 30, 31] The Cramer Rao Lower Bound (CRLB) is a measure of the uncertainty of the fit. A lot of different factors influence the result, so this percentage is not very trustworthy when measuring the uncertainty for one fit, but it is a good method to do group comparisons between groups of metabolites. [38] The dataset with the lowest error rate after ML will be selected and will function as the ConcD.

Table 3.1: The three possible datasets for ConcD. (A) All metabolites from which concentrations were quantified in the spectra with LCModel. (B) Selected metabolites based on prior knowledge. (C) Metabolites with more than 50% of the data with a 50% CRLB or lower.

A. All Metabolites	B. Selected Metabolites	C. Metabolites CRLB ≤ 50
Alanine	Alanine	Creatine + Phosphocreatine
Aspartate	Aspartate	Glutamine
Creatine	Creatine + Phosphocreatine	Phosphocholine
Phosphocreatine	GABA	Choline + Phosphocholine
GABA	Choline + Phosphocholine	Glutamine+Glutamate
Glucose	Glutamine+Glutamate	Glutathione
Choline	Glutathione	Myo-Inositol
Phosphocholine	Myo-Inositol+Scyllo	NAA
Glutathione	NAA + NAAG	NAA + NAAG
Myo-Inositol	Lactate	MM09
NAA	Taurine	MM20
NAAG	MM09+Lip09	Lip13a+b
Scyllo	MM13+Lip13a+b	MM09+Lip09
Taurine	MM20+Lip20	MM13+Lip13a+b
-CrCH2		MM20+Lip20
Glutamine+Glutamate		
Lipide09		
Lipide13a		
Lipide13b		
Lipide20		
Macro Molecules 09		
Macro Molecules 12		
Macro Molecules 14		
Macro Molecules 17		
Macro Molecules 20		

.

3.1.5.2 Data source 2: LineSD and Peak/HeightD

Since the radiologists has easy access to screenshots of the data, using these screenshots could possibly result in an accessible tool. To derive information out of the screenshots, two different methods were used.

Firstly, an effort was made to recreate the raw data by segmenting the yellow line from the spectrum. This line is meant to show the raw spectral data from 0 to 4.2 ppm. The red line shows the baseline and the blue line represents the fit, produced by software from the MR vendor. To decrease the amount of information provided by the screenshot, the spectral part of the image was cropped and used for analysis (Figure 3.2). Because there was a small change in the location of the spectra vertically on the image, the beginning of the x-axis was manually determined for every screenshot by locating the y-axis.

Because the image had to be transferred to a grayscale image for a segmentation algorithm, the maximum of contrast had to be created for the specific yellow line. Additive

color mixing helped with preserving the yellow line, while removing other colors from the image (Fig 3.2.C). [39] After preserving the line, the image could be transferred to grayscale and a relief map of the image could be created for the algorithm. After following this path, an array of the segmentation was obtained and shown in Figure 3.2.D.

The line segmentation was performed with a Viterbi algorithm. This algorithm is based on the probability density for different paths. White pixels indicate a larger height than black pixels, and the path follows the ridges. The algorithm determines the likelihood for every step and it finds the path from left to right. [40]

The parameter σ changes the size of the steps the algorithm takes. When σ is set high, the algorithm is not allowed to take jumps across datapoints. The parameter σ varied between 5 and 10, depending on the amount of noise present in the spectrum. If there seemed to be a more straight line, with less peaks, σ was chosen to be 5. The parameter σ was chosen to be 10, if the yellow line (Figure 3.2) in the screenshot had a path with a lot of noise present to segment.

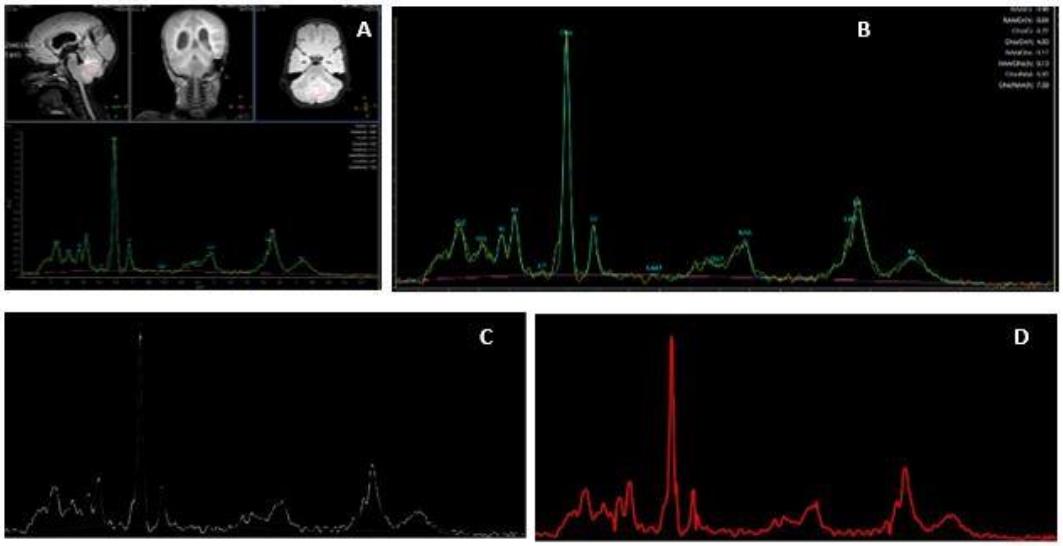

Figure 3.2: (A) The original screenshot from Case I. (B) The cropped image. (C) The yellow line is extracted and a grey value image is created, while other information is removed. (D) A Viterbi algorithm follows the path based on the probability density.

Table 3.2: The measured ratios of the chosen metabolites for the Peak/HeightD

Metabolites
Cre/Cho
NAA / Cho
mI / Cho
NAA / Cre
mI / Cre
mI / NAA
Lip + MM (0.9 ppm) / Cho
Lip + MM (1.3 ppm) / Cho
Lip + MM (0.9 ppm) / Lip + MM (1.3 ppm)

The second dataset from the screenshot data consists out of peak/height ratios. The following metabolites were included; Lip and MM (0.9 and 1.3 ppm), NAA (2.01 ppm), Cre (3.01 ppm), Cho (3.2 ppm) and mI (3.55 ppm) (Table 3). [10] The ratios were determined by an in-house developed MATLAB program where manual detection of the peak and baseline were needed. (MATLAB ver. R2019b (MathWorks; Natick, Massachusetts, USA)). This was done for all selected metabolites for every screenshot. With the measured length between the height of the peak and the baseline, the ratios were determined. An example can be seen in Figure 3.1.D, where the red lines simulate the measured distance between the peak height and baseline.

3.1.6 Statistical Analysis

SPSS v. 25.0 (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.) was used to perform statistical analysis. The correlation between the LineSD and Peak/HeightD was determined to observe a relation between an increase and decrease of metabolite concentration from both datasets. The interclass correlation coefficient (ICC) was determined to assess the test-retest reliability between measurements from different reviewers for determining peak/height ratios. The intraclass correlation coefficient (ICC) was determined to assess the test-retest reliability between measurements from the same examiner for determining peak/height ratios. Two different examiners determined the peak/height ratios for six different cases and one of the examiners repeated this operation six times for the same case. Both were determined based on a 95% confident interval, with absolute agreement. The interpretation of the results is as follows, results below 0.5 indicate a poor reliability, values between 0.5 and 0.75 indicate a moderate reliability, values between 0.75 and 0.9 show good reliability, results above 0.9 show an excellent reliability and 1.0 demonstrates perfect reliability. [41]

3.2 Results

3.2.1 Patient data

A total of 85 screenshots were included. Raw data was available for 45 of the screenshots. The dataset contains different tumor types, which are represented in Table 3.3.

Table 3.3: The number of cases for different tumor types.

Tumor type	Number of screenshots	Number raw data
Medulloblastoma	11	5
Ependymoma	6	3
Pylocytic astrocytoma	17	8
High grade glioma	8	3
Low grade glioma	22	11
DIPG	14	10
GCT	3	2
Pineoblastoma	2	2
ATRT	2	2
TOTAL	**85**	**46**

3.2.2 Data source 1: Raw data

Raw data (i.e. act.sdat) files could be used to generate the RawD. All items in the RawD consists out of 1024 complex numbers, and contain the full spectrum from -3.5 ppm to 12.5 ppm (Figure3.1.A). The residual water peak (at 4.7 ppm) remains visible when the spectrum is visualized.

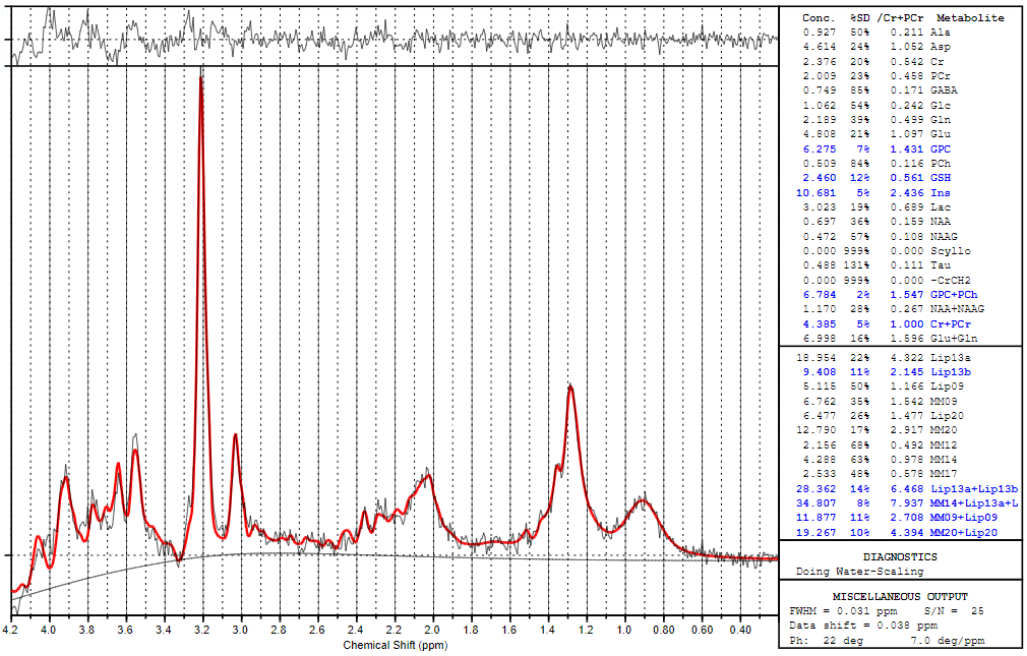

Figure 3.3: The total fit of LCModel of Case I. On the right the different amount of metabolite concentrations are seen.

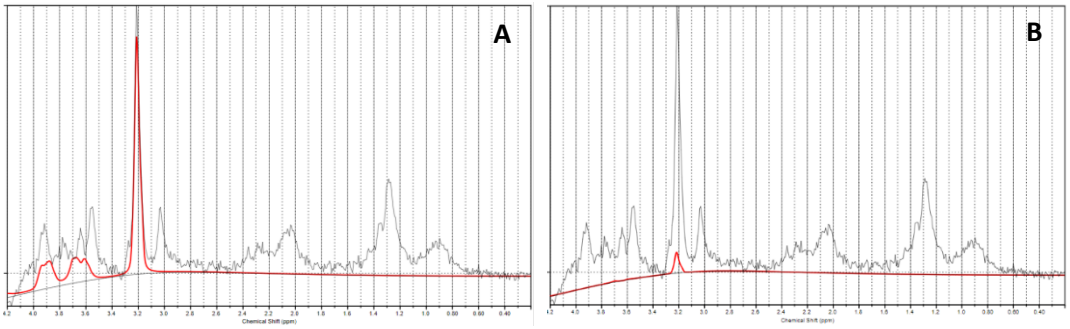

Figure 3.4: (A) The fit of GPC in the spectrum of Case I by LCModel. (B) The fit of PCh in the spectrum of Case I by LCModel.

Figure 3.3 and 3.4 show the different fits the LCModel program appoints to the peaks of the spectrum. The peak in Figure 3.3 at 3.2 ppm consists out of two metabolites, quantified with a concentration of 6.275 mM GPC and 0.509 mM PCh. After a summation of the two peaks the total choline concentration is 6.784 mM.

3.2.3 Data source 2: Screenshots

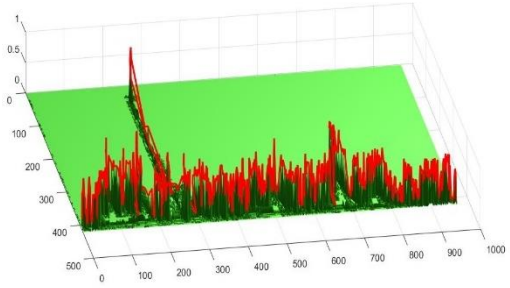

Figure 3.5: A relief map of the screenshot shows the height of the different datapoints. The algorithms follows the path which has the highest value.

Line segmentation was visually approved for all cases and contained 970 data points. Figure 3.5 and 3.6 show the segmentation in the screenshot. A relief image is seen in Figure 3.5, the line follows the path of the highest grayscale values. The full screenshot is shown in Figure 3.6, but during the segmentation the image was cropped as seen in Figure 3.5. Since values as ppm and concentration levels are not part of the segmentation, they disappear when the line is visualized.

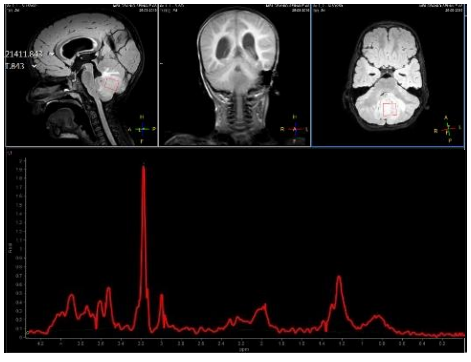

Figure 3.6: A segmented line, following the original yellow line.

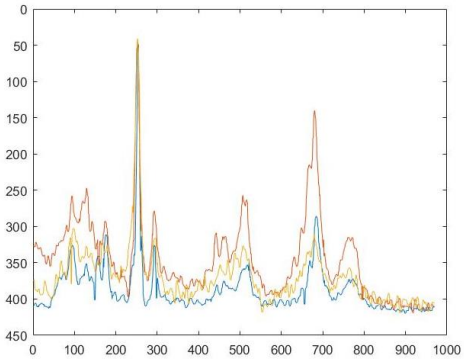

Figure 3.7: Segmented line from Case I (blue line) and two other medulloblastomas

Table 3.4: Example of results from a measurement of the peak/height ratios for case I.

Metabolites	Ratios
Cre/Cho	0.1997
NAA / Cho	0.1601
mI / Cho	0.2427
NAA / Cre	0.8018
mI / Cre	1.2160
mI / NAA	1.5160
Lip + MM (0.9 ppm) / Cho	0.0982
Lip + MM (1.3 ppm) / Cho	0.2995
Lip + MM (0.9 ppm) / Lip + MM (1.3 ppm)	0.3278

Figure 3.7 visualizes the plot of the segmented arrays for three different cases of medulloblastomas. As can be seen, the highest and lowest value overlap and the segmentations have the same length.

The peak/height ratios were determined and noted in Excel. An example of the results are shown in Table 3.4.

3.2.1 Statistical analysis

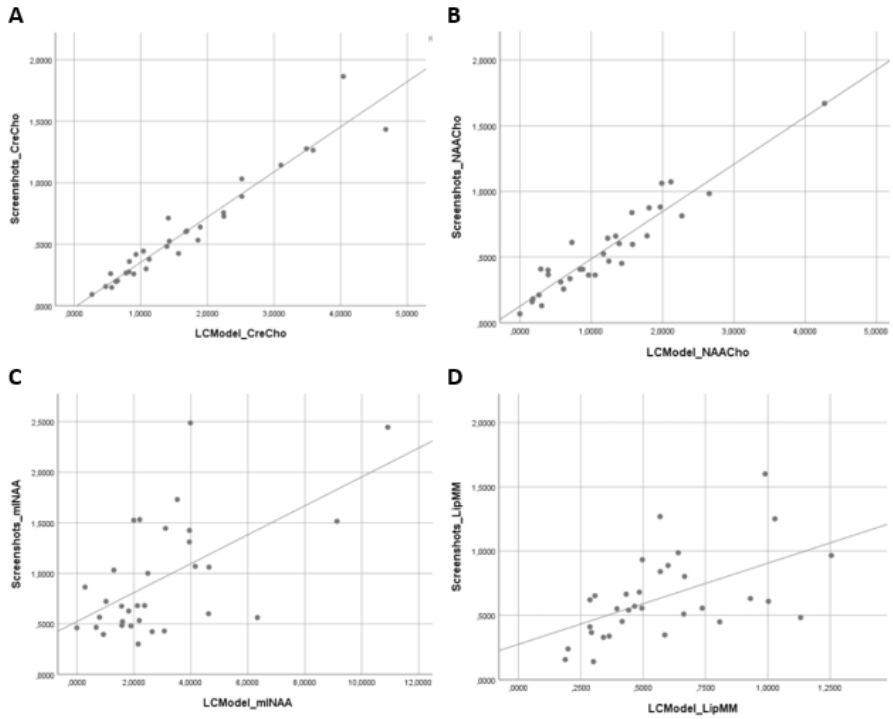

Figure 3.8: Correlation of different metabolite ratios between measurements from the screenshots (peak/height ratios (Peak/HeightD)) and LCModel (concentrations (ConcD)). (A) Correlation of Cre/Cho, (B) correlation of NAA/Cho, (C) mI/NAA, (D) Lip/MM.

Figure 3.8 visualizes the correlation of different metabolites for the Peak/HeightD and ConcD. As can be seen in Figure 3.8.A, the correlation between the Peak/HeightD and ConcD for Cre/Cho is 0.934. All of the Peak/HeightD Cre/Cho results remain below 2.0, while ConcD results show an increase up till 5.0. Also, within the results of other metabolite ratios a likewise difference can be seen. As shown in Table 3.5, most

correlations are higher than 0.5 and even higher than 0.7 when the ratios were divided over Cho. When observing the distribution of mI/NAA or Lip/MM, the correlation is poor.

Table 3.5: Metabolite with the correlations between Peak/HeightD and ConcD.

Metabolites	Correlation Peak/HeightD - ConcD
Cre / Cho	0.934
NAA / Cho	0.892
mI / Cho	0.543
NAA / Cre	0.667
mI / Cre	0.303
mI / NAA	0.345
Lip + MM (0.9 ppm) / Cho	0.749
Lip + MM (1.3 ppm) / Cho	0.860
Lip + MM (0.9 ppm) / Lip + MM (1.3 ppm)	0.295

Table 3.6 shows the inter- and intraclass correlation coefficient for determining peak/height ratios. Table 3.6.A shows the ICC for multiple tests performed by two raters. The result shows an ICC of 0.999 for single measures. Table 3.6.B shows the ICC for multiple tests performed by one rater. The outcome shows an ICC of 0.998 for single measures. Both these results indicate an excellent result.

Table 3.6: Inter- and intraclass correlation coefficient for the measurement tool to determine peak/height ratios. (A) The interclass correlation coefficient. (B) The intraclass correlation coefficient.

A

Interclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,999[a]	,999	1,000	3267,709	53	53	,000
Average Measures	1,000[c]	,999	1,000	3267,709	53	53	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

B

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,998[a]	,994	,999	2586,362	8	40	,000
Average Measures	1,000[c]	,999	1,000	2586,362	8	40	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

3.3 Discussion

In this chapter, data was acquired through two data sources and transformed into four different datasets. Each data source provided two datasets. Data obtained directly from the MR scanner, the raw data, resulted in a dataset with a vector of the full spectrum (RawD) and concentrations of the metabolites through fits by LCModel (ConcD). The data obtained using MR vendor software, are the screenshots. They resulted in a dataset with a vector of the yellow line/"raw spectral data" out of the image (LineSD) and peak/height ratios of metabolites present in the screenshot (Peak/HeightD).

3.3.1 Data source 1: Raw data

The strength of the RawD is the amount of features (1024), which provide a machine learning algorithm with a lot more freedom, compared to nine (Peak/HeightD) or twenty (ConcD) features. With so many features the algorithm has more freedom to identify distinctive features by feature selection and extraction. A lot of the data provided by the RawD might be indistinctive, but possible unknown features of MRS can be determined. The features from the Peak/HeightD or ConcD have been chosen based on literature, but this means the algorithm has less freedom in identifying distinctive features itself, since they already have been determined. This aspect of the raw data is important to take into account and it is exactly the reason why no post-processing has been performed. However, it also increases the risks for overfitting when implementing in a ML algorithm, especially when there are noisy features. Noise can be identified as high variance features and contribute to classification, hence the increased chance of overfitting.

The basissets used for quantification were pre-installed within the library of LCModel. As described in the method the 1.5T 38 ms TE basisset was provided to LCModel, while the MR parameters of the MRS data had a TE between 35 - 40 ms. As described in the method the 3T 35 ms TE basisset was provided to LCModel, while the MR parameters of the MRS data had a TE between 35 - 40 ms. Therefore, some data did not correspond to the simulated basisset. The program LCModel did not give an error for this mismatch, but a better match could be generated with a sequence specific basisset. This could improve accuracy of the quantification of metabolite concentrations.

For this research, different metabolites were selected to create different datasets for the ConcD. It remains unclear which dataset gives the lowest error rate. Brain tumors are known for the presence of uncommon increase of metabolites in healthy brain tissue, for instance Tau or Lac. These metabolites can be taken into account in the data collection of metabolite concentrations, but this information is not useful if the CRLB presents a very high uncertainty. Then the question remains what the algorithm will do with these features. Further exploration is done in Chapter 5, Data Analysis.

3.3.2 Data source 2: Screenshots

The segmented lines from the LineSD exists out of 970 datapoints. Since this line is segmented from an image, the datapoints contain real numbers. As can be seen in Figure 3.7, the lowest and highest datapoint for all three array plots overlay. The interest of this

study is focused on the ratios, therefore, no correction has been performed. This could be performed easily when correcting for the y-axis, but the concentrations depend on voxel size. Exact voxel sizes are only known when there is raw data. Since this is not present for all screenshots, this correction cannot be performed on the whole dataset.

The measurements of the peak/height ratios were performed in a research by Manias et al. [10] Within their research they invented a flowchart which was used to differentiate between three different tumor types. With this research the value of using screenshots with help of the peak/height ratio was shown. The same metabolites used in Manias et al. were used in the current research. The different ratios of metabolites were successfully determined with measurements in the screenshots, but only the metabolites named in paragraph 3.1.5.2 were measured. These metabolites are also fitted by software from Philips and result in a clear fit. Other metabolites are harder to indicate on the screenshots. As mentioned in the previous paragraph, Tau and Lac are big indicators of a disturbance in the metabolism of cells, especially in brain tumors. These metabolites are not included in the fit provided by the MRI scanner. This might become a problem when the algorithm shows a large contribution of these metabolites in differentiation tumor types in the other datasets. The fit, provided by the Philips software on the MRI host, could be altered in future research to pick up these metabolites and report them back to medical personnel.

3.3.3 Statistical analysis

The correlation between the Peak/HeightD and ConcD demonstrated a very good correlation between the Cre/Cho ratios from both datasets. As shown in Table 6, all ratios with choline result in a correlation above 0.5. Low correlation between ratios from both datasets can be explained by the shapes of some peaks. The linewidth and presence of peaks of mI, Lip and MM(0.9 and 1.3) can be very capricious. With a bigger linewidth, LCModel might account a different metabolite concentration than with peak/height ratios. This is because LCModel quantifies the area under the curve, but peak/height ratios only take the height of the peak into account. Height can be related to the concentration, but this is not the case when the peak is broadened. In general, Cho, Cre and NAA have a narrow peak, while mI, Lip and MM tend to have a more broadened peak.

As described in the results section, most of the Peak/HeightD results remain below 1.0. This causes a different interpretation from the ascertained presence of metabolites. LCModel is a well-known and widely used program to quantify metabolites and the Peak/HeightD ratios deviate from corresponding LCModel ratios. To improve the interpretation of the peak/height ratios, it would be useful to develop correction factors.

The ICC for the performance with multiple raters gave an excellent score. If an examiner reads the instructions, the test can be performed in a reliable manner. The ICC for the performance of one rater gave an excellent score as well. This means that if an examiner knows how to conduct the test, this can be performed in a reliable manner.

4. PHANTOM STUDY: CORRECTION FACTOR PEAK/HEIGHTD - CONCD

In the previous chapter, a correlation was found between the Peak/HeightD and the ConcD for multiple metabolites. However, interpretation of these results in a clinical context might be difficult. A ratio for Cre/Cho with LCModel resulted in many ratios above one. This means a higher presence of Cre than Cho in the sample. A lot of the ratios in Peak/HeightD remained below one. This means a lower presence of Cre than Cho in the sample. These results have a different meaning, since a ratio below one indicated a sample with a higher concentration of Cho than Cre. Therefore, a correction factor was determined to be able to compare results from peak/height ratios and LCModel with MR spectroscopy data from a phantom.

4.1 Method

As described in 3.4.4 Statistical analysis, the relationship between concentration ratios from ConcD and Peak/HeightD are different for each metabolite. This indicates that a different correction factor is needed for the different metabolites. Since the exact metabolite concentration in vivo is unknown, a phantom was used from which the concentrations are known. The Brain-O phantom was used to obtain scans with prior knowledge of these metabolite concentrations.

Three scans were obtained with the same sequence applied for routine scanning of pediatric brain tumors. The raw data and screenshots were both saved and two types of data (described in 3.1.B and 3.1.D) were derived. The average metabolite ratios from both data sources were compared. Dividing the average concentration through the result of the peak/height ratios provided correction factors for the different metabolites. With the determined correction factor, the peak/height ratios can be interpreted as the LCModel ratios.

4.2 Results

After determining the peak/height ratios from the screenshots and the ratios for the corresponding metabolites from LCModel, a correction factor was determined. The correction factors are shown in Table 4.1.

	Peak/Height ratios	LCModel ratios	Correction factor
Cre/Cho	1.14 (± 0.04)	2.76 (± 0.02)	2.41
NAA/Cho	1.66 (± 0.08)	3.70 (± 0.03)	2.23
mI/Cho	0.41 (± 0.05)	2.95 (± 0.09)	7.17
NAA/Cre	1.46 (± 0.01)	1.34 (± 0.02)	0.92
mI/Cre	0.36 (± 0.03)	1.07 (± 0.02)	2.97
mI/NAA	0.25 (± 0.02)	0.80 (± 0.03)	3.23
Lip/Cho	0.01 (± 0.00)	-	-
MM/Cho	0.18 (± 0.09)	-	-
Lip/MM	0.05 (± 0.03)	-	-

Table 4.1: Peak/height ratios and LCModel ratios of the phantom. The correction factors are shown in the right column to improve interpretation of Peak/Height ratio results.

4.3 Discussion

With the determined correction factor, the peak/height ratios can be interpreted as the LCModel results. These correction factors have better results when there is a good correlation between both datasets. When looking at Table 4.1, the correction factor for Cre/Cho is 2.41, and this works well for this correlation. For instance the blue dots in the orange circle in Figure 3.1.A, $1.3 * 2.41 = 3{,}133$. This comes close to 3.5. The correction factor for NAA/Cho is 2.23, which results in $1.7 * 2.23 = 3.8$. (Figure 3.1.B)

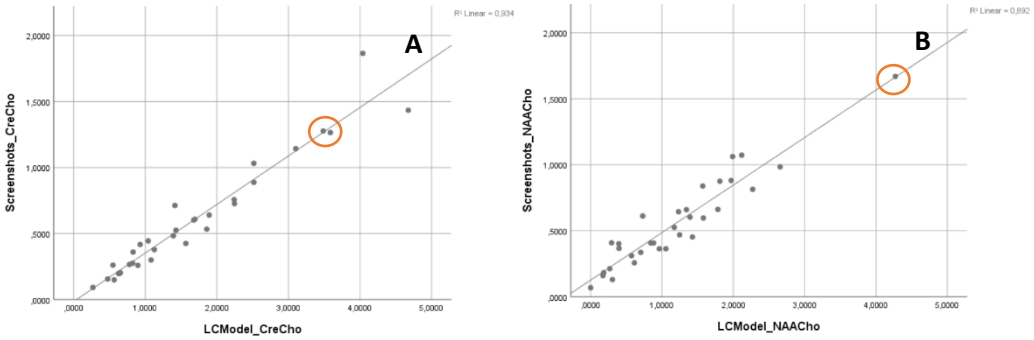

Figure 3.1: (A) Correlation of Cre/Cho ratios between Peak/HeightD (Screenshots) and ConcD (LCModel). (B) Correlation of NAA/Cho ratios between Peak/HeightD (Screenshots) and ConcD (LCModel).

Unfortunately not all metabolites were present in the phantom. Therefore, some correction factors could not be determined. Next to this, not all correction factors work as well for the ratios from the two datasets, this is because of a low correlation between the two datasets for these metabolites. An explanation for this difference in correlation might be the broadening of these peaks. But despite this flaw, it still helps with better interpretation of the results from peak/height ratios where a good correlation is present. The use of correction factors could be useful when looking at peak/height ratios, when raw data has not been saved, but there is still the need for interpretation of the concentrations. The missing correction factor could be calculated by obtaining data from a phantom in which all metabolites are present.

5. DATA ANALYSIS

5.1 Method

As mentioned in chapter 3 Data Acquirement, it is important to guarantee quality of data when working with machine learning. Therefore, a quality control was performed before creating the datasets. In the following chapter, the methods of the quality control and machine learning is described.

5.1.1 Quality control

After collecting the data, quality control of the data took place. Based on literature, a voxel size of 3 cm^3 was used as cut off value for this research for the raw data. Estimations of voxel size were made for screenshot data, since the selected voxel is visible on the screenshots. Spectra with small voxel sizes contained much noise. Poor water suppression resulted in an ascending baseline from 3.0 to 4.2 ppm. These spectra were rejected based on visual inspection. The amount of collected cases which passed the quality control are shown in Table 5.1. A total of 63 cases with screenshots, from which 34 cases with raw data was available.

The most samples were collected for medulloblastomas, DIPG and pilocytic astrocytomas. These three tumor types were labelled and used for classification in an algorithm.

Table 5.1: Total database after quality control.

Tumor type	Number of screenshots	Number raw data
Medulloblastoma	10	5
Ependymoma	6	3
Pilocytic astrocytoma	11	4
High grade glioma	6	3
Low grade glioma	13	6
DIPG	11	8
GCT	3	2
Pineoblastoma	1	1
ATRT	2	2
Total	**63**	**34**

5.1.2 Machine Learning

Down sampling was performed on RawD and LineSD, and $^1/_3$ of the data was maintained. To prevent overfitting, the most significant features were selected. Therefore, after down sampling, feature selection was used to filter the most useful features and implement them in the algorithm. Feature selection was applied to the RawD (20 features) and LineSD (30 features). After feature selection, feature extraction was applied with a PCA or LDA.

The LDA was used when N ≥ 10 for at least three classes, otherwise PCA was used. Since at least three classes of N ≥ 10 were needed, the data from screenshots could be used for this specific dimensionality reduction algorithm, because the three selected classes have ten or more samples. With LDA, the dimensions are reduced to classes –1. This indicated two dimensions for the use of three classes. When four classes are used, the dimensions are reduced to three. When using PCA, the amount of features for the maximum variance is determined. A PCA is applied to all four datasets, while an LDA is applied to the LineSD and Peak/HeightD.

The optimal hyperparameters for KNNC and SVM were determined on trial-and-error base with the use of cross validation. After dimension reduction, the dataset was implemented in LDC, QDC, KNNC, SVM and FFNN. Cross validation prevented undetected overfitting and provided a more reliable error rate. The performance of the classifiers was based on the error rates.

The error rates were evaluated for the three different datasets for ConcD, to determine which selection of metabolites would provide the best classification. The dataset with the lowest error rate was selected to be the final ConcD. The ML algorithms could not perform with complex numbers, therefore the magnitude for all datapoints from the RawD was determined.

Error rates for all datasets were calculated, and also distinctive components were made visible. An overview of these calculations can be seen in Figure 5.1. The classifiers LDC and QDC are not visualized in this figure, but an error rate was obtained for these classifiers as well. The algorithms mentioned above were implemented in MATLAB with the toolbox PRTools (MATLAB R2016b (MathWorks; Natick, Massachusetts, USA)). [42]

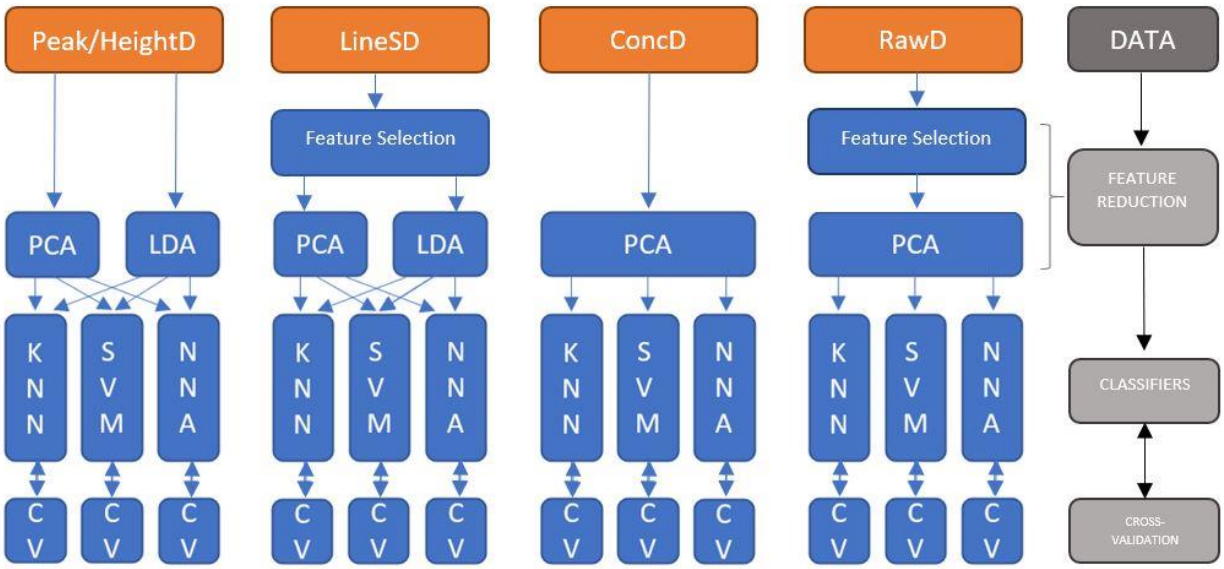

Figure 5.1: The paths that are taken by the different datasets are made visible within this flowchart. They all end with cross validation, which provides an error rate. Peak/HeightD and LineSD produce two error rates for every machine learning algorithm, because feature reduction in both datasets is performed by a PCA and an LDA.

Because of the small datasets, it is more difficult to obtain reliable error rates for the classifiers. The error rates might be too positive, because there was no separate test set. Despite the size of the datasets, five different training and test sets were created for the LineSD. 2 samples were taken randomly from every class. This resulted in five different training sets with 26 samples and five different test set with 6 samples. The same feature selection and LDA was performed for both datasets. After training the classifier, the datasets were combined and the classifier was tested.

5.2 Results

5.2.1 Cumulative variance

The cumulative variance was calculated to determine the amount of needed features. After a certain amount of features, the cumulative variance equals 1. From this point onwards, addition of other features does not improve variance. The amount of features was chosen as follows: RawD (12 features), ConcD (10 features), LineSD (30 features), Peak/HeightD (8 features). Plots can be found in Appendix A.

5.2.2 Feature selection

After determination of the cumulative variance, feature selection was performed. To compare the original dataset and the down sampled dataset, a feature selection of 30 features was compared. Figure 5.2.A shows which features were selected during the feature selection performed on the original LineSD, while Figure 5.2.B shows the selected features after feature selection performed on the down sampled LineSD.

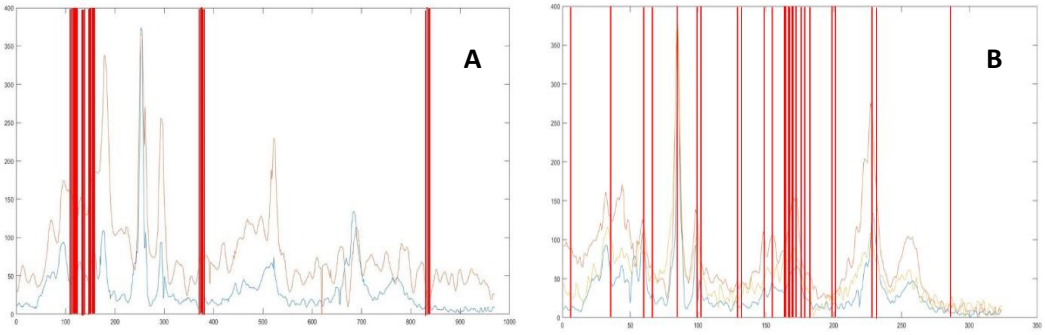

Figure 5.2: 30 selected features after feature selection. (A) The selected features in the original dataset, consisting out of 970 datapoints. (B) The selected features in the down sampled dataset, consisting out of 324 datapoints.

During the feature selection on the original LineSD, selected features remained in the same range. The selected features could be divided into five groups; 106 – 116, 126 – 137, 152 – 157, 381 – 386, 826 – 833. When looking at the ppm range, these features range from 3.7 – 3.8 ppm, 2.65 ppm and 0.65 ppm. Glutamate and glutamine peak around 3.7 – 3.8. Aspartate peaks around 2.65. The feature selection on the down sampled LineSD show a more distributed selection of features. The features selected in the down sampled

data are selected at more evenly distributed ppms. Some peaks of metabolites which are selected are Cho, Cre, NAA, mI, MM and Lip.

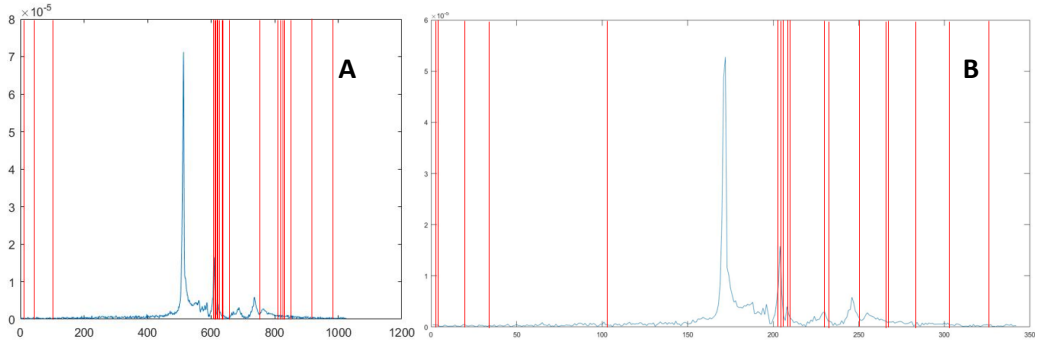

Figure 5.3: Feature selection on RawD before and after downsampling. (A) The feature selection algorithm selected 20 features, mostly between 600 and 900 (3.4 – 0.4 ppm). (B) After down sampling it was more distributed, but most of the features were still selected between 200 and 300 (3.4 – 0.4 ppm).

This was also performed for the RawD. The selected features were mostly distributed from 600 to 900, which correlate with a 0.4 – 3.4 ppm range (Figure 5.3.A). After down sampling, most of the features were still selected between this range, especially between 2.9 – 3.3 ppm, where Cre and Cho peak (Figure 5.3.B). A few features were selected around 2.0 ppm, where NAA peaks.

5.2.3 Feature extraction PCA

A PCA was performed to decrease the amount of data for all four datasets. The PCA determines distinctive components and these patterns help with the differentiation of data. The y-axis demonstrates the deviation of different data point for a component. The peaks indicate a deviation for this data point for the three different classes. The deviation of the first two components of the PCA are presented in Figure 5.4.

The RawD components show an increase of the Cho peak, while other peaks decrease in the spectrum. When inspecting the screenshots, Cho peaks tend to go up for all three tumor types. For the LineSD the area between NAA and Lip + MM (1.3 ppm) increases. The 1.3 ppm area remains very low for medulloblastomas after inspection of the screenshots. DIPGs show a higher peak at mI and Lip + MM (1.3 ppm), compared to the other two.

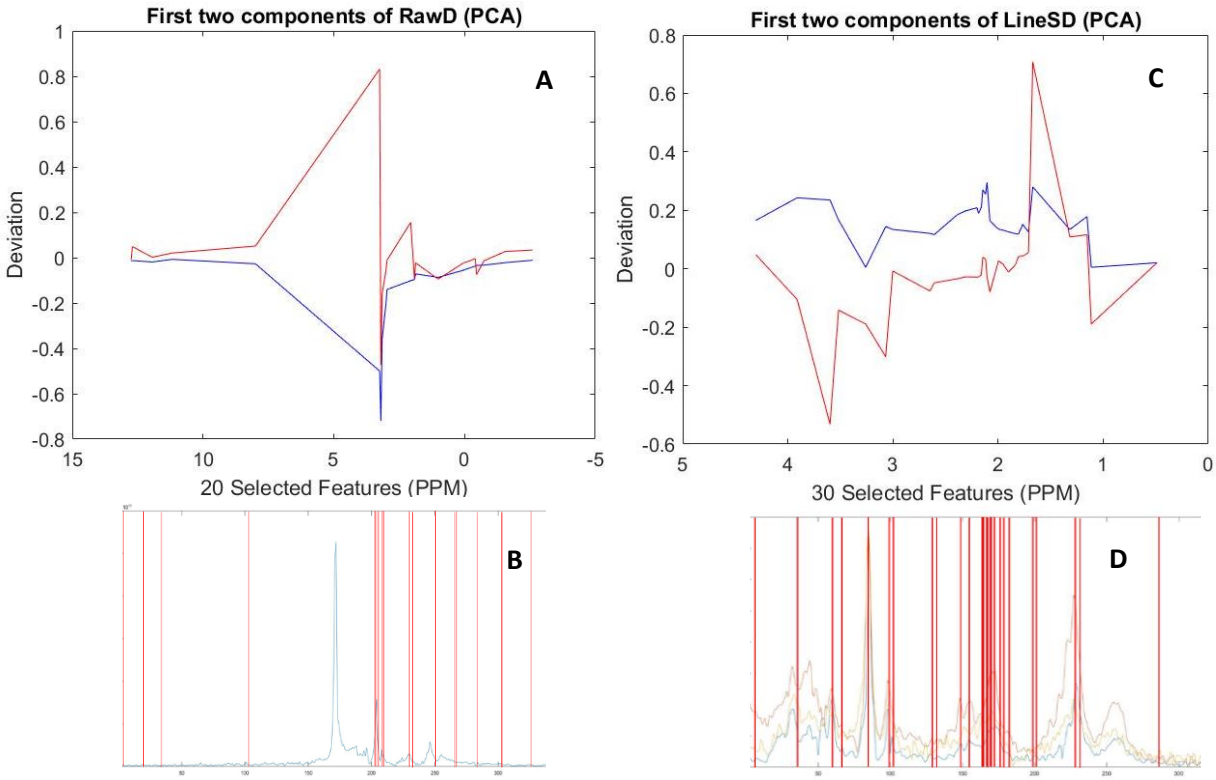

Figure 5.4: The deviation of the first two components of the RawD and LineSD after a PCA. (A) Deviation of RawD: The blue line shows a correlation between Cho and Cre decreasing. The red line shows a correlation between a part of the Cho peak increasing, while peaks decrease. The peak of Cre decreases, while the peaks around NAA increase. (B) Plot RawD with selected features. (C) The deviation of LineSD: The blue line shows a correlation between an increase of mI, increase of Cre, NAA and around 1.7 ppm. The red line shows a correlation between a decrease of most of the peaks in the spectrum (mI, Cho, Cre, NAA and Lip), but an increase around 1.7 ppm. (D) Plot LineSD with selected features.

Figure 5.5 shows the first two components for the ConcD. The data points for ConcD in Figure 5.5 are shown in Table 5.4. The LCModel concentrations show a much higher concentration for all the MM + Lip (0.9 ppm, 1.3 ppm, 2.0 ppm) for medulloblastomas. Deviation of these peaks is also seen in Figure 5.5. The concentrations also show a lower Cre concentration for pilocytic astrocytomas, as well as a lower Cho concentration compared to medulloblastomas. DIPGs show the lowest mI, GSH and Cho concentrations.

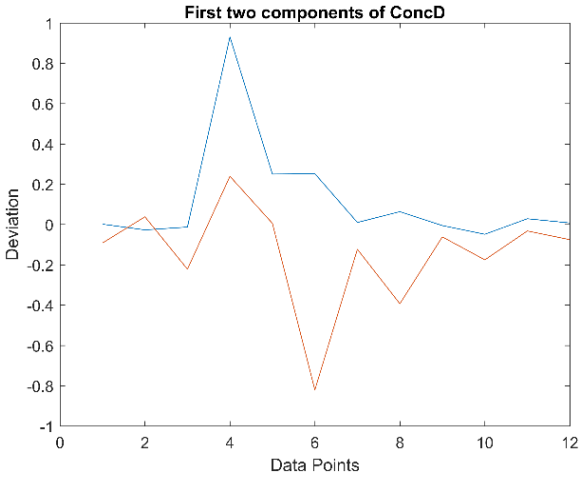

Figure 5.5: First two components of ConcD: The blue line shows a correlation between a higher concentration for MM and Lip at 0.9 ppm, 1.3 ppm and 2.0 ppm, while other metabolites show less deviation in this component. The orange line shows a correlation between a higher concentration for MM and Lip at 1.3 ppm, a lower concentration of MM and Lip at 2.0 ppm and a lower concentration for Cho.

Table 5.4: Data points with corresponding metabolites.

Data point	Metabolite
1	Lac
2	NAA + NAAG
3	Cr+PCr
4	MM13+Lip13a+b
5	MM09+Lip09
6	MM20+Lip20
7	Tau
8	GPC+PCh
9	GSH
10	mI+Scyllo
11	Glu+Gln
12	GABA

Figure 5.6 shows the first two components for the Peak/HeightD. The data points for Peak/HeightD in Figure 5.6 are shown in Table 5.5. The spectra show higher mI and a higher Lip + MM (1.3 ppm) peaks for DIPGs, compared to medulloblastomas and pilocytic astrocytomas. DIPGs have a higher value for mI/Cho than medulloblastomas or pilocytic astrocytomas after inspecting the ratios in the dataset. Medulloblastomas show lower values for NAA/Cho compared to the other two tumor types.

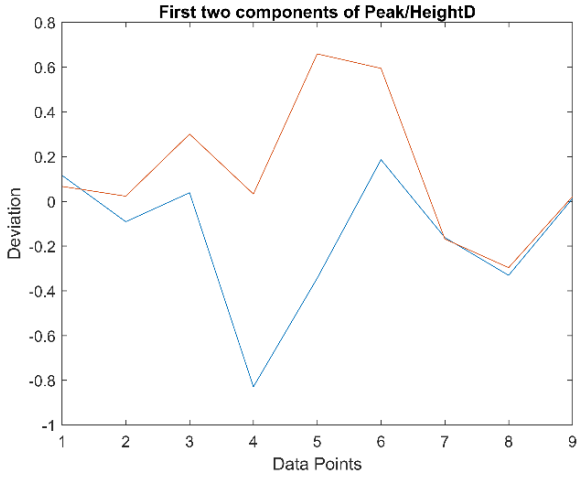

Figure 5.6: First two components of Peak/HeightD: The blue line shows a correlation between a lower NAA/Cre ratio, a higher mI/NAA and a lower Lip+MM(1.3 ppm)/Cho ratio. The orange line shows a correlation between a higher mI/Cre, a higher mI/NAA ratio and a lower Lip+MM(1.3 ppm)/Cho ratio.

Table 5.5: Data points with corresponding metabolites ratios.

Data point	Ratio
1	Cre/Cho
2	NAA / Cho
3	mI / Cho
4	NAA / Cre
5	mI / Cre
6	mI / NAA
7	Lip + MM (0.9 ppm) / Cho
8	Lip + MM (1.3 ppm) / Cho
9	Lip + MM (0.9 ppm) / Lip + MM (1.3 ppm)

5.2.4 Classifiers after PCA

The values for the hyperparameters k, r and C were chosen based on the lowest error rates determined with a cross-validation method. Plots can be found in Appendix B. The following classifiers were trained: LDC, QDC, KNNC, SVM and FFNN. Their performances were examined with cross-validation and based on the error rates.

The ConcD exists out of metabolite concentration based on prior knowledge. This dataset provided the lowest error rate. A table with all the ConcD error rates can be found in Appendix C.

The error rates for the datasets with the classifiers can be found in Table 5.6. The LineSD and RawD have comparable results with a KNNC.

Table 5.6: Error rates for all datasets after a PCA and cross-validation, for the three multiple classifiers.

	Peak/HeightD	LineSD	ConcD	RawD
LDC	0.4250 (± 0.0574)	0.4437 (± 0.0751)	0.2118 (± 0.1392)	0.3765 (± 0.1045)
QDC	0.4719 (± 0.0758)	0.4344 (± 0.0344)	0.5824 (± 0.0585)	0.3706 (± 0.0588)
KNNC	0.3937 (± 0.0264)	0.2031 (± 0.0593)	0.1412 (± 0.0304)	0.2000 (± 0.0186)
SVM	0.2562 (± 0.0506)	0.1656 (± 0.0331)	0.1294 (± 0.0372)	0.5118 (± 0.0186)
FFNN	0.3775 (± 0.0836)	0.3594 (± 0.0695)	0.4765 (± 0.0937)	0.4882 (± 0.0736)

5.2.5 Feature extraction LDA

There was enough data to perform an LDA to decrease the numbers of features for both screenshot datasets. The LDA determined two distinctive discriminants, which help with differentiation of data. The distribution of the data after application of an LDA is shown in Figure 5.7. A clear distribution is seen for the LineSD (Figure 5.7.A). The distribution for the Peak/HeightD is less clear distributed (Figure 5.7.B).

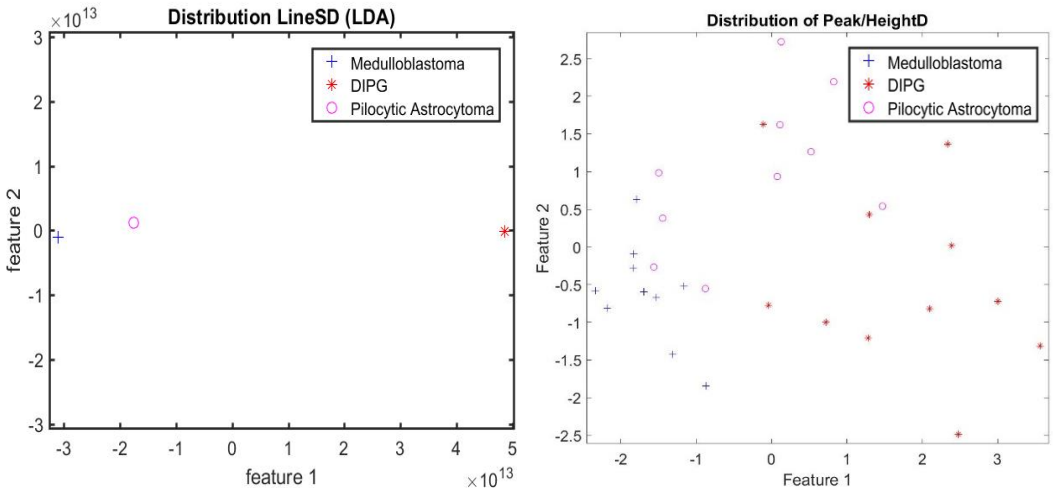

Figure 5.7: Distribution of two datasets. (A) The distribution of the LineSD after an LDA. (B) The distribution of the Peak/HeightD after an LDA.

Figure 5.8 shows the discriminants by the LDA that help with classification of the LineSD. The x-axis show the 30 selected features that were selected, plotted on the ppm scale, and the y-axis shows the deviation of these features. Figure 5.8.B shows the correlated features, visualized on the spectrum. There is not much deviation in the Cho peak (3.2 ppm). A lot of features were selected around 2.0 ppm. When observing the spectra, there is more deviation around 1.7 and 2.3 ppm for DIPGs and pilocytic astrocytomas, compared to medulloblastomas.

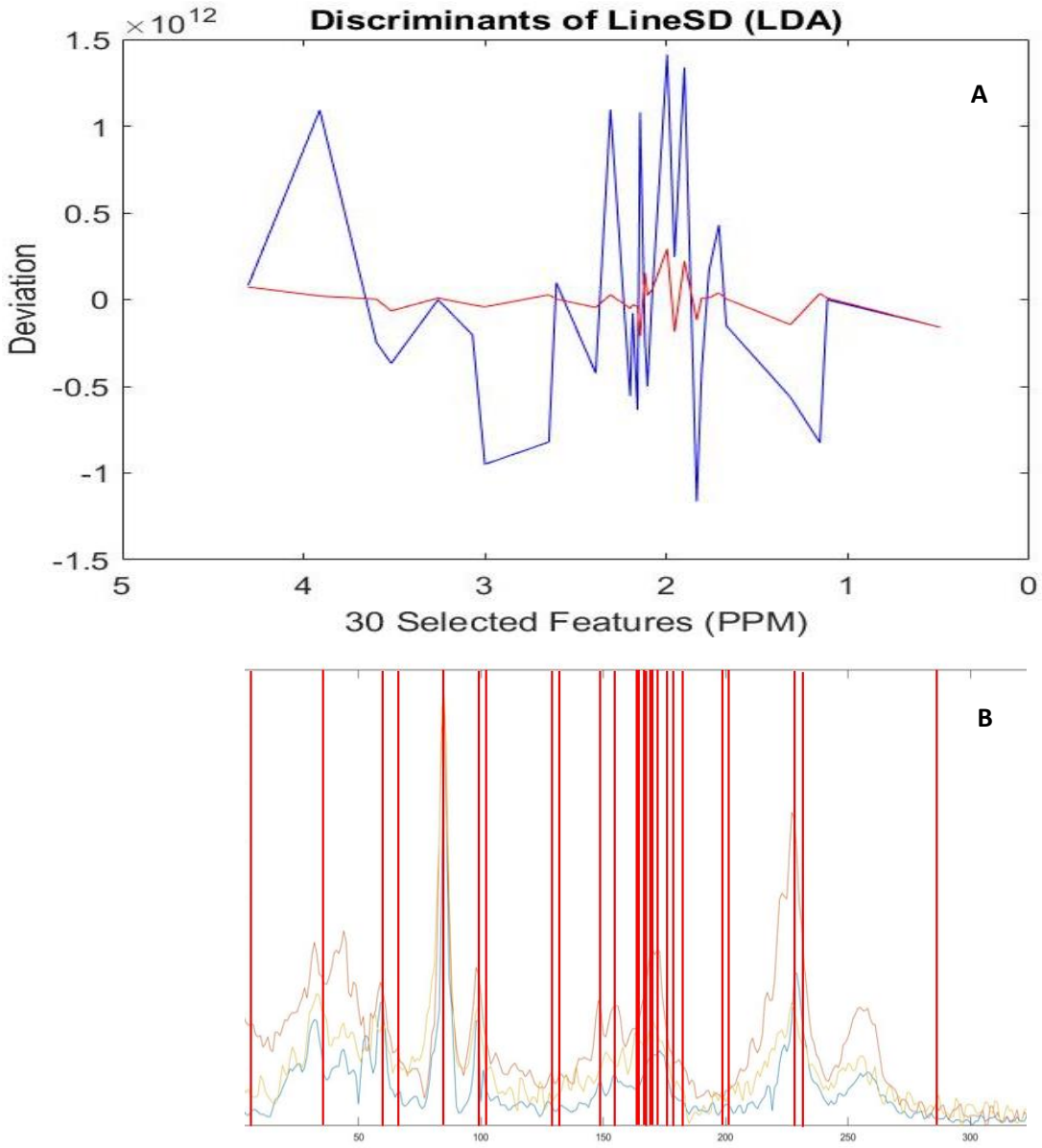

Figure 5.8: (A) The deviation between the two discriminants of LineSD: The blue line shows a correlation between an increase of the spectral line around 4 ppm, a decrease around 3.2 and 3.0 ppm (Cho and Cre), a decrease around 1.1 ppm and a lot of deviation around 2.0 ppm (NAA). The red line shows a correlation between a lot of deviation around 2.0 ppm (NAA), and a decrease around 1.2 ppm (MM + Lip). (B) Plot LineSD with selected features.

Figure 5.9 shows the first two components for the Peak/HeightD. The data points for Peak/HeightD in Figure 5.9 are shown in Table 5.7.

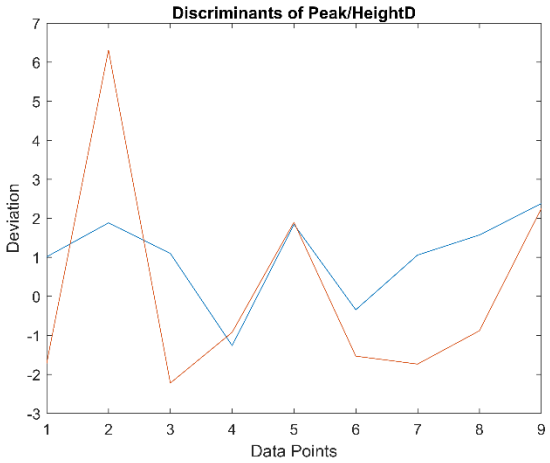

Figure 5.9: Discriminants of the Peak/HeightD after LDA: The blue line of the two discriminants of Peak/HeightD shows a correlation between a higher Cre/Cho ratio, a higher NAA/Cho and a lower Lip + MM (0.9 ppm/Lip + MM (1.3 ppm) ratio. The orange line shows a correlation between a higher NAA/Cho ratio, mI/Cre and a higher Lip + MM (0.9 ppm)/Lip + MM (1.3 ppm) ratio, while the other ratios have deviation below zero.

Table 5.7: Data points corresponding to the metabolite ratios.

Data point	Ratio
1	Cre/Cho
2	NAA / Cho
3	mI / Cho
4	NAA / Cre
5	mI / Cre
6	mI / NAA
7	Lip + MM (0.9 ppm) / Cho
8	Lip + MM (1.3 ppm) / Cho
9	Lip + MM (0.9 ppm) / Lip + MM (1.3 ppm)

5.2.6 Classifiers after LDA

All hyperparameters were optimized by examining different values for k, r and C. This was done with cross-validation (Appendix B). Figure 5.10 and 5.11 shows the distribution of the Peak/HeightD and different classifiers (LDC, QDC, KNNC, SVM and FFNN). The black lines show the final decision boundaries for the classifiers. The black lines of the SVM and FFNN classifier show very smooth lines, while the line of KNNC is less smooth.

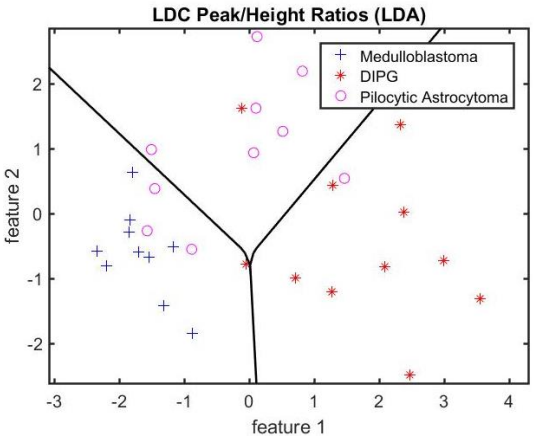

Figure 5.10: Classification of Peak/HeightD with a LDC.

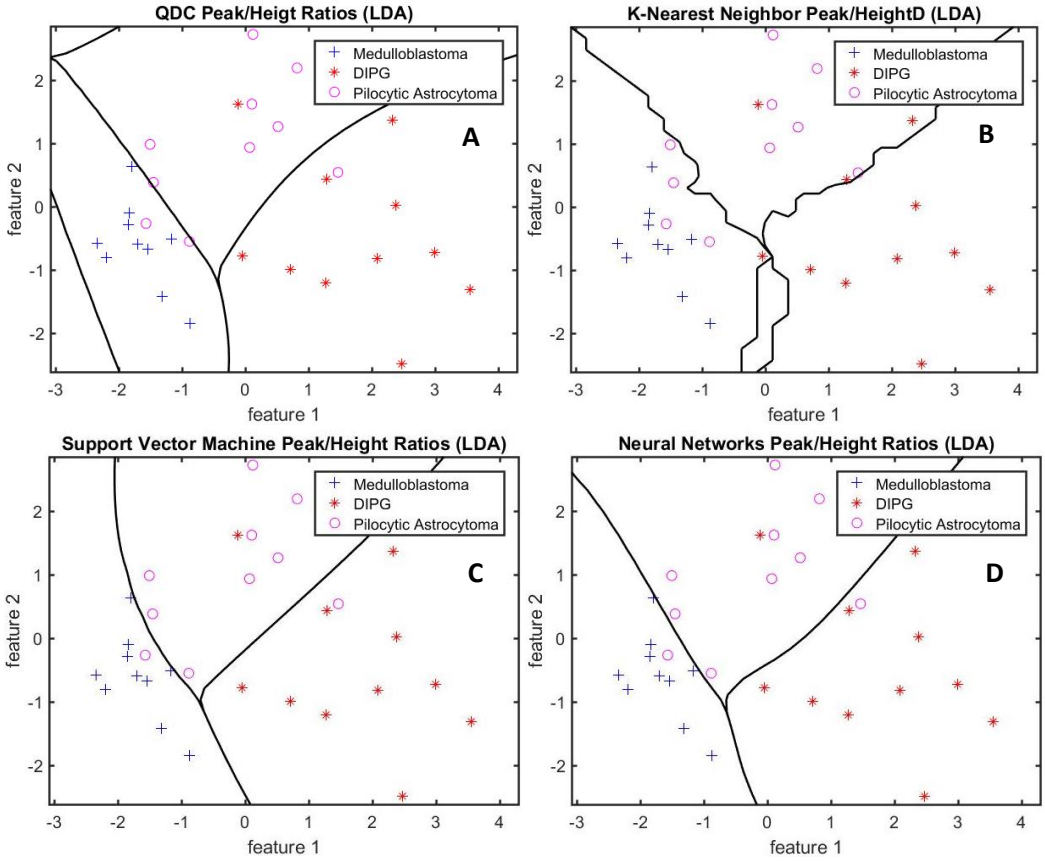

Figure 5.11: Distribution and classification of Peak/HeightD with different classifiers. (A) Classification of Peak/HeightD with a QDC. (B) Classification of Peak/HeightD with a KNNC. (C) Classification of Peak/HeightD with a SVM. (D) Classification of Peak/HeightD with an ANN.

Table 5.8 shows the error rates for both datasets, for the three different classifiers. The results for the Peak/HeightD after classification with feature reduction by an LDA are 10% lower compared to classification with feature reduction by a PCA. The KNNC shows an error rate of 0.1718. The error rate for the LineSD shows an error rate of 0.000 for the KNNC and SVM, while FFNN shows an error rate of 0.0181. The SVM shows a very high error rate compared to error rates provided by other classifiers.

Table 5.8: Error rates for the Peak/HeightD and LineSD after an LDA and cross-validation for multiple classifiers.

	Peak/HeightD	LineSD
LDC	0.2188 (± 0.0255)	0.000 (± 0.000)
QDC	0.2406 (± 0.0418)	0.000 (± 0.000)
KNNC	0.2375 (± 0.0124)	0.000 (± 0.000)
SVM	0.1844 (± 0.0231)	0.3000 (± 0.0264)
FFNN	0.2875 (± 0.0321)	0.0063 (± 0.0198)

Table 5.9: Error rates for the LineSD after an LDA for multiple classifiers with a training (26 samples) and test (6 samples) set.

	Test 1	Test 2	Test 3	Test 4	Test 5
LDC	0.1259	0.1259	0.1259	0.0944	0.1259
QDC	0.1259	0.1259	0.1259	0.0944	0.1259
KNNC	0.0944	0.0930	0.0622	0.0629	0.0944
SVM	0.3706	0.3077	0.3706	0.3706	0.3706
FFNN	0.0937	0.0622	0.0622	0.0937	0.0944

The error rates in Table 5.9 show the results for multiple tests with a different training and test sets. Results vary from 0.0944 to 0.0622 for the KNNC and FFNN, and from 0.0944 to 0.1259 for LDC and QDC. 2 out of the 6 samples were correctly classified for a classification with an error rate of 0.1259, 3 for an error rate varying from 0.0944 to 0.0930 and 4 for an error rate varying from 0.0629 to 0.0622. The SVM shows the same error rate as before. Medulloblastomas are not classified correctly by this classifier and it only provides labels for pilocytic astrocytomas or DIPGs. This happened for all training and testing data.

5.3 Discussion

A lot of different results were obtained from the acquired data. Different feature reduction techniques and multiple classifiers resulted in many options for a classification tool. With these results, and especially the amount of data, such a tool could not be developed yet. But results are promising, especially since error rates are comparable between the raw and screenshot data. This means the inclusion of screenshot data to replace raw data is possible. As mentioned earlier, it is important to guarantee the quality of data when implementing it in an ML algorithm. Therefore, quality control was performed first.

5.3.1 Quality Control

After quality control, 63 screenshots and 34 raw datasets remained. 22 screenshots were excluded due to the voxel size, or a remaining water signal. This is a loss of 26% of the data. MR spectra obtained with a small voxel size, resulted in spectra with a lot of noise and peak broadening. Maybe voxel sizes should be bigger when obtaining MRS data from small tumors, since it is believed that tumorous cells could have infiltrated in normal-appearing tissue. [43] This is thought to be the difference between the brain being the original site of the tumor and metastases. After communication with the University Hospital in Frankfurt about a collaboration, there was told that the whole hemisphere where the tumor is located is denoted as tumorous tissue, because of the possibility for infiltration. [44, 45] They use 2D CSI MR spectroscopy, where more voxels are selected. These ideas might make it possible to select a bigger voxel size for SVS as well. But before implementation of such a method, this should be tested first. Spectra from the voxels in the same hemisphere, located within the tumor (as seen on MR imaging) and voxels located within healthy tissue, should be compared. Because it might be possible to image tumor infiltration with MR spectroscopy, but there is also signal from metabolites from healthy brain tissue cells. The signal from healthy brain tissue, or even low grade parts of a tumor, might average the metabolite levels. [7]

During voxel placement, peripheral fat (skull, scalp), or water signal from the ventricles should be avoided, because of water/lipid signal which will distort the MRS signal. Avoiding these areas is especially difficult when a smaller tumor is located at the demarcation of brain tissue.

Next to this, voxel placement is highly user dependent. Therefore, it remains important to train medical personnel for obtainment of spectra and how to judge spectral data. There is also the possibility of automatic recognition of artefacts in a spectra after obtainment with help of another ML algorithm. This could help medical staff to judge the quality of the spectra directly after acquiring and if needed, they are able to obtain a new MR spectra. [46]

5.3.2 Feature Extraction

RawD

A lot of features which were selected during feature selection of the RawD were in the range between 600 and 900. After down sampling, the range changed from 600 – 900 to 200 – 300. Most of the features ranged between 0.6 – 3.4 ppm, this included peaks from Cho, Cre and NAA and confirms the ppm range where peaks are used for clinical interpretation. Also 1.1 ppm was selected, and this feature lies between Lip + MM 0.9 ppm and Lip + MM 1.3 ppm. The deviation of this area shows a small decrease in both components (Figure 5.4.A). It remains unclear why this feature is selected in the raw data, as it remains unclear why the other seven from the twenty features were selected outside of the 0.6 – 3.4 ppm range. But when observing the plot of the components, a very small deviation is noticeable at the selected features outside of this range, while the features inside the range cause for much higher deviation.

LineSD

When feature selection took place for the LineSD, features were selected in a close range, around 3.7 – 3.8 ppm, 2.65 ppm and 0.65 ppm. These results could not be linked to metabolites used in other datasets of this analysis, although these ppm values are very similar to the ones where Gln and Asp peak. These metabolites are significant features for pilocytic astrocytomas in a study by Andronesi et al. [47]

After down sampling, feature selection was repeated. The new features were more spread over the spectrum and highlighted peaks used in other datasets (mI, Cho, Cre, NAA, Lip and MM). The earlier selected features could be noisy features which cause a lot of variance to the spectrum, or they could be features to distinguish pilocytic astrocytomas. But the error rate improved when a down sampled dataset was used. Therefore, it is supposed that the features did not necessarily distinguish pilocytic astrocytomas, but were most likely noise. Next to this, since more data samples also add computational time, it is better to use the less data samples.

The area between NAA and Lip + MM (1.3 ppm) increases, when looking at the two components (PCA). This might be a distinguishment between medulloblastomas and the other two tumor types, since this area remains very low when inspecting the screenshots of medulloblastomas. DIPGs show a higher peak at mI and Lip + MM (1.3 ppm), compared to the other two. Therefore, the decrease at this ppm might indicate a component that is helpful to classify DIPGs.

The two deviations (LDA) show the Cho peak remaining around zero (Figure 5.8). This can be explained because the Cho peak is almost always the highest point in the spectrum and this causes it to be the almost always the same value. The decrease of mI, Cre and Lip + MM (1.3 ppm) can contribute to the differentiation of DIPGs from the other tumors. Figure 5.8 also shows a lot of deviation between 1.7 and 2.3 ppm. The lowest value of the plots is seen for medulloblastomas, but there is also a question in most spectra about the

presence of Glu/Gln. The signal looks a bit noisy sometimes, or have effect of peak broadening of the NAA peak.

ConcD

Both components, shown in Figure 5.5, provide the classifier information for classification. The blue line shows a connection between an increase of all MM + Lip. Only medulloblastomas have a much higher concentration of all these three MM + Lip, and this line helps differentiating the classes based on this knowledge. Also the orange line of the first two components helps differentiating medulloblastomas. The orange line shows a decrease of Cr, and while the deviation of Lip + MM at 1.3 ppm goes up, the deviation of Lip + MM at 2.0 ppm decreases a lot, together with Cho. This is what differentiates medulloblastomas from DIPGs and pilocytic astrocytomas when observing the ConcD. DIPGs have the lowest concentrations of MM + Lip at 1.3 according to the ConcD. Also mI and GSH and Cho show the lowest results for DIPGs, but that is not particularly shown in these two components.

The possible role of Gln and Asp was suggested earlier, but when observing data point 10 Glx + Gln), there is nearly no deviation seen. Although this was suggested to be a feature helping to classify pilocytic astrocytomas. [47] Another suggested feature was the presence of Tau, since this is also seen as an indicator for medulloblastomas. A small decrease is seen, but a more reliable result should be obtained with a bigger data collection. [7]

Peak/HeightD

A lower NAA/Cre and Lip + MM (1.3 ppm)/Cho ratio with a higher mI/NAA ratio are expected when observing an MR spectra from tumorous tissue. NAA decreases and Cho increases for all three tumor types, while mI mostly increases for DIPGs. When looking at the screenshots, there is also an increase in Lip + MM (1.3 ppm) for DIPGs. This can also be seen as a decrease for medulloblastomas and pilocytic astrocytomas in comparison to DIPGs.

The orange line (Figure 5.6) shows a correlation between a higher mI/Cre, a higher mI/NAA ratio and a lower Lip + MM(1.3 ppm)/Cho ratio. But since these components derived from a PCA, the results are not specific for one class.

The LDA reduces the dataset to two dimensions and helps with supervised classification of the data. As decribed in the result section, the blue line of the two discriminants of Peak/HeightD in Figure 5.6 showed a correlation between a higher Cre/Cho ratio, a higher NAA/Cho and a lower Lip + MM (0.9 ppm/Lip + MM (1.3 ppm) ratio. The highest Cre peaks are seen with pilocytic astrocytomas, which also explains there decrease of NAA/Cre ratio. Compared to DIPGs, pilocytic astrocytomas and medulloblastomas have lower results for mI/NAA, caused by the higher mI peaks in the DIPG screenshots. The orange line (Figure 5.6) showed a correlation between a higher NAA/Cho ratio, mI/Cre and a higher Lip + MM (0.9 ppm)/Lip + MM (1.3 ppm) ratio, while the other ratios have deviation below zero.

Overall, DIPGs have a highest value for mI/Cho, and medulloblastomas show lower values for NAA/Cho compared to the other two tumor types. So, the algorithm is able to use these features to classify the data and distinguish these three tumor types.

5.3.3 Classifiers

The lowest error rates after application of a PCA for the datasets were 0.2000 for RawD, 0.2031 for LineSD, 0.1274 for ConcD and 0.2562 for Peak/HeightD. The ConcD had the lowest error rate after a PCA of all four datasets. Since LDA is a preferable feature reduction method, it is expected that after more data acquiring, the error rate could become even lower after application of an LDA over ConcD. A KNNC resulted in the lowest error rate for the RawD and LineSD. Since the recreated raw data resulted in the same error rate as the original raw data, screenshots can be used to recreate this. SVM resulted in the lowest error rate for ConcD and Peak/HeightD.

The lowest error rates after application of a LDA for the two datasets were 0.00% for LineSD and 0.22% for Peak/HeightD. Since the distribution of the LineSD was so clear, only the classification lines for the Peak/HeightD were plotted. The error rate for the QDC was worse, compared to other error rates. When looking at the lines of the classifier, a different gaussian function might have improved the result. The error rate of the LDC resulted in the lowest error rate for the Peak/HeigtD, which indicates that a more robust classifier performs well for this classification problem as well. This algorithm accounts for less computational time, but remains more simple than the other classifiers. For now, there might not be enough data to determine which classifier performs best for a classification tool.

The error rate for the LineSD turned out to be very low for most of the classifiers. Even though a cross-validation method was applied, it was expected that this method would not function as well with new data. Therefore a new experiment was run, where five different training set and test set were obtained from the dataset. The experiment was run five times. The results show very low error rates, but when examining the wrongly classified labels, an error rate of 33%, 50% and 66% is seen for data samples from the test set. KNNC and FFNN show the most promising results, but further training should be performed after more data acquisition.

The SVM did not seem to function on this dataset, since it only divided the set into two classes (pilocytic astrocytomas and DIPGs). The feature space might be too big for this type of classifier. Many different hyperparameters were given as input, but none of them seemed to improve the error rate.

6. Discussion

6.1 Overall discussion

From the raw data and screenshots, four different datasets were derived. After quality control, 63 screenshots and 34 raw datasets were collected. Training sets with medulloblastomas, DIPGs and pilocytic astrocytomas were acquired from the four datasets.

When observing the error rates for the different classifiers after a PCA, the results for the LineSD and RawD were almost the same. The results show the possibility of including screenshots and recreating raw data with this data source. This might be of great value for a lot of clinical centers, since a lot of raw data has never been saved. This method is used on screenshots from a Philips MR system. When using screenshots from different vendors, the new data should be examined closely. It is important to test the new data and determine the error rates for the classification of the combined data from different vendors.

The performance of the LineSD even improved after performing an LDA. Since this resulted in an error rate of 0.000, a training and test set were obtained and they showed a different result. 33% to 50% of the test set was incorrectly classified with a KNNC or FFNN algorithm. These results were higher with a LDC, QDC and SVM. Simpler classifiers, such as an LDA or QDC are less sufficient to perform the classification of this dataset. More freedom of the KNNC and FFNN helped with the classification, but distribution free classifiers have a higher risk of overfitting. Therefore, cross validation and different training and test sets should be continuously used to test the classifier. It would be of great value to perform this type of testing for all datasets, but the dataset should be enlarged and might need more variance of tumor types to become a functional tool in the clinic. A standard method of classification has not been developed yet. Classifiers that are frequently used to classify MR spectra from pediatric brain tumor in other studies are KNNC, SVM and ANN. [13, 31, 48, 49] In this research SVM performed very well for most datasets, and this is also noted during more studies. [31, 48]

The three chosen classes were chosen because of the practical consideration aiming for the largest amount of data. From clinical point of view, these type of tumors can be distinguished based on other factors as well. DIPG tumors are high grade tumors, located in the pons. They can be distinguished solely on MR imaging already. A lot of studies show the possibility of differentiating high and low grade tumors, but the benefit of a classification tool would be the possibility of differentiating difficult tumors or even subtypes within tumor types. From a clinical point of view, differentiation between GCTs, ATRTs, plexus tumors is difficult and an ML algorithm capable of doing this would be of added value. But more data is needed to train a classifier for this type of multi-classification. Also differentiation of subtypes would be helpful. One study showed the

possibility of differentiating subtypes from medulloblastomas, but the dataset used in this study did not contain enough data to train a classifier. [14]

To improve and build a tool, procedures for data acquisition and data analysis should be automated. For now, it was important to determine the value of screenshots and compare many different data types. Automation of data acquisition and quality control would be of great help in the future, as would automation for the data analysis be.

6.2 Recommendations and future perspectives

The ultimate goal of this research would be to build a tool able to distinguish different tumor types or even histological subtypes within tumor types. [14] There are over a hundred different histological subtypes. Therefore, it is necessary to start developing a tool with the ability to differentiate tumor types. [50] To do so, other features, such as age, gender and location of lesion are needed to be implemented as well. Other features that might be interesting are family history, cancer related syndromes and nationality, but also other imaging techniques. Values of Apparent Diffusion Coefficient (ADC) from Diffusion Weighted Imaging (DWI) demonstrate higher outcome when severity of tumor grades increase. [50, 51] MR perfusion and the degree of contrast enhancement in the tumor can contribute in classification of tumors. [47]

In retrospect the collected data was not enough to provide the algorithm enough data to distinguish many different tumor types. Sample sizes of other studies vary between 30 to 90 patients when trying to classify pediatric brain tumors. These researches try to differentiate ependymomas, pilocytic astrocytomas and medulloblastomas. [30, 31, 48] In many of them there seems to be a disbalance between the number of cases for different tumor types, but there are numerous ways to correct for the unbalanced distribution. [48] But when trying to build a robust tool, more data is needed to classify the amount of tumor types daily observed in the PMC. Next to collaboration with other facilities, there are multiple ideas to enlarge the dataset and they will be discussed next.

Single Voxel VS Multiple Voxel

The data that has been collected at the PMC contains solely single voxel spectra (SVS). When looking at other research sites, there is also a great interest in the spectra of 2D or 3D chemical shift imaging (CSI). With these techniques, spectra from multiple voxels can be obtained, and information about the spatial distribution of metabolites can be provided. This technique can be especially useful when imaging heterogeneous lesions. However, CSI remains technically challenging, it is more difficult to perform automatic quantification of metabolites and compared to SVS, it is a more time consuming technique. [52] Still a lot of sites are enthusiastic about this technique, since it can provide more information per time unit. Therefore, it is important to search for methods to use information from CSI as well as information from SVS in an algorithm. There might be possibilities to take an average spectrum of the multivoxel spectra, or they could be

implemented separately. In this last manner, more data would be generated, but especially with heterogeneous lesions it is important to know which voxel is studied.

Imaging of organoids or biopsies

Scanning biopsies or organoids can be another example of collecting data. Since patient-derived organoids have become important to determine drug response, this could also be used as data to be obtained for a ML tool. More insight is gained in producing multicellular tumor spheroids (MTS). These 3D models exhibit characteristics from in vivo tissue, and are thus more realistic in vitro models. Unfortunately, research also shows the difference between in vivo and in vitro metabolism. More improvement is needed for organoids to gain full potential and mimic the organs from which they have been derived. [53, 54] Less literature can be found on MR spectroscopy on brain tumor biopsies, but Andronesi et al. showed a method from which ex vivo data can be used. 2 mg tissue was needed to perform HRMAS 1H MR spectroscopy. After performing the MR spectroscopy, the tissue remained available for subsequent analysis. They were able to perform tumor classification based on 16 metabolites. [47]

Both the use of organoids and biopsies give the possibility to investigate therapy response with help of MR spectroscopy. Possible benefits of MR spectroscopy should be extended to have a purpose after a noninvasive diagnosis. With this manner, more insight can be gained on the effect of therapy during treatment. [47] MR spectroscopy is capable of differentiating between radiation necrosis and tumor recurrence. [55] Whereas multiple obtainment of biopsies during treatment to map therapy response is too invasive and comes with great risks, MR spectroscopy could be implemented during routine scanning. However, it is important to determine the changes in tissue under more controlled circumstances and to gain knowledge on the processes on lower scale.

7. Conclusion

In conclusion, this study demonstrates that screenshots can be used as a replacement for raw data. Classification is possible for all four datasets. Low error rates were obtained with the LineSD and the ConcD after a PCA. The error rates for the Peak/HeigtD and LineSD further reduced after an LDA. KNNC and SVM showed promising results to classify the data, but with this amount of data all classifiers are more prone to overfitting.

No unknown features were noticed in the four datasets, but the datasets provided more insight in ML algorithms to differentiate pediatric brain tumors. The data included three very different tumor types, but this research should be repeated with tumor types which are more difficult to classify. Although the datasets were not big enough to build a reliable and accessible tool to implement in a clinical setting, it showed the possibility of MRS in combination with machine learning to be a useful technique to differentiate pediatric brain tumors in a non-invasive manner, and could become a diagnostic classification tool for clinical practice.

APPENDIX A: Feature Reduction

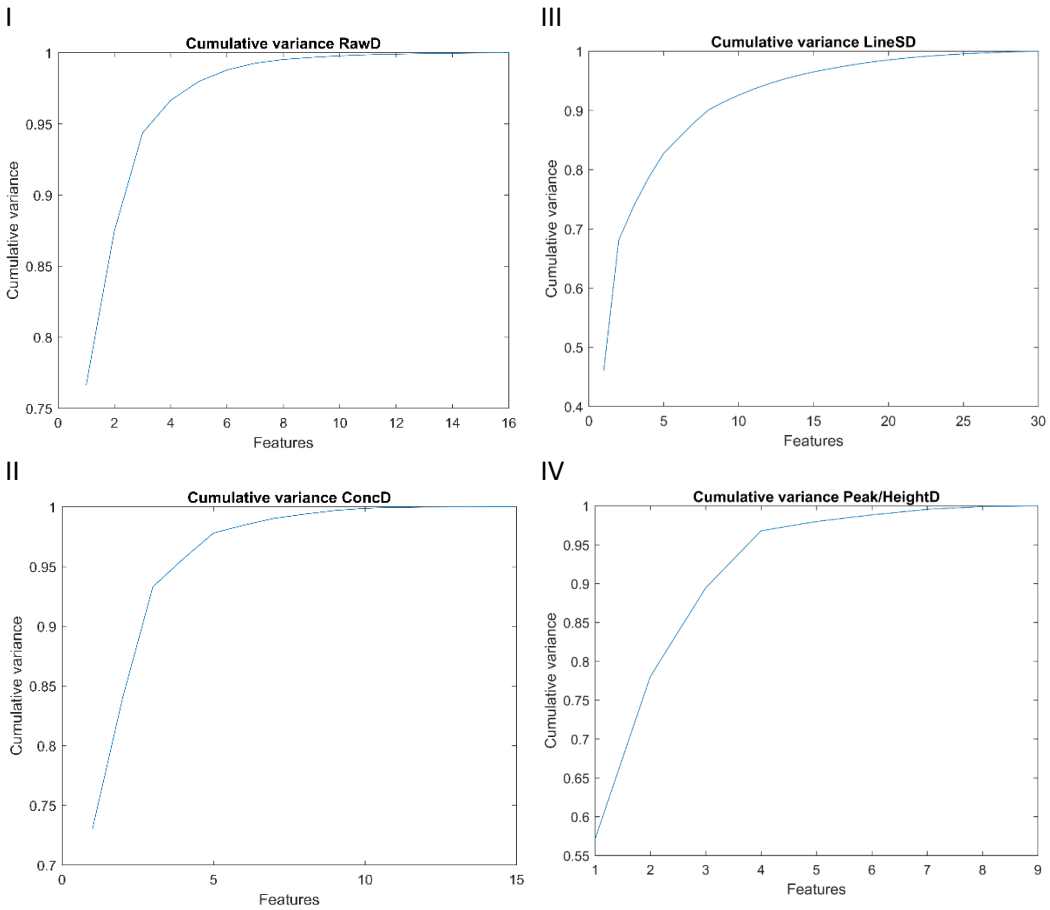

Figure A-1: Cumulative variance for all datasets as preparation to perform a PCA. I. RawD. II. ConcD. III. LineSD. IV. Peak/HeightD.

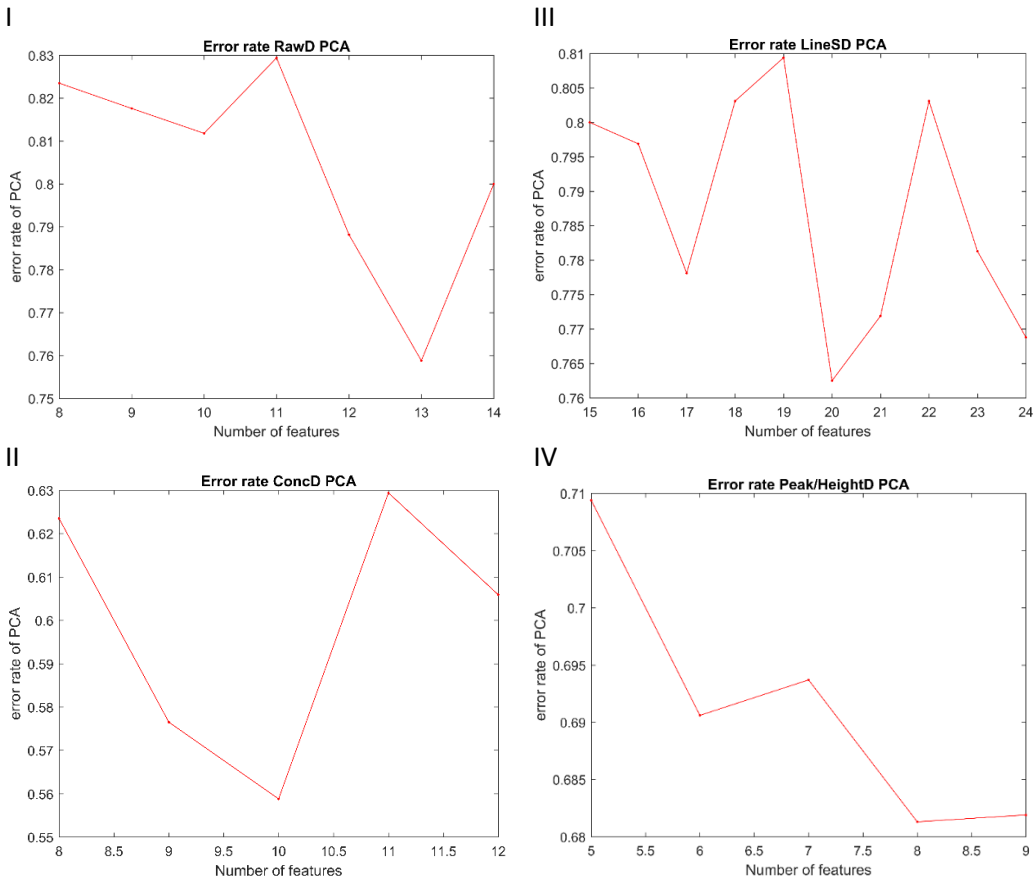

Figure A-2: The error rates of different numbers for feature reduction with PCA, determined with cross-validation. I. RawD. II. ConcD. III. LineSD. IV. Peak/HeightD.

Tablee A-1: The error rates for the two feature reduction methods.

	LineSD Error	STDS	Peak/HeightD Error	STDS
PCA	0.7620	0.0357	0.6815	0.0515
LDA	0.6875	0.0467	0.7688	0.0628
	ConcD Error	STDS	RawD Error	STDS
PCA	0.5765	0.0464	0.7412	0.0928
LDA	NA	NA	NVT	NA

APPENDIX B: Hyperparameters

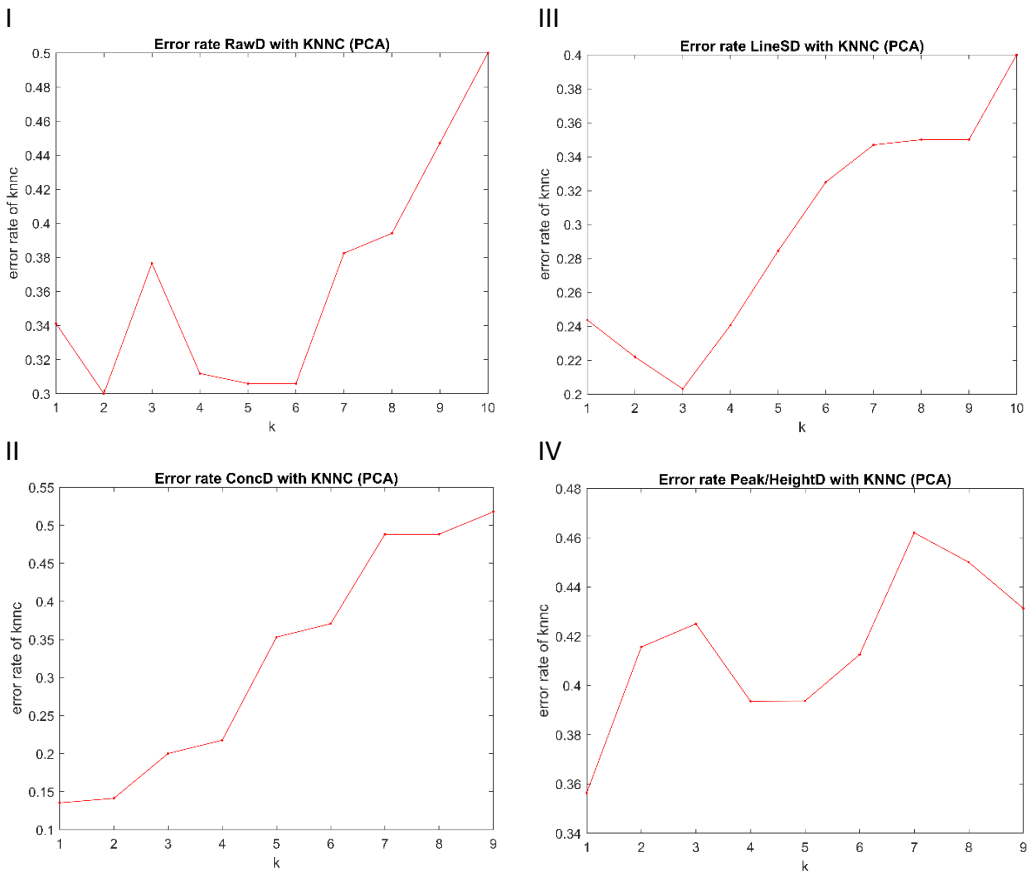

Figure B-1: Error rate of datasets for different k, performed with cross validation, after a PCA. I. Error rates for the RawD. II. Error rates for the ConcD. III. Error rates for the LineSD. IV. Error rates the for Peak/HeightD.

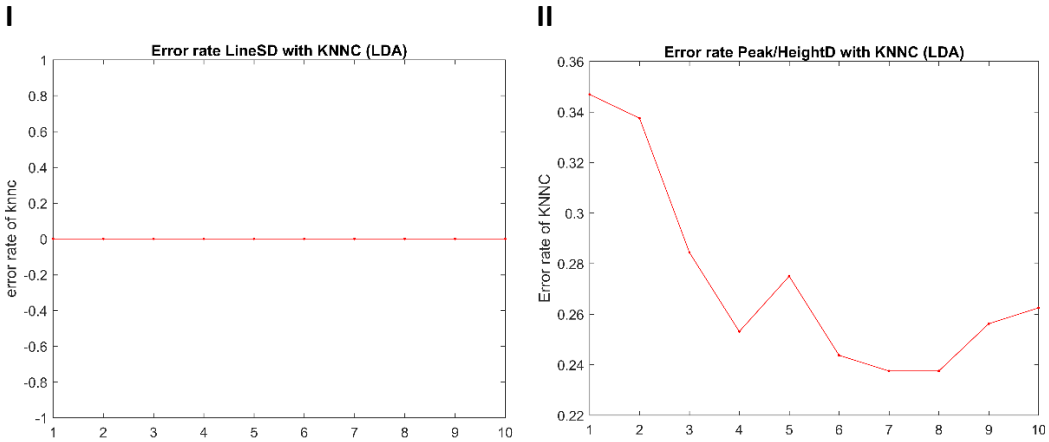

Figure B-2: Error rate of datasets for different k, performed with cross validation after an LDA. A. Error rates for the I. Error rates for the LineSD. II. Error rates the for Peak/HeightD.

C was chosen to be 0.1, 1, 10, 100, 1000, 1500 2000, 4000. The best error rates with different values for r are plotted in Figure B-3 and B-4.

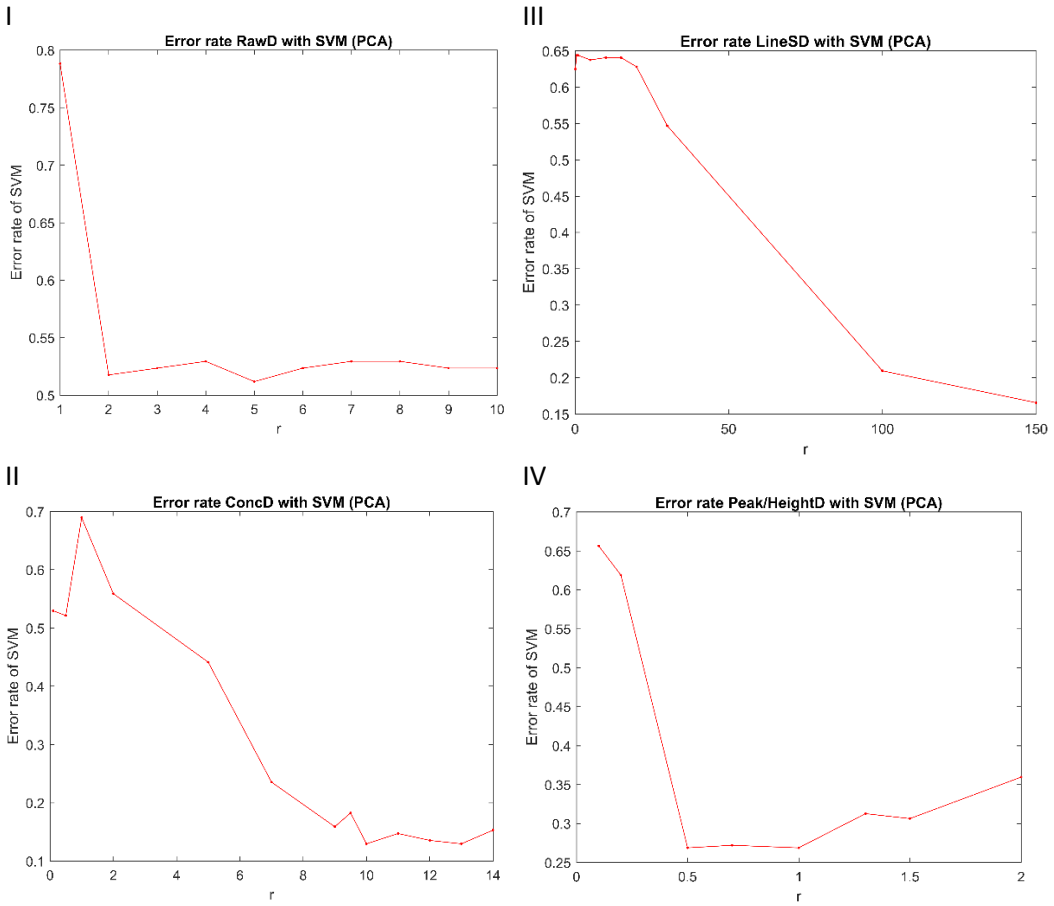

Figure B-3: Error rate of datasets for different r, performed with cross validation, after a PCA. I. Error rates for the RawD. II. Error rates for the ConcD. III. Error rates for the LineSD. IV. Error rates the for Peak/HeightD.

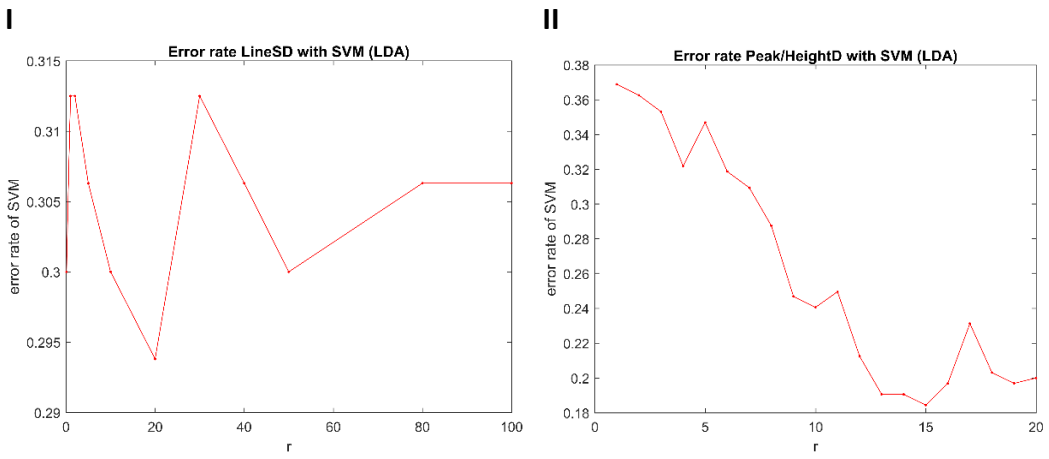

Figure B-4: Error rate of datasets for different r, performed with cross validation after an LDA. A. Error rates for the I. Error rates for the LineSD. II. Error rates the for Peak/HeightD.

APPENDIX C: Datasets ConcD

Table C-1: Error rates of the three differenct datasets for the ConcD, performed with cross validation, after a PCA.

	Data CRLB < 50	All LCModel Conc.	Chosen Conc
KNNC	0.2381	0.1905	0.1952
SVM	0.3381	0.2952	0.2476
ANN	0.4000	0.4571	0.3476

References

1. Byer, L., C. Kline, and S. Mueller, *Clinical trials in pediatric neuro-oncology: what is missing and how we can improve.* CNS Oncol, 2016. **5**(4): p. 233-9.
2. Fischer, C., et al., *Improving Care in Pediatric Neuro-oncology Patients: An Overview of the Unique Needs of Children With Brain Tumors.* J Child Neurol, 2016. **31**(4): p. 488-505.
3. Suh, E., et al., *Late mortality and chronic health conditions in long-term survivors of early-adolescent and young adult cancers: a retrospective cohort analysis from the Childhood Cancer Survivor Study.* The Lancet Oncology, 2020. **21**(3): p. 421-435.
4. Armstrong, G.T., et al., *Late mortality among 5-year survivors of childhood cancer: a summary from the Childhood Cancer Survivor Study.* Journal of Clinical Oncology, 2009. **27**(14): p. 2328.
5. Jagt- van Kampen, C.T., M.D. van de Wetering, and A.Y.N. Schouten–van Meeteren, *The timing, duration, and management of symptoms of children with an incurable brain tumor: a retrospective study of the palliative phase.* Neuro-Oncology Practice, 2015. **2**(2): p. 70-77.
6. Karajannis, M., J.C. Allen, and E.W. Newcomb, *Treatment of pediatric brain tumors.* J Cell Physiol, 2008. **217**(3): p. 584-9.
7. Lequin, M. and J. Hendrikse, *Advanced MR imaging in pediatric brain tumors, clinical applications.* Neuroimaging Clinics, 2017. **27**(1): p. 167-190.
8. Majós, C., et al., *Brain Tumor Classification by Proton MR Spectroscopy: Comparison of Diagnostic Accuracy at Short and Long TE.* American Journal of Neuroradiology, 2004. **25**(10): p. 1696-1704.
9. Naser, R.K.A., et al., *Role of magnetic resonance spectroscopy in grading of primary brain tumors.* The Egyptian Journal of Radiology and Nuclear Medicine, 2016. **47**(2): p. 577-584.
10. Manias, K.A., et al., *Prospective multicentre evaluation and refinement of an analysis tool for magnetic resonance spectroscopy of childhood cerebellar tumours.* Pediatr Radiol, 2018. **48**(11): p. 1630-1641.
11. Vieira, B.H., A.C.d. Santos, and C.E.G. Salmon, *Pattern recognition of abscesses and brain tumors through MR spectroscopy: Comparison of experimental conditions and radiological findings.* Research on Biomedical Engineering, 2017. **33**: p. 185-194.
12. Lundervold, A.S. and A. Lundervold, *An overview of deep learning in medical imaging focusing on MRI.* Zeitschrift für Medizinische Physik, 2019. **29**(2): p. 102-127.
13. Tandel, G.S., et al., *A Review on a Deep Learning Perspective in Brain Cancer Classification.* Cancers (Basel), 2019. **11**(1).
14. Bluml, S., et al., *Molecular subgroups of medulloblastoma identification using noninvasive magnetic resonance spectroscopy.* Neuro Oncol, 2016. **18**(1): p. 126-31.
15. Durmo, F., et al., *Multivoxel 1H-MR Spectroscopy Biometrics for Preoprerative Differentiation Between Brain Tumors.* Tomography, 2018. **4**(4): p. 172.

16. van der Graaf, M., *In vivo magnetic resonance spectroscopy: basic methodology and clinical applications.* European Biophysics Journal, 2010. **39**(4): p. 527-540.
17. Alger, J.R., *Quantitative proton magnetic resonance spectroscopy and spectroscopic imaging of the brain: a didactic review.* Topics in magnetic resonance imaging: TMRI, 2010. **21**(2): p. 115.
18. Lin, A. *Making the 'virtual biopsy' a reality with MR spectroscopy*. 2018; Available from: https://www.dotmed.com/news/story/44982/#pid=1.
19. De Graaf, R.A., *In vivo NMR spectroscopy: principles and techniques*. 2019: John Wiley & Sons.
20. Ahuja, S. and S. Scypinski, *Handbook of modern pharmaceutical analysis*. Vol. 10. 2010: Academic press.
21. Zeinali-Rafsanjani, B., et al., *Mrs Shimming: an important point which should not be ignored.* Journal of biomedical physics & engineering, 2018. **8**(3): p. 261.
22. MRIQuestions. *PRESS*. 2020; Available from: http://mriquestions.com/press.html.
23. Lin, A., et al., *Efficacy of proton magnetic resonance spectroscopy in neurological diagnosis and neurotherapeutic decision making.* NeuroRx, 2005. **2**(2): p. 197-214.
24. Castillo, M., L. Kwock, and S.K. Mukherji, *Clinical applications of proton MR spectroscopy.* American journal of neuroradiology, 1996. **17**(1): p. 1-15.
25. Seymour, Z.A., et al., *Citrate in Pediatric CNS Tumors?* American Journal of Neuroradiology, 2008. **29**(5): p. 1006-1011.
26. Choi, C., et al., *In vivo detection of citrate in brain tumors by 1H magnetic resonance spectroscopy at 3T.* Magn Reson Med, 2014. **72**(2): p. 316-23.
27. El-Habr, E.A., et al., *A driver role for GABA metabolism in controlling stem and proliferative cell state through GHB production in glioma.* Acta neuropathologica, 2017. **133**(4): p. 645-660.
28. Zhou, W., L.A. Liotta, and E.F. Petricoin, *The Warburg Effect and Mass Spectrometry-based Proteomic Analysis.* Cancer Genomics Proteomics, 2017. **14**(4): p. 211-218.
29. Nakamura, H., et al., *The Significance of Lactate and Lipid Peaks for Predicting Primary Neuroepithelial Tumor Grade with Proton MR Spectroscopy.* Magn Reson Med Sci, 2018. **17**(3): p. 238-243.
30. Davies, N., et al., *Identification and characterisation of childhood cerebellar tumours by in vivo proton MRS.* NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo, 2008. **21**(8): p. 908-918.
31. Zarinabad, N., et al., *Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy.* Magnetic resonance in medicine, 2017. **77**(6): p. 2114-2124.
32. Hira, Z.M. and D.F. Gillies, *A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.* Advances in Bioinformatics, 2015. **2015**: p. 198363.
33. Sharma, A. and K.K. Paliwal, *Linear discriminant analysis for the small sample size problem: an overview.* International Journal of Machine Learning and Cybernetics, 2015. **6**(3): p. 443-454.
34. Cady, E., *Clinical Magnetic Resonance Spectroscopy*. 1990.

35. Belkic, D., *Quantum-mechanical signal processing and spectral analysis*. 2019: CRC Press.
36. Near, J., *FID-A: The FID Appliance, Version 1.2, User Manual*. 2018.
37. Provencher, S.W., *Automatic quantitation of localized in vivo 1H spectra with LCModel.* NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo, 2001. **14**(4): p. 260-264.
38. Kreis, R., *The trouble with quality filtering based on relative C ramér-R ao lower bounds.* Magnetic resonance in medicine, 2016. **75**(1): p. 15-18.
39. Abele, T., et al., *Diagnostic accuracy of screening MR imaging using unenhanced axial CISS and coronal T2WI for detection of small internal auditory canal lesions.* American Journal of Neuroradiology, 2014. **35**(12): p. 2366-2370.
40. Van der Heijden, F., *Viterbi Algorithm Hand Out*. 2019: Enschede.
41. Koo, T.K. and M.Y. Li, *A guideline of selecting and reporting intraclass correlation coefficients for reliability research.* Journal of chiropractic medicine, 2016. **15**(2): p. 155-163.
42. Pekalska, E.a.D., R.P.W. *Pattern Recognition Tools*. Understanding, Teaching, Services; Available from: http://37steps.com/37-steps/.
43. Huang, R.Y. and A. Lin, *Whole-Brain MR Spectroscopy Imaging of Brain Tumor Metabolites*. 2020, Radiological Society of North America.
44. Polomac, N., *Universitätsklinikum Frankfurt: Collaboration UMCU - PMC, Sharing MR Spectroscopy Data*, M.F. Brasz, Editor. 2020: Frankfurt.
45. Zhang, J., et al., *Metabolic approach for tumor delineation in glioma surgery: 3D MR spectroscopy image–guided resection.* Journal of neurosurgery, 2016. **124**(6): p. 1585-1593.
46. Menze, B.H., et al., *Mimicking the human expert: pattern recognition for an automated assessment of data quality in MR spectroscopic images.* Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2008. **59**(6): p. 1457-1466.
47. Andronesi, O.C., et al., *Molecular classification of brain tumor biopsies using solid-state magic angle spinning proton magnetic resonance spectroscopy and robust classifiers.* International journal of oncology, 2008. **33**(5): p. 1017-1025.
48. Zarinabad, N., et al., *Application of pattern recognition techniques for classification of pediatric brain tumors by in vivo 3T 1H-MR spectroscopy—A multi-center study.* Magnetic resonance in medicine, 2018. **79**(4): p. 2359-2366.
49. Hsu, C.-W. and C.-J. Lin, *A comparison of methods for multiclass support vector machines.* IEEE transactions on Neural Networks, 2002. **13**(2): p. 415-425.
50. Johnson, K.J., et al., *Childhood brain tumor epidemiology: a brain tumor epidemiology consortium review.* Cancer Epidemiology and Prevention Biomarkers, 2014. **23**(12): p. 2716-2736.
51. Aggarwal, A., et al., *Role of multivoxel intermediate TE 2D CSI MR spectroscopy and 2D echoplanar diffusion imaging in grading of primary glial brain tumours.* Journal of clinical and diagnostic research: JCDR, 2017. **11**(6): p. TC05.
52. Gupta, R.K., Kumar, S., *Magnetic Resonance Imaging of Neurological Diseases in Tropics*. 2014.

53. Borja, M.J., et al., *Conventional and advanced MRI features of pediatric intracranial tumors: supratentorial tumors.* American Journal of Roentgenology, 2013. **200**(5): p. W483-W503.
54. Jamieson, L.E., D.J. Harrison, and C.J. Campbell, *Raman spectroscopy investigation of biochemical changes in tumor spheroids with aging and after treatment with staurosporine.* Journal of biophotonics, 2019. **12**(5): p. e201800201.
55. Weygand, J., et al., *Can an Organoid Recapitulate the Metabolome of its Parent Tissue.* A Pilot NMR Spectroscopy Study. J Cancer Prev Curr Res, 2017(7): p. 00307.