TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Artificial intelligence unveils tumor diversity in brain cancer through Raman spectroscopy

Machine Learning for Glioma Subtype Classifications

Joel Edvard Christian Sjöberg

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 753 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2026

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ARTIFICIAL INTELLIGENCE UNVEILS TUMOR DIVERSITY IN BRAIN CANCER THROUGH RAMAN SPECTROSCOPY

Machine Learning for Glioma Subtype Classifications

Joel Edvard Christian Sjöberg

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. AI OSA TOM. 753 | ASTRONOMICA CHEMICA PHYSICA MATHEMATICA | TURKU 2026

University of Turku

Faculty of Science
Department of Mathematics and Statistics
Mathematics
Doctoral Programme in Exact Sciences (EXACTUS)

Supervised by

Prof. Ion Petre
University of Turku, Finland

Prof. Vesa Halava
University of Turku, Finland

Reviewed by

Prof. Nicolae Leopold
Babeș-Bolyai University, Romania

Assoc. Prof. Isaac O. Afara
University of Eastern Finland

Opponent

Prof. Adrian Iftene
Alexandru Ioan Cuza University of Iaşi, Romania

ISBN 978-952-02-0549-2 (PRINT)
ISBN 978-952-02-0550-8 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)
Painosalama, Turku, Finland 2026

UNIVERSITY OF TURKU
Faculty of Science
Department of Mathematics and Statistics
Mathematics
SJÖBERG, JOEL: Artificial Intelligence Unveils Tumor Diversity in Brain Cancer through Raman Spectroscopy
Doctoral dissertation, 109 pp.
Doctoral Programme in Exact Sciences (EXACTUS)
March 2026

ABSTRACT

Applying machine learning (ML) methods as diagnostic classification models can accelerate the process of diagnosing cancers. For glioma, a brain cancer possessing diverse genetic makeup, ML can capture the different genetic characteristics in tumor environments, providing a diverse and precise mapping of heterogeneous gliomas. The need for methods capable of computing these kinds of predictions with high reliability is of special interest when considering the rapid deterioration of health in glioma patients. A promising avenue for developing these models can be found through Raman spectroscopy, a vibrational spectroscopic technique capable of capturing the genetic traits of gliomas through tumor-wide scanning. ML can be utilized to curate Raman spectra, a necessary procedure for quality assurance of Raman spectroscopy datasets. In larger datasets formed through combinations of different cohorts, there is a considerable risk of the batch effect occurring. The batch effect is descriptive of the bias present within datasets which results from assumptions and methodologies carried out during their extraction. Curating Raman data from the batch effect is important to ensure model reliance on cancer specific patterns rather than acquisition-related effects.

In this thesis, we present mathematical models capable of forming predictions for tumor-wide classifications of genetic characteristics. We develop methods for curating Raman spectra through ML and synthetic data generation and demonstrate their effectiveness on a dataset of glioma tumors. Furthermore, we develop a method to improve dataset quality through ML for removal of the batch effect, promoting model detection of cancer-specific patterns. Our contributions to the field of glioma classification comes in the form of classifier models and the strategies which have enabled them. We present a deep learning (DL) architecture which utilizes a minimal number of parameters to provide consistent outputs for correction of spectra. We also present a strategy to reduce the batch effect through adversarial learning while measuring the features relevant for genetic classifications. Through this work, we show how applying our methods can improve the performance of classification models for gliomas.

KEYWORDS: Machine Learning, Glioma, Raman Spectroscopy, Feature Extraction, Deep Learning

TURUN YLIOPISTO
Natural Sciences
Mathematics And Statistics
Pure Mathematics
SJÖBERG, JOEL: Artificial Intelligence Unveils Tumor Diversity in Brain Cancer through Raman Spectroscopy
Väitöskirja, 109 s.
Exactus
Maaliskuu 2026

TIIVISTELMÄ

Koneoppimismenetelmiä, voidaan soveltaa diagnostisissa luokittelumalleissa nopeuttamaan syövän diagnosointiprosessia. Geneettisesti monimuotoisen aivosyövän, eli glioman kohdalla koneoppimisella voidaan tunnistaa kasvaimen eri geneettisiä ominaisuuksia. Koneoppiminen tarjoaa monipuolisen ja tarkan kuvan heterogeenisistä glioomista. Otettaessa huomioon glioomapotilaiden terveyden nopea heikkeneminen on tarpeellista käyttää menetelmiä, joilla voidaan tehdä luotettavia arvioita koneoppimisella. Lupaava keino näiden menetelmien kehittämiseen löytyy värähtelyä mittaavasta spektroskopiatekniikasta, eli Raman-spektroskopiasta, jolla voidaan havaita glioomien geneettiset piirteet skannaamalla koko kasvain. Koneoppimista voidaan käyttää Raman-spektrien kuratointiin, joka on välttämätöntä Raman-spektroskopian datajoukkojen laadunvarmistuksessa. Suuremmissa tietokannoissa eri kohorteista muodostuvissa yhdistelmissä on suuri riski erävaikutuksen esiintymiselle. Erävaikutus kuvaa tietokannoissa esiintyvää harhaa, joka johtuu kohorttien poimimisen aikana tehdyistä oletuksista ja menetelmistä. Raman-tietojen kuratointi erävaikutuksesta on tärkeää, jotta malli perustuu syöpään liittyviin malleihin eikä poimintaan liittyviin vaikutuksiin.

Tässä opinnäytetyössä esittelemme matemaattiseja malleja, jotka pystyvät muodostamaan ennusteita kasvaimen geneettisten ominaisuuksien luokittelusta. Kehitämme menetelmiä Raman-spektrien kuratoimiseksi koneoppimisen ja synteettisen datan generoinnin avulla ja osoitamme niiden tehokkuuden glioomakasvaimia koskevilla datajoukoilla. Tämän lisäksi kehitimme menetelmän datajoukon laadun parantamiseksi, jossa käytämme koneoppimista poistamaan erävaikutuksia, mikä edistää syöpään liittyvien indikaattoreiden havaitsemista mallissa. Meidän osuutemme glioomien luokittelun alalla ovat luokittelumallit sekä niiden mahdollistavat strategiat. Esittelemme syväoppimisarkkitehtuurin, joka käyttää minimaalisia parametreja spektrien korjaamiseen. Esittelemme myös strategian, jolla voidaan vähentää erävaikutusta vastakkainasettelun oppimisen avulla ja mitata samalla geneettiseen luokitteluun liittyviä ominaisuuksia. Tässä opinnäytetyössä osoitamme, kuinka menetelmiemme soveltaminen voi parantaa glioomien luokittelumallien suorituskykyä.

ASIASANAT: Koneoppiminen, Glioma, Raman-spektroskopia, Ominaisuuksien poiminta, Syväoppiminen

Acknowledgments

This work would not have been possible without the support of individuals to whom I am extremely grateful. First, I wish to express gratitude to my supervisor Ion Petre, whose knowledge is vast, and empathy appears infinite. Thank you for letting me join this project, which allows me to put all I have studied into practice, for challenging me on the things I have misunderstood, and for your endless support. I could never have asked for a better supervisor. To my second supervisor, Vesa Halava, and my research director Jarkko Kari, I am grateful for the support and assistance you have given me through the years. I also wish to thank my opponent, Adrian Iftene, who put significant time and effort into studying my work, and for his valuable comments on my thesis. Isaac Afara and Nicolae Leopold have reviewed my thesis and provided feedback on the quality of my work, for which I am extremely grateful. For providing a productive working environment and for funding the finalization of this thesis, I want to thank the Department of Mathematics and Statistics. It has been an honor to study alongside some of the most driven people I have ever met. I also wish to thank my co-authors, especially Mioara Larion and Adrian Lita, whose extensive knowledge in neuro-oncology and Raman spectroscopy enabled my work. It has been a true privilege to work with all of you. I wish to thank the EXACTUS programme at the University of Turku, Turku University Foundation, and the Swedish Cultural Foundation for funding my research work. Special thanks are extended to my partner Farima Ajdari, who helped produce the Finnish translation of the abstract, and for supporting me on my journey.

Finally, I want to thank my parents, extended family, and all my friends and peers. I dedicate this work to all of you.

March 2026
Joel Edvard Christian Sjöberg

Disclosure on the use of AI: ChatGPT was used to critique my writing style and to identify spelling errors. The content included here is my own, developed with feedback from my advisor.

Table of Contents

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Adrian Lita, Joel Sjöberg, David Păcioianu, Nicoleta Siminea, Orieta Celiku , Tyrone Dowdy, Andrei Păun, Mark R. Gilbert, Houtan Noushmehr , Ion Petre , and Mioara Larion. Raman-based machine-learning platform reveals unique metabolic differences between IDH^{mut} and IDH^{wt} glioma. Neuro-Oncology, 2024; 11: 1994-2009; URL: 10.1093/neuonc/noae101.

II Joel Sjöberg, Nicoleta Siminea, Andrei Păun, Adrian Lita, Mioara Larion, and Ion Petre. RADAR: Raman Spectral Analysis Using Deep Learning for Artifact Removal. Advanced Optical Materials, 2025; Volume 13: 2500736; URL: 10.1002/adom.202500736.

III Joel Sjöberg, Adrian Lita, Mioara Larion, and Ion Petre. Towards Unbiased Raman Spectroscopy: Deep Learning Solutions for the Batch Effect. Under review.

1 Introduction

Glioma is a type of cancer manifesting within the brain; the most aggressive form of glioma is glioblastoma, which possesses a median survival of 10.4 months [1]. Treatment of gliomas includes surgical resection, drug treatment and chemotherapy to manage tumor progression [2; 3; 4; 5]. Research in this context is focused on identifying genetic properties within gliomas which can separate glioma subtypes. The identification of genetic features functioning as treatment targets is an active field of study which can improve personalized glioma treatment [6; 7; 8; 9; 10]. For this purpose, Raman spectroscopy has been applied as a medium for translating tumor tissues into numerical form. The molecular fingerprint extracted using Raman spectroscopy includes informative features of material properties. However, Raman spectra require precise and time-consuming efforts to properly curate and analyze. Machine learning (ML) models are computational methods capable of rapid processing of large feature-rich datasets. Analysis through Raman spectroscopy combined with ML methods is a field of study with the potential to improve our diagnostic capabilities of gliomas. The ability of ML models to learn complex data properties makes them suitable for learning tasks associated with Raman spectroscopy. One such learning task is data curation, which requires extensive knowledge about data patterns. Leveraging ML to learn these patterns in combination with synthetic data generators designed to produce Raman-spectrum-like data has led to models for spectrum curation [11; 12; 13]. The ability to generate synthetic Raman spectra enables the training of large ML models.

In this thesis, we approach these fields with the goal of developing reliable ML models for the curation and classification of Raman spectroscopy data extracted from gliomas. We explore a dataset consisting of 46 unique patient glioma tumors, represented by Raman spectra extracted from their microscopic surfaces. The dataset is provided by our collaborators Dr. Mioara Larion and Dr. Adrian Lita from the National Cancer Institute (NCI) branch of the National Institutes of Health (NIH) who are experts on glioma and Raman spectroscopy. Our contributions to this field come in the form of RADAR, the first DL solution for extracting known artifacts and peaks from Raman spectra of variable length. Our second contribution within this field is through a proof-of-concept showing how an adversarial learning method can reduce the batch effect in Raman spectrum data. We also present a feature importance layer for DL models designed to identify and rank important data features akin to random

forest (RF) model capabilities. We apply our methods to the glioma dataset and show how classification of the genetic properties is improved using our methods compared to conventional ones.

Raman spectroscopy is a technique which captures and describes the vibrational states of molecules. This technique has gained traction in applications to biological tissues; notably cancerous tissue, due to the information-rich features extracted through Raman scattering. Raman spectroscopes measure the change in photon frequency caused by molecular interactions inside the material. The molecular information of materials is encoded into Raman spectra, containing thousands of features, which represent a molecular fingerprint of the measurement point. This technique is label-free and rapid, enabling point-wise examinations and whole tumor analysis in clinical settings. Application to entire tumor samples can capture genetic information which enables analysis of the heterogeneous tumor environments within gliomas. However, Raman scattering is easily affected by factors outside the materials' molecular properties. Photons can be easily affected by environmental light and cosmic rays, which disturb the Raman process and result in spectral artifacts. These artifacts obscure Raman peaks by introducing spectral noise, baseline signals and sporadic spikes. Proper curation of these artifacts relies on analytical methods which require manual parameter tuning for determining appropriate parameterization of said methods.

Current ML methods for classification of glioma through Raman spectra focus on capturing patterns related to genetic properties such as gene mutations and chromosome loss. For example, the ability to confirm IDH mutation status of tumors is important due to their correlation with favorable prognosis in contrast to IDH wild-type tumors. ML methods have successfully modeled the relationship between Raman spectra and IDH mutation status through linear discriminant analysis (LDA) and principal component analysis (PCA) [14]. Providing proof for the ability of Raman spectra to capture genetic properties in glioma tumors. Examples of avenues for research into ML models utilizing other genetic features captured by Raman spectroscopy include classification of 1p/19q codeletion, EGFR amplification and G-CIMP status. However, problems related to the curation of Raman spectra and training models for genetic classifications in glioma still limit the ability to perform detailed analysis of tumors. Current problems within the field of modeling tumor genetics through Raman spectra center on the lack of understanding internal model operations. This is due to the complex computational flow used by ML models to form predictions. One potential remedy for this issue is to require explainable behaviors through feature importance scoring of the trained models. Other issues regarding classification of glioma subtypes include batch effects present in large combined datasets and heterogeneous genetics within gliomas. Gliomas have heterogeneous genetics which necessitates tumor-wide analysis as opposed to sampled measurement spots [15].

Data availability is a considerable problem within the field of ML which can be partially solved with data generation provided the generator can produce diverse data points akin to real Raman spectra. ML applications utilizing Raman spectroscopy have also resulted in successful models for the classification of cancers and their subgroups [16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26]. The ability to apply ML models on large glioma datasets of Raman spectra can help further understanding of the diverse glioma landscape and reveal Raman features expressive of genetic properties. Developing models capable of feature interpretation and scoring is essential for providing informative predictions. This is especially relevant in medical settings where decisions must be verifiable to ensure trust in medical decisions by patients and experts alike. Training ML models includes developing strategies for mitigating these problems in part by selecting modeling methods with robust performance on small datasets.

Three main model architectures were used in this thesis: random forests, support vector machines, and convolutional neural networks. RF models are ensemble models composed of decision trees capable of learning classification tasks while also computing the feature importance driving their decision-making [27]. They offer explainable behavior through their compositions of decision trees and feature importance attributes, enabling explainable decision-making based on Raman features. Their feature scores can be utilized to discard irrelevant data features which reduces memory requirements. Support vector classifiers (SVC) are robust classification models capable of transforming non-linear classification problems into a linear domain [28]. By training SVC models on the feature-reduced data, SVCs can provide increased predictive accuracy compared to the initial random forest models. Another model capable of non-linear classification is convolutional neural networks (CNNs). Through parallelized tensor operations they form an efficient method for processing large, feature-rich Raman spectroscopy datasets. Due to the customizable architectures of deep learning (DL) models, CNNs can be trained to perform advanced data transformations [29]. These models are suitable candidates for providing pattern recognition of Raman spectra to capture genetic features.

The thesis is structured as follows. We begin in Chapter 2 by covering research regarding glioma and Raman spectroscopy. We present common concepts within both fields and give background on the established subtypes of glioma. The chapter concludes with an introduction to Raman spectroscopy and how ML methodology is applied to it. In Chapter 3 we introduce our first contribution to this field in the form of RADAR, a DL solution enabling curation of Raman spectra with unconstrained spectral lengths. Furthermore, we present a method for reducing the batch effect in Raman spectroscopy datasets; an effect which causes non-trivial bias patterns in datasets, obscuring relevant data patterns. Chapter 4 covers classification models capable of classifying glioma subtypes with high accuracy using Raman spectra. The chapter concludes by applying our methods presented in Chapter 3 to the ML

pipeline and shows how the classification methods benefit from batch-effect-reduced data. The thesis concludes in Chapter 5 with a discussion on possible future prospects of continued studies in this field.

2 Background

Neuro-oncology is the study of neoplasms manifesting inside the central nervous system (CNS), which develop into tumorous growths. Over 100 different types of tumors have been identified to date, each with distinguishable traits and varying rarity [5; 30]. Of specific interest are gliomas, characterized by the potential for aggressive growth and reduced life expectancy in patients [31].

2.1 Glioma

Gliomas originate from glial cells exhibiting abnormal behavior and can in general be divided into multiple histological subtypes: astrocytoma, oligodendroglioma, ependymoma and glioblastoma. Each subtype exhibits different degrees of aggressiveness in terms of their spread and growth [4; 32]. The World Health Organization (WHO) categorizes glioma tumors according to molecular properties into four different grades (grades I to IV). Low-grade glioma (LGG) encompasses grades I and II; these exhibit less aggressive clinical behavior in contrast to high-grade glioma (HGG) which possess more aggressive traits (grades III and IV) [4; 33]. The definitions of these grades have been expanded and clarified over time to better explain the molecular properties of tumors rather than immediately observed behavior [32].

Between the years 2017 and 2021 51.5% of all malignant brain and CNS tumors in the United States were identified as glioblastoma, the most aggressive form of glioma [31]. On average, 17,411 patient deaths in the United States of America occur annually from malignant brain and CNS tumors [30]. In 2022 the global estimate of both malignant and benign brain and CNS tumor cases was 322,000. It is estimated that the yearly number of global cases will increase to approximately 474,000 by the year 2045 [33]. In Finland, a study published in 2019 [34] was conducted on statistics of glioma incidence in data from between the years 1990 to 2006 and 2007 to 2016. The incidence rate was observed to be 7.7 per 100,000 people in the period between 1990 and 2006 and 7.3 per 100,000 in 2007 to 2016. An increasing incidence in malignant glioma among those 80 years or older was discovered in the period between 1990 and 2006.

Clinical behavior alone cannot sufficiently characterize the complex genetic makeup in glioma tumors. The WHO also requires classification according to established biomarkers [35]. The identification of biomarkers which can properly identify known

and novel subgroups of glioma is an active field of study [4; 32; 36; 37; 38; 39]. The need for biomarkers is especially relevant when considering the heterogeneous makeup of CNS tumors [40; 41]. Standard practice in diagnosing and treating tumors has evolved to utilize biomarkers for informed treatment plans [42].

Heterogeneous tumor environments can complicate diagnosis and treatment [40; 43]. Heterogeneity also complicates surgical resections due to tumor region infiltration into healthy brain tissue. The diffuse region causes uncertainty when locating tumor borders during surgery. Excessive removal can leave lasting damage to the patient, while conservative resection risks reemergence of tumors post-resection. [38; 44; 45; 46]. Despite this, resection remains an essential part in treatment, and is often combined with medicine to extend life expectancy [47]. Continued efforts are made to further develop existing and novel surgical techniques to increase accuracy and efficiency of invasive procedures [48; 49; 50].

Diagnosis and treatment of highly heterogeneous tumors requires an understanding of the diverse characteristics within different tumor areas. Prompt diagnosis and treatment are essential to avoid tumor progression. For example, grade 2 gliomas may transform into higher grade gliomas in a period of 5 to 10 years [49; 50; 51]. Targeted treatment of specific tumor types can suppress homogeneous tumor areas but may also fail to suppress heterogeneous areas which then develop resistance to previously applied treatments. To properly suppress glioma with medical treatment, a highly personalized approach is necessary to address the diverse tumor environment [46; 7; 52]. This has led to molecular targeted treatment strategies being applied to glioma patients [5; 37; 53]. By examining the molecular microenvironment of tumors, surgeons are able to map out heterogeneous regions of tumors to better understand the unique genetic makeup of individual tumors [40; 54; 55]. The following are examples of factors with proven characterization of glioma categories.

IDH mutation. Isocitrate dehydrogenase (IDH) gene mutations are commonly associated with LGGs and improved prognosis. IDH mutant tumors usually signify increased survival rates compared to IDH wild-type tumors i.e. IDH without mutations. IDH mutations are important markers for categorizing gliomas making them essential for predicting survivability [42].

G-CIMP Status. A subset of the IDH mutant glioma cohort have G-CIMP (CpG island methylator phenotype) properties. This subset can be further divided based on DNA methylation, known as G-CIMP-high and G-CIMP-low for high and low methylation respectively. G-CIMP-high is associated with better prognosis compared to the low variants. [56; 57]. In [58] they discovered that recurrent G-CIMP-low variants shared resemblances with glioblastoma [59]. The ability to differ between these subtypes would provide a valuable prognostic marker for surgeons [60].

1p/19q Codeletion. Deletion of the short arm and long arm of chromosomes 1 and 19 respectively along with IDH mutant status is characteristic of the LGG oligodendroglioma. It is a strong marker indicating oligodendroglioma status which includes better prognostic outcomes [61; 62; 63; 64; 65].

Glioblastoma Subtypes. Glioblastoma tumors can be subdivided according to [66]. These were based on identified genetic differences in the genes NF1, EGFR and IDH1. Their differences led to the so-called mesenchymal, classical and proneural-like tumors [67]. These subcategories provided an insight into the genetic diversity possessed by glioblastoma multiforme tumors and provided characteristics for their identification [68; 69]. According to [70], between 30 to 49 % of glioblastoma tumors are characterized by the presence of mesenchymal stem cells. The mesenchymal-like tumors are characterized by lower expression levels of the NF1 gene. Classical tumors had frequent EGFR amplifications and a lack of TP53 mutations. Proneural tumors were more likely to contain IDH1 and PDGFRA mutations often seen in LGGs.

Methylation Clusters. Investigation carried out by [71] through machine learning approaches with glioma samples identified 6 unique clusters which were validated as separating tumors according to their methylations. These classes were organized in a hierarchical structure. All tumors could be separated into one of the two methylation classes IDH mutant or IDH wild-type. Within each of these cohorts, tumors could be further divided into three subgroups depending on IDH mutant status. Within the IDH mutant cohort the methylation clusters LGm1-3 consisted of tumors with G-CIMP-low (LGm1), G-CIMP-high (LGm2) and 1p/19q codeletion (LGm3). Within the IDH wild-type cohort were the classic-like (LGm4) and mesenchymal-like (LGm5) tumors along with pilocytic astrocytoma (PA)-like tumors (LGm6). PA-like tumors are still a subject of uncertainty regarding their proper classification, they are known to possess BRAF alterations and are generally considered a grade I tumor [72]. These clusters showed correspondence with established markers required by the WHO [32].

2.2 Raman Spectroscopy

Raman spectroscopy is a vibrational spectroscopic technique which measures the inelastic scattering of photons, capturing the molecular vibrations of materials. The Raman spectrum is often referred to as the molecular fingerprint due to its ability to express molecular compositions of measurement points. It captures this information by measuring energy changes in photons after molecular interaction. These energy changes describe molecular vibrations within the material caused by interaction with the incident photon [73; 74]. Due to its ability to transform molecular vibrations into

measurable spectra, Raman spectroscopy has been applied to the study of cancers in multiple cases to analyze tumors at the molecular level [17; 21; 75; 76; 77]. Material properties are analyzed through the widths and positions of high-intensity peaks inside spectra called Raman peaks. These are often modeled using Gaussian and Lorentzian curves and Voigt-profiles [11; 12]. Comparison between peak positions and intensities can be performed to identify different material properties. This makes Raman spectra a suitable medium for describing the molecular makeup of gliomas in numeric form. The analysis of molecular information in Raman spectra has the potential to inform decisions on treatment lines fine-tuned uniquely for each patient [66; 78].

Analysis of Raman spectra is a laborious task; from a statistical perspective, they are high-dimensional data which contain thousands of features. Analysis of spectra usually requires identification of informative spectral regions whose frequencies contain relevant information. For example, silent regions in Raman spectra are regions devoid of Raman signal; these regions are best removed to avoid interpreting noise as informative signal [79]. This step is crucial in cases where computational resources are limited. To further reduce the computational load and analytical work, spectra are often reduced to smaller sections where established relevant information is present. These areas depend on the research context, but the full Raman spectrum is seldom included for analysis. Since the identification of relevant Raman peaks for molecular identification is an active field of study, there is potential to discover new relevant peaks within the research context. Novel peaks with relevant information are therefore at risk of being ignored when only a subset of the spectrum is analyzed. We hypothesized that including the entire Raman spectrum could lead to better understanding of the information carried in Raman spectra. With the omission of the silent regions, we wanted to analyze the importance of all peaks for glioma classification.

It is worth noting that Raman scattering is easily affected by the environmental conditions around the instrument used to measure it. Fluorescence and other environmental light may introduce baseline signals into spectra, which obscures Raman peaks [80]. Heat is also a factor during measurements, as the laser may heat up the materials, risking damage to samples which affects their molecular properties [81; 82]. The components that disturb spectral signals are usually categorized into three main component groups: baseline signals, cosmic rays and noise. Of these, cosmic rays are by far the rarest, originating from high energy particles which appear as sharp spikes in spectra. These components can mimic or obscure peak patterns causing uncertainty when analyzing spectra and potentially leading to misinterpretations of peak regions. Removal of the disturbing components is vital to avoid false positives during analysis.

Baseline correction. The baseline signal of a spectrum can be represented as a continuous polynomial line of variable degree. These can be caused by environmental

light and fluorescence which cause elevations of entire spectral regions, heightening the Raman signal within said regions. Baseline correction methods aim to approximate this signal with high precision to then remove it from the affected spectrum. The polynomial fitting methods modPoly and imodPoly [83; 84] both require a polynomial degree as input parameter to fit a polynomial line with input degree to the input. However, given sufficiently large degrees, these methods can fit to the entire spectrum. This can result in Raman peaks being identified as part of the baseline signal. To address this setback, the adaptively reiterated weights penalized least squares (airPLS) [85] method introduced intensity weights in combination with Whittaker smoothing to adapt and ignore peak regions. This method is reliant on two parameters related to the order of the polynomial curve and the smoothness of the Whittaker smoothing algorithm. Recent methods for baseline correction have been provided in the ORPL-package which introduced the bubblefit algorithm [86]. Their method approximates baselines by fitting ellipses to the underside of the spectra. The ellipses are increased in size until they reach a collision point with the spectrum. The method then leaves the upper side of the ellipse as the baseline, provided the ellipse collided with the spectrum or the maximum ellipse size was reached. Their method requires the ellipse size be provided as a parameter.

Cosmic ray removal. Cosmic rays (CR) are rare high energy particles which can enter the instrument during extraction. Their appearance as high-intensity spikes can dwarf Raman peaks resulting in statistical differences between affected and unaffected spectra regarding the maximum intensity frequencies. One possible way to remove CRs in a dataset of spectra is via outlier correction as outlined in [87]. Computing the standard deviation on each frequency in the dataset can identify spectra and frequencies which deviate from the majority. Deviations can then be replaced by the mean or median depending on the amount of noise in the dataset. The approach proposed by [87] utilizes the derivative of the spectrum signal and the median absolute deviation (MAD) of the data to compute a modified z-score of the location around the potential CR. This method is expressed as

$$Z = \frac{(0.6745 \cdot \nabla S) - M}{|MAD|}, \qquad MAD = \texttt{median}(|\nabla S - M|),$$

where the median M is subtracted from the derivative ∇S which is used to compute the MAD value for each frequency of the spectrum S. Spikes can then be removed by replacing values of Z where $|Z| > 3.5$ as suggested by [87].

Alternatively, the interquartile range can be applied in place of the standard deviation method to allow for higher frequency varieties. However, this method is reliant on spectral uniformity and can lead to catastrophic intensity removal if the dataset consists of spectra with varying intensities. An alternative method for use on single spectra is to utilize a sliding window approach by considering a local area of a

few frequencies. The standard deviation or interquartile range can then be utilized to identify if the center of the area deviates significantly or not from the surrounding area. Note that the decision on what defines significant deviance often is context dependent and may require parameter tuning. The Savitzky-Golay filter [88] is a method for incorporating the sliding window approach to spectra. While fitting a polynomial (with variable degree) through the window, this method can remove CRs and also lead to denoising. The method manages to remove these components by exchanging the spectral line with the interpolated polynomial line.

Denoising. Noise in spectra is characterized by sporadic intensity changes between adjacent frequencies. The intensity of which can be reduced by increasing the exposure time for the measurement spot. In extreme cases, the noise can dwarf peak regions, which significantly complicates the analysis of spectra. Noise affecting the spectral intensities is commonly caused by environmental factors hard to control during extraction. It is therefore necessary to have reliable methods available for post-processing of spectra to salvage spectral signals. The effect of noise on the spectral signal is often measured using the signal-to-noise ratio (SNR). This ratio is defined as the ratio between the mean signal and the standard deviation of the noise. For spectral data, the mean signal can be applied to the peak regions exclusively, avoiding reduced signal mean by including the silent regions. The ratio is formally expressed as

$$\texttt{SNR} = \frac{\mu_s}{\sigma_n}.$$

Proper measurements may require the ability to identify peak regions to measure the mean signal and conversely, the silent areas for the standard deviation of the noise.

2.3 Raman-Based Deep Learning for Oncology

To analyze the patterns in Raman spectra, experimentalists require data analysis methods for unraveling relevant information while ignoring patterns irrelevant to their studies. Identifying relevant Raman peaks is a challenge with potentially unique solutions for each individual study. It is not always certain which peaks should be analyzed for a given research question. Furthermore, analysis of data containing spurious components like baseline signals can lead to bias due to the amount of unique patterns available. This means that Raman spectra require qualitative curation to remove erroneous patterns, decrease spurious patterns and to reduce the risk of wrongful conclusions.

ML methodology can be applied on Raman spectroscopy data to solve these issues. The ability to process large tabular datasets with spurious patterns makes models ideal for curating and analyzing Raman spectra. Recent literature has provided insight into how ML, and more specifically DL models, can be applied to Raman

spectra to produce high accuracy predictions. We review recent methods for curating and classifying Raman spectra to give an overview of the field.

2.3.1 Spectrum Correction

Examining dataset patterns is required to select the appropriate curation methods for each component type. With the selection of curation methods comes inevitable parameter tuning for each method selected. For larger datasets, it can be infeasible to find the best parameter, requiring a grid search by utilizing each method with different parameters. Furthermore, large datasets of spectra containing components with high variation can lead to suboptimal parameter selection.

In contrast to the established analytical methods designed to curate Raman spectra, ML models can be trained to provide complex, non-linear curation of spectral data. The benefit of ML for data curation in this case lies in their optimized computational performance, allowing for rapid parallel curation of Raman spectra. Another factor is the ability of these models to learn concrete data patterns based on training data. This alleviates the need for data analysis required to determine appropriate method parameters. This process is instead inferred through parameter tuning performed by the model during the training process. Through large datasets with appropriate labels, the model can learn to process complex patterns and provide precise curation for them. One limiting factor in ML is the acquisition of large datasets. To train models for Raman correction would require extensive datasets of Raman spectra. These spectra must be contaminated by the problematic components and be paired with the corresponding target values represented by corrected spectra. Gathering a dataset of this type is considered infeasible, since the variety of data requires many different materials to scan and experimentalists for scanning them using various Raman instruments. Furthermore, the practical work required to determine the correct curation method for each spectrum is infeasible for research teams with limited resources. There is no established lower limit for the number of data points required to train ML models. For DL models a rule of thumb is to have ten times the number of learnable parameters to avoid model overfitting [89]. For models meant to provide curation for diverse types of Raman spectra, large datasets containing spectra from multiple different cohorts are needed. These datasets would ideally contain components originating from varying environmental conditions. To increase the robustness of model predictions on rare patterns, examples of rare patterns must likewise be collected in considerable magnitude to allow for model adaptability.

By far the most efficient way to gather data along with their target values is to generate them. Data generation is an active field with great value for ML modeling as it promises to solve the data limitation problem for model training. Within the context of data generation for Raman spectroscopy this has been done by empirical observations made from available datasets to estimate the data distributions.

The successful efforts of [11; 12; 13; 90] show that synthetic Raman-like data can be generated efficiently to address the need for large datasets. Data generation has an additional benefit of requiring less working memory as data can be generated dynamically for training and then deleted to make space for the next generated batch. In [11] synthetic spectra are generated through a linear combination of spectrum and background signals, each containing synthetic peaks generated as Lorentzian peaks. Baseline signals are introduced as polynomial lines with random degree. Poisson distributed noise is then added to each frequency to produce the final synthetic spectrum. Their generator concludes by adding cosmic rays which are generated as sharp Lorentzian curves. The difference between peaks and cosmic rays is dependent on the γ-parameter of the Lorentzian function defined as

$$\frac{\gamma\pi^{-1}}{(x-p)\gamma^2},$$

where γ is the scale of the peak intensity and p is the position of the peak on the spectrum. Both parameters are drawn from a Gaussian distribution with mean of 0 and a standard deviation of 10. Their work asserts that any smooth mathematical line could be used as a background signal. Similarly, Gaussian noise could replace the Poisson noise. The choice between these methods appears arbitrary.

The approach by [90] developed a peak library consisting of raw Raman peaks extracted from real spectra. The utilization of real data to construct synthetic data is an efficient way to generate new data reminiscent of data augmentation methods. They showed that their adversarial model was sufficient for learning the synthetic data by evaluating it on real data. Evaluation on real data is essential in projects utilizing synthetic data to ensure that model training on the former variant can enable real-world applicability.

While the inner workings of large CNN models remain unknown, both models by [11] and [12] contained over 10 million parameters, and in simple terms utilized convolutional layers to process the input before transforming the output using fully connected layers. Informally the signal flow through both networks relied on layers computing latent features which could then be used by the fully connected layers to produce an output. We hypothesized that their high performance was possible in large part due to the utilization of fully connected layers as they allow for feature association through the entire spectrum. The limitation introduced by this flow is that the densely connected layers require defined input-, and output-sizes. Both models exclusively process spectra with 1024 frequencies. In the work by [91] a similar limitation is imposed for their denoising model, forcing an input length of 600. This forces the users to limit or transform their data to fit into the models, which can be error-prone if done carelessly.

The conclusion drawn from our observations was that higher accuracy is achieved in producing outputs if the layers producing it can reach global information from the

input. Thus, densely connected layers have an advantage in producing higher accuracies due to access to the global input context. A higher accuracy is also achieved if the model is forced to learn patterns separate from the Raman peaks. The work in [12] gained higher accuracy when compared to [11] in part due to the capability of learning the baseline patterns along with the Raman peaks. In the second Paper presented in this dissertation, we examine the feasibility of achieving similar accuracies as [11; 12; 13] by exclusively utilizing CNN layers to allow for variable input sizes. The availability of a model free of input constraints would increase the applicability of ML models in Raman research.

2.3.2 Classification Models

The usability of Raman spectroscopy combined with ML methodology has been well established in recent years [92]. Models aimed at leveraging the molecular fingerprint for assisted research work in cancer data have been successful for gastric [17; 18], breast [20; 19], lung [93] and bladder cancer [21] to name a few. While the majority of works are aimed for specific classification tasks such as predicting genetic features and malignancy status, other works have focused on data exploration. For example, in [17] ML was used to explore and establish markers to differentiate between healthy and cancerous tissues in gastric cancer. In [22] they managed to establish changes induced in the microenvironment following immunotherapy, providing a predictive model for patient therapy response. The cancer grading of chondrosarcoma (bone cancer) has also been achieved through the use of ML and Raman spectroscopy [94]. These results prove that Raman spectroscopy can be leveraged to unveil genetic factors through ML which can be further studied to deepen our understanding of cancers.

The applicability of DL on cancer data through spectroscopy has also been extended to glioma classification. In this context there have been efforts for differentiation between cancer types as in the case of CNS lymphoma and glioblastoma [26] and even models capable of application on different types of cancer [93]. One of the earliest models for IDH mutant classification [95] was achieved through PCA and LDA modeling. Their efforts hinted at the value of dimensionality reduction for computational efficiency while enabling relatively small classification models to achieve high accuracy. The differentiation between IDH mutant and wild-type tumors has also been achieved using extreme gradient boosting trees and support vector machines with radial-basis function kernels. The tree-based models could be further utilized to establish novel Raman features which could be use for their classification [96]. In [71] unsupervised clustering and random forest models were used to establish groupings of different glioma subtypes. Their work showed that through clustering sequenced glioma tumors they arrived at six distinct clusters. Through evaluation of the cluster relevance they found that each cluster contained combining

factors relating to various genetic factors. In Paper III, we explore the possibility to model for different clusters identified by [71]. We established the possibility of utilizing Raman imaging combined with random forest and support vector machine models to predict IDH mutant status and G-CIMP status in a dataset consisting of 59 unique glioma tumor samples (total of approximately 250,000 Raman spectra).

3 AI Methods for Raman Curation

In the first phase of our work covered in this thesis, we explored data analysis and curation methods for Raman spectra datasets. We examined the functionality and effectiveness of established preprocessing methods for Raman spectra and evaluated them using our dataset. In doing so, we noted several open problems within the field of preprocessing and modeling for Raman spectra. Our contributions presented in this context are aimed to address these problems. We also aim to provide models capable of curating data with satisfactory accuracy for future projects utilizing Raman spectra.

The heterogeneous nature of glioma tumors provides multiple challenges in diagnosing tumor types. For example, treatment lines can become inefficient if they do not respond to the entire genetic landscape within gliomas. ML models trained for glioma classification must utilize the genetic profiles of gliomas to produce accurate predictions. To this end, we utilize a dataset consisting of 46 patient tumors extracted by Dr. Adrian Lita and Dr. Mioara Larion at the NIH/NCI. Each patient sample consists of between one and four glioma tumor samples. All tumors have been sequenced and applied to the methods introduced by [71] to provide labels for their genetic properties (LGm1-6). Each surface of the available patient tumors has been scanned using Raman imaging, resulting in a dataset consisting of over 300,000 Raman spectra. The tumor samples have been preserved in formalin-fixed paraffin embeddings (FFPE) to shield them from environmental effects. This step is essential to avoid sample degradation over time due to environmental exposure. Ensuring the usability of FFPEs for analysis of gliomas would enable larger studies as tumor samples can be preserved without deteriorating the Raman signal. We seek to confirm the effectiveness of FFPE preservation and promote glioma preservation through our work. We do this by showing that models trained on gliomas encased in FFPE can categorize their genetic properties. Our work is aimed at the following three concepts within data preparation.

RADAR: Removal of Spectrum Artifacts. The first set of questions concerns the correction of Raman spectra. Established methods in this field rely on parametric methods which require an intuitive understanding of the components present within datasets of Raman spectra. Furthermore, methods which see popular use for Raman data are also contextual, i.e. baseline correction methods can remove Raman peaks or

maintain parts of the baseline signal as an erroneous peak if suboptimal parameters are provided. For larger datasets, these methods also require long periods of computation, making parameter tuning for large spaces of parameters infeasible. Recent works have trained DL models for alleviating the need for intuitive understanding of the dataset. These models have been proven to outperform the previously established methods within their own study limitations. These DL models often rely on data generators to satisfy the need for large datasets during model training. However, these models have also been limited by their availability, and restrictions regarding the spectrum length. Another concerning factor regarding these models is their sizes. The majority of the published models rely on large architectures with millions of parameters. While this can promote complex preprocessing behaviors for spectra, it can also lead to overfitting within the data domain and uncertainty regarding overall model functionality. This issue is further exacerbated by the fact that many solutions aim to outright return the corrected spectrum rather than the extracted components to be removed. This can lead to uncertainty following model application on spectra since the exact change becomes unverifiable by complex computations. This exemplifies the necessity to develop methods capable of complex computations while providing a degree of rationalization regarding their final outputs.

One case where verifiability is essential in correction models can be seen when correcting spectra of lower exposure times. Exposure time refers to the time the spectroscope lingers on a given measurement spot to extract as clean a signal as possible. In projects looking to extract large sets of Raman spectra, the exposure time can become a bottleneck leading experimentalists to optimize for minimal exposure time. This leads to compromised signal integrity which can affect later analytical work. Through reliable DL models for spectrum correction the exposure time could be minimized while signal integrity is maintained. This requires that the correction model performance can be verified to prove that peaks are indeed maintained. Furthermore, verification on model performance for different exposure times can establish a lower bound on exposure time where the model can be guaranteed to perform accurate corrections.

Our contribution to the problem of spectrum correction comes in the form of DL models capable of correcting spectra possessing any length. Our model is the first to attempt complete extraction of baseline signals, cosmic rays, noise, and Raman peaks. Through component extraction in this way our model can provide clean Raman peaks with explanations on what other components were removed in the process. This provides a degree of explainability not seen in other models aimed at spectrum correction. Furthermore, our models are designed to contain as few learnable parameters as possible. With this, we aim to lessen the risk of overfitting while providing consistent performance on spectra from different data cohorts. We demonstrate that our models can be utilized to lessen the exposure time by approximately 90% on glioma spectra, reducing the maximum exposure time of 5 seconds within

our dataset to 0.5 seconds. Another contribution we provide is through our synthetic data generator, capable of producing Raman-like spectra. Our generator can generate synthetic versions of all spectrum components to give an overview of the entire spectrum composition. These can be added to produce a synthetic Raman spectrum.

Unsupervised Identification of Tumor Areas. We tested methods for clustering and dimensionality reduction of spectra to establish an efficient pipeline for glioma curation. While most ML methods are robust and can handle small sets of outliers in data, it is important to determine the number of present outliers and remove them. Using our methods to correct the available spectra we set out to utilize computational methods for visualizing and clustering the tumor environments. Through unsupervised density clustering algorithms and dimensionality reduction methods we managed to identify the presence of approximately 50,000 outliers in our dataset. We showed how the outliers affected the dimensionality reduction of the data by comparing the reduction to the version performed without outliers. Doing so unveiled the true diversity of the tumor microenvironments, enabling further analysis on tumor heterogeneity. Dimensionality reduction is also used as a tool to visualize the tumor surfaces with high detail. We evaluated the outlier detection procedure and determined their origin as spectra originating from blood, paraffin, support substrate and over-heated measuring points. This confirmed the need to remove them from further analysis of gliomas.

Adversarial Training for Reducing the Batch Effect. Creating large datasets of cancerous tissues is a tremendous effort requiring time and money. Analyses provide better results the more data is available and including patient tumors in datasets require explicit permission from the patients themselves. Provided the patients give permission, there is still a risk of mistakes when handling the samples which may result in unusable data. Preservation is essential due to the relatively low incidence of gliomas available for study. Many institutes examine patients with glioma which easily leads to variation among studies regarding extraction, handling and preservation. Differences in these stages can result in a phenomenon commonly referred to as the batch effect [97; 98; 99; 100]. The batch effect is commonly characterized by the difference in institute practices when handling biological data. Though these practices are methodologically correct in isolation, they make samples incomparable across combinations of multiple datasets due to non-uniform practices. An example of this effect was examined in a study conducted by [101]. They identified the batch effect in tumors extracted from different acquisition sites due to color staining, a common practice in identifying tumor surface properties. The staining caused tumors to be categorized according to the color of the samples rather than their surface patterns. The potential in amassing and properly analyzing large datasets of biological data completely depends on our ability to remove the batch effect present

in combined datasets. Another problematic factor present in surgical practice is the possible non-tumor areas present in the tumor samples. These may be areas populated by necrotic cells or areas where blood has been concentrated, obscuring tumor surface during extraction. We investigated the batch effect in context of the diverse tumor environments available to us. While this dataset comes from one institute, we confirmed that the Raman spectra contained features descriptive of their tumor of origin. This indicated that even corrected Raman spectra contain information independent from the environmental factors responsible for baselines, cosmic rays and noise. We set out to curate our dataset of the batch effect in this context while being careful not to remove features necessary for classification of genetic properties. Our contribution comes from demonstrating an adversarial approach to remove the batch effect via feature scaling on Raman spectra. We developed a feature importance layer designed to learn feature importance of the input during model training. By training classification models and the feature importance layer adversarially we converged at a feature transformation capable of reducing the batch effect in our dataset.

3.1 RADAR: Removal of Spectrum Artifacts

In our dataset, we identified three artifacts commonly seen in Raman spectra; namely baselines (B), cosmic rays (CR) and noise (N). B signals are often described as a polynomial line of random degrees on which the Raman peaks are located. These lines are primarily responsible for changing the peak ratios and sometimes even dwarf the peaks, rendering them undetectable during visualization [102].

Available DL models for correction have been designed for specific spectrum lengths in part due to their reliance on densely connected layers [11; 12; 90]. This requires users to either remove part of the spectral frequencies, padding them or squeezing them together to fit the model input requirements. Furthermore, none of the models we found could identify each of the artifacts specified here and separate them such that their relative intensities were maintained. Most models were designed for one component [11; 91; 13], we found one model designed to extract two-components in [12]. Their two-component model focused on extracting Raman peaks and baselines. Few of the models utilized data normalization, a common processing step in ML practice. Normalization is used to specify the signal intensities which the models were trained for, thus guaranteeing stability for a defined input range. Instead, the existing models were trained to process raw signals which have no clear upper or lower limit regarding signal intensities.

We identified a need for DL methods capable of extracting each of the spectral components. Our aim was to develop DL models capable of processing spectra of variable length while maintaining component integrity. We hypothesized that through framing the learning problem as an extraction task of each component, we could ensure the explainability of the model predictions through component analy-

sis. This would enable the user to see exactly how much of each feature intensity is allocated to either B, CR, N or P. To enable input length variability, we elected to develop our models without the use of fully connected layers which require set input size. This meant we could not rely on the ability of fully connected layers for global feature utilization throughout the input. The scope of feature usability within our models would be bounded by the sizes of the combined convolution kernels. Along with this challenge, we aimed to make the models as small as possible without compromising on model performance compared to the established models which contained millions of parameters. Furthermore, we train our models on min-max normalized data, stabilizing the training process and guaranteeing consistent performance on all spectra within the normalized domain. Training a model capable of comparable performance to other state-of-the-art models required large datasets of diverse spectra. With this in mind, we developed a diverse synthetic spectrum generator to remove the need for gathering large training sets.

3.1.1 Synthetic Data Generation

We designed our Raman spectrum generator based on observable characteristics of spectra within our glioma dataset. We also examined other available datasets of Raman spectra visually to gain a broader perspective of the possible shapes seen in real datasets. We utilized mathematical modeling of the different components we aimed to generate. In our generator, Raman spectra are represented as the sum of four independent vector components. A spectrum (S) is the sum of B, CR, N and P formally expressed as

$$S = B + CR + N + P.$$

To stabilize the training and use of our models, we resolved to use normalization to limit the maximum and minimum elements in the component vectors. This is made possible by dividing the maximum of the spectrum sum by each component vector except N. N is excluded from the normalization to allow for greater noise fluctuations in the spectra. The final formula for the generated vector is then given by

$$S = \frac{(B + CR + P)}{\texttt{max}(B + CR + P)} + N.$$

To generate B, we utilized the work presented by [103] for spectral augmentation during training of DL models. We modified their suggested parameters to stabilize the behavior of the proposed functions on the interval between 0 and 1. The functions suggested for baseline representation were the linear, sinusoidal, polynomials of the nth degree, Gaussian and Lorentzian functions. These were also utilized in the generator designed by [13]. To increase the complexity of our generated B, we

generate a random candidate of each function type and combine their scaled representation linearly. The generator first produces a list of values $\alpha \in \mathbb{R}^5, 0 \leq \alpha_i \leq 1$ such that $\sum_{i=1}^{5} \alpha_i = 1$. The candidate functions responsible for B are then generated as

$$f_1(x) = (xa_1 + b_1),$$

$$f_2(x) = sin(a_2 x + b_2),$$

$$f_3(x) = (\sum_{i=0}^{10} a_{(3,i)} x^i),$$

$$f_4(x) = a_4 e^{\frac{-(x-b_4)^2}{2c_4^2}},$$

$$f_5(x) = \frac{1}{\pi} \frac{a_5}{a_5^2 + (x - b_5)},$$

which are then summed using their respective scaling values as

$$B = \sum_{i}^{5} \alpha_i f_i(x).$$

The parameter values a_{1-5}, b_{1-5} and c_4 were randomly generated for each function to add further variety among the B candidates. While we provide default ranges for the random values assigned to all parameters, we found that their selection was largely arbitrary.

CRs were generated as random, singular large intensity values over random spectral frequencies. The intensities were drawn from a uniform distribution $U(0.1, 1)$. N was generated as normally distributed noise on each frequency using the normal distribution $N(\mu = 0, \sigma \sim U(0.1, 0.3))$. We generate a random number of peaks in the spectra by random choice between Gaussian and Lorentzian curves and Voigt profiles; all of which mimic the shapes of Raman peaks. The formulas for these, containing the randomized parameters a and b are expressed by

$$P_{\text{Gauss}}(x) = e^{-x^2},$$

$$P_{\text{Lorentz}}(x) = \frac{1}{\pi} \frac{a}{a^2 + (x - b)^2 + \epsilon},$$

$$P_{\text{Voigt}}(x) = \frac{Re[\frac{x+ib}{\sqrt{2b}}]}{a\sqrt{2\pi}}.$$

3.1.2 Optimizing RADAR for Spectrum Correction

We developed two DL models in parallel to separate the components of S. Both models trained to optimize the sum of two objective functions: the root mean square error and a modified version of the R^2-score. We modified the R^2-score by adding 1 and multiplying the entire expression by -1, thereby transforming the score to have a global minimum of 0. These metrics are formulated as

$$\texttt{RMSE}(y, p) = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - p_i)^2},$$

$$\texttt{R}^2(y, p) = \frac{\sum_i^N (y_i - p_i)^2}{\sum_i^N (y_i - \mu_y)^2}.$$

The loss function we optimized to train RADAR: $L(y, p)$ is defined by their addition:

$$L(y, p) = \texttt{RMSE}(y, p) + \texttt{R}^2(y, p).$$

The architectures of the RADAR models are designed to enable processing of spectra with a variable number of frequencies and to extract the components of the given spectra while maintaining their relative intensities. The resulting outputs can then be added together to reconstruct the input spectrum. This can be done to validate the model predictions and establish how each Raman frequency was divided among the components. To accomplish this, we found that the models benefited from residual layer connections and a layer block aimed at encoding the global characteristics of the entire input spectrum. The RADAR models consist purely of convolutional layers, which allows for unrestricted input sizes, since the activations of each layer are carried out by kernel convolutions. Of the two models developed in RADAR, the smaller model designed to iteratively extract B and CR followed by P and N achieved superior performance over its sister model by allocating half of the parameters to extracting B and CR. Let $B_{pred}, CR_{pred}, N_{pred}$ and P_{pred} denote the model predictions for B, CR, N and P respectively, the model can then subtract them from the input, creating the latent representation $(S - (B_{pred} + CR_{pred}))$. The model then extracts N and P from the residual. This forces signal transformations such that the predicted components satisfy the equivalence $P_{pred} = S - (B_{pred} + CR_{pred} + N_{pred})$. The second model aims to extract B, CR and P at once, which are then subtracted from the input spectrum to produce the residual N_{pred} with the goal to minimize the difference $N - N_{pred}$.

Since no other model solution exists to extract the components of Raman spectra in this manner, we resolved to evaluate RADAR on one component at a time. This allowed us to perform detailed comparisons with established baseline correction methods ([12; 85; 86]), cosmic ray removal algorithms ([87]) , denoising strategies

([88; 91]) and peak extractors [11; 12]. In doing so, we also compare RADAR to competing DL methods for the purpose of correcting Raman spectra. Our results showed that RADAR was able to outperform all methods on every metric for extracting each component in Raman spectra. Furthermore, we applied RADAR and the model by [12] to Raman spectra from one sample, extracted at different exposure times. The resulting peak regions retained signal integrity by using all models while decreasing the extraction time to 0.5 seconds (a 90% increase from the maximum extraction time samples with 5 seconds of exposure time).

For evaluation on labeled data where the components had been generated or manually extracted, we compared the methods by using the SNR and *maximum error* metrics. SNR was primarily used to evaluate the denoising abilities of our models while the *maximum error* was used to evaluate the extraction of the other components. The *maximum error* was useful for identifying the largest error between the ground truth and the extracted components. It is expressed as

$$\textit{maximum error} = \max_{1 \leq f \leq N} |Y(f) - Y'(f)|.$$

The metric returns the largest error between components Y and Y' by comparing each wavelength between 1 and N.

We trained both RADAR models for 12 epochs, with 200,000 spectra generated in each epoch. In total, 240 million spectra were generated during training. The models were trained on Nvidia Titan X GPUs for approximately 5 days. The in-depth details of our evaluation and training setup are outlined with greater detail in Paper II.

3.2 Identification of Tumor Areas

The ratio of signal intensities between spectra in a single tumor can be processed to display the microscopic tumor environment. For example, principal component analysis (PCA) can be applied on sample spectra to compute a dimensionality-reduced representation of spectra in 3 dimensions. We found that the variance between spectra was maintained to adequately visualize the tumor environment by treating the 3 computed components as color vectors. By normalizing them to a range [0, 1], they could be treated as image data. For example, sample HF-442 is visualized using this method in Figure 1.

Utilizing this method shows areas of drastically different Raman spectra. Through manual validation, we discovered that these spectra were non-tumor spectra, originating instead from a number of different factors (blood, necrotic areas, substrate due to this tissue, heated sample spots, etc.). Through unsupervised clustering, the spectra possessing deviating patterns can be flagged as outliers (Figure 1a.). These outlier spectra obfuscate the genetic heterogeneity due to their radically different patterns as

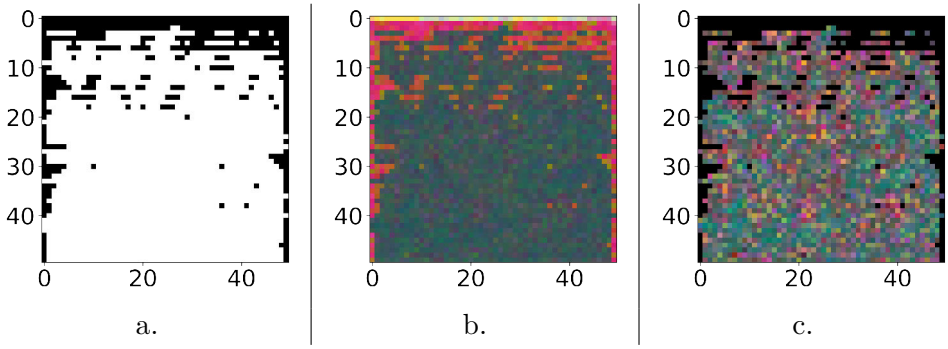

Figure 1. The tumor surface of sample HF-442 throughout our processing pipeline. a. The tumor surface (white) and non-tumor spectra (black) as separated by the DBSCAN algorithm. b. The tumor surface with all spectra present, colors are dimensionality reduced points as computed by PCA. c. The diversity within the tumor environment becomes clear after removing non-tumor spectra. PCA was applied to the tumor spectra exclusively to produce the color map of the surface.

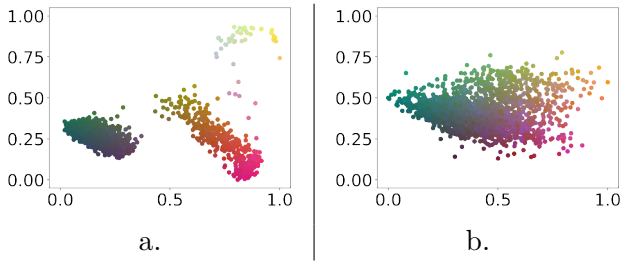

Figure 2. PCA can differ between drastically different spectra. a. Non-tumor spectra are well separated from tumor spectra (colors are consistent with Figure 1b.) The tumor spectra (darker spots) are linearly separable from the non-tumor spectra (multi-colored clusters). b. PCA cannot show separations between clusters when computing 2 components. The colors are consistent with Figure 1c.

seen by PCA (Figure 1b.). By removing these spectra and re-computing the principal components we get a clearer mapping of the tumor environment (Figure 1c.). In Figure 2 dimensionality reduction to only 2 components is enough to separate tumor from non-tumor spectra (Figure 2a.). To adequately visualize the tumor environment and curate the samples, we set out to remove these spectra using unsupervised clustering. By removing outlier spectra, we were able to focus our continued work exclusively on tumor spectra. The scatter plot of the dimensionality reduced tumor spectra is shown in Figure 2b.

Since all spectra contain 1738 features and an unresolved number of different kinds of non-tumor spectra were present in the dataset, we utilized the DBSCAN algorithm to cluster the data. DBSCAN [104] proved suitable for this task due to its ability to compute the optimal number of clusters based on the number of groups identified in its computations. This made the method favorable in comparison to the

agglomerative clustering, which requires the number of clusters as a hyperparameter. DBSCAN is also capable of forming density-based clusters rather than the centroid-based clusters used in K-means. Through a grid search for parameter optimization, we found that the epsilon-parameter (responsible for setting a maximum allowed distance between cluster elements) was optimal on L2-normalized spectra when using a value of 0.28. The min_samples argument, responsible for setting the minimum required number of points which constitutes a cluster, was set to 10. This method allocated many non-tumor spectra into an outlier class, represented by spectra which are too far away from established clusters to become cluster members.

We evaluated the outlier detection of DBSCAN using the silhouette index [105], a metric designed to validate cluster belonging through dissimilarity computation. Given a data point i and a cluster G, the dissimilarity w.r.t. i is computed by

$$d(G_i) = \sqrt{\sum_{g \in G} (i_r - g_r)^2}.$$

The silhouette index computed for $i, s(i)$ using two clusters A and B is formalized as

$$s(i) = \frac{d(B_i) - d(A_i)}{\texttt{max}(d(A_i), d(B_i))}.$$

Using the silhouette index we verified that tumor spectra contained silhouette indices ≥ 1, indicating good cluster belonging for their assigned cluster. Outlier spectra gained silhouette indices close to 0 which indicated no proper cluster belonging was possible in the identified cluster. This maintained that outliers were correctly removed from clusters which had good internal cluster similarity.

The performance of DBSCAN is fully dependent on the data and its hyperparameters. For novel datasets where non-tumor spectra may be present, DBSCAN would need to be reapplied which potentially includes finding the optimal parameters again if the previous hyperparameters are unsuitable for the novel samples. To avoid this, we trained a Random forest model on the data to create a classifier capable of discriminating between tumor and non-tumor spectra. This classifier was trained using class-weights to address the imbalance between tumor spectra (250,000) and non-tumor spectra (50,000). Our classifier managed to achieve 99% accuracy on the classification problem, providing suitable classifying performance on unseen glioma samples. We trained the Random forest model by using 5-fold cross-validation. We merged tumor samples originating from the same patient to avoid bias towards potential patient specific cancer patterns. Our results on identifying tumor spectra are covered in Paper I. The global median of the entire dataset of Raman spectra, along with the medians of the tumor spectra and outlier spectra are displayed in Figure 3.

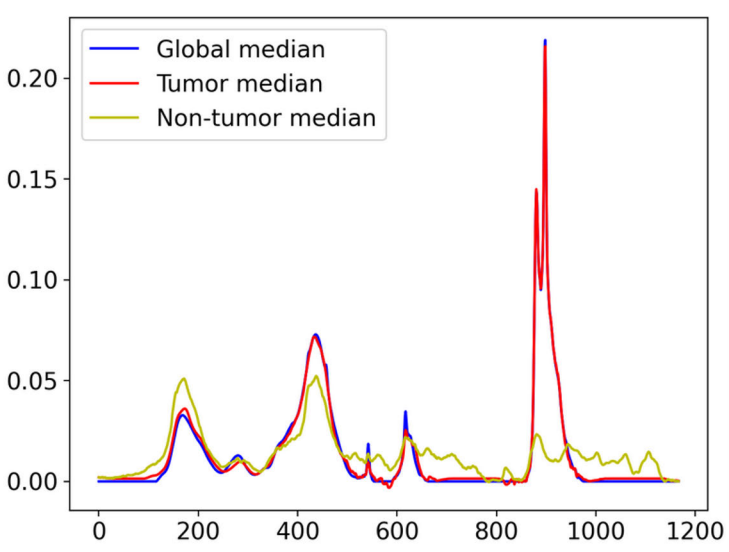

Figure 3. The medians of spectra within the glioma dataset. In blue is the median spectrum of all available spectra. In red is the median spectrum of the tumor spectra as identified by DBSCAN. In green is the median of the outlier spectra which were removed from further analysis.

3.3 Reducing the Batch Effect

The batch effect is a prevailing problem within projects aimed toward analyzing large biological datasets despite careful data management and preparation. This is especially vital within the medical context as the batch effect risks producing biased models. With the potential for optimized analysis promised by machine learning, diverse datasets are often combined to create larger datasets which enable the training of more complex DL models. For data analysts taking on tasks outside their field of expertise, lack of knowledge about the intricate field-specific data management practices can lead to batch effects occurring in their analyses. The batch effect is often non-trivial, leading to promising results in machine learning projects through reliance on patterns originating from acquisition-specific practices rather than relevant patterns. The ultimate problem stemming from the batch effect is the release of models with high accuracy that become unreliable on novel datasets.

In machine learning, there exist different tools to limit the batch effect from influencing the final model performance. One such method is the leave-one-patient-out cross-validation strategy. This paradigm alleviates batch effects in datasets by allocating entire samples of similar origin into the validation set, separated from the other training samples. In a medical context, this can avoid the batch effect in datasets where samples are highly heterogeneous. If the trained model manages to attain high accuracy on all validation samples in this paradigm, there is some certainty that the model is able to identify patterns relevant to the classification task. In

cases where data imbalance is of concern, leave-k-samples-out cross-validation can be used to construct balanced validation folds. This method allows for analysis of balance in model predictions, ensuring that the models do not suffer from skewed prediction biases. However, to address the batch effect on data from different acquisition sites, the validation split must take said sites into account. This is done in addition to the patient split to avoid bias towards site-specific patterns and study specific patterns.

Another way of limiting the batch effect lies in normalization methods and other preprocessing methods for data. For example, min-max normalization forces data points to assume a set minimum and maximum signal, removing relative signal intensities between points which could potentially be utilized to discriminate between two subsets of the dataset. Removal of environmental factors can also alleviate the batch effect. Our RADAR models function as methods for reducing the batch effect in this sense by removing baselines, cosmic rays and noise which could potentially be used to discriminate between tumor samples. However, despite these methods potentially removing the batch effect from data, there is currently no generalized method for guaranteed batch effect removal.

A promising method for determining batch effects can be seen in feature extraction and feature importance methods. To establish whether or not the batch effect influences model accuracy, model explainability can be leveraged to analyze feature importance according to the models themselves. This can help determine if models utilize relevant features for their predictions or if their attention is on proven irrelevant features. However, this relies on established feature relevance before feature importance is computed.

In the dataset available through the APOLLO project (Paper I) , we identified the ability to predict individual tumor samples based on single spectra extracted from them. This was confirmed through a 10-fold cross-validation training experiment. In each fold, a random selection of spectra from each patient sample was separated into a training set to train a DL model. The remaining spectra were allocated into a validation set. The training task of each model was to classify the patient and the tumor category given one spectrum. The resulting validation accuracy for patient identification, taken as the mean balanced validation accuracy across all 10 folds, was 68%. This confirmed the presence of bias within the dataset, enabling identification of patients based on isolated spectra. The IDH mutant classification tasks presented high accuracy across all validation folds (85%), likely caused by the batch effect.

Our contribution for reducing the batch effect in this context is in the form of an adversarial training solution which learns feature importance while reducing the intensity of features responsible for the batch effect in Raman spectra. We designed a novel feature learning layer capable of scaling the spectral intensities according to their contribution to model predictions. In parallel, our layer also learns to scale down features which are unimportant for the classification task. The layer contains

as many learnable parameters as the number of frequencies in the training spectra. Each learnable parameter functions as a multiplier to their respective input feature, meaning they scale the intensity according to feature contribution. The features are also reduced by L_1-regularization on the parameters which we found to work well for removing unimportant features in our datasets. The L_1-norm is a stronger option for forcing parameter minimization when their values are between 0 and 1 compared to L_2-regularization. The L_1-norm is computed as the sum of absolute parameter values $X_{L_1} = \lambda \sum_{x \in X} |x|$ which is bounded by the parameter λ. This regularization method is effective for forcing a decrease in values while λ can be set to function as a lower bound for the learnable parameters. The computation performed by the layer is expressed as the Hadamard product

$$f(x) = \frac{1}{m}(x \odot W),$$

where W is the vector containing parameters learning feature-wise importance. The parameter m is a maximum scaling factor, included to stabilize the feature scaling of the product $x \times W$. During training, it is set to be a combination of the maximum value of x and the previous values. Formally, it is a linear interpolation $m_t = m_{t-1} \cdot n_t + \texttt{max}(x) \cdot (1 - n_t)$ where t is the current training epoch and n is a list of floating-point values enumerated by t. Let $\epsilon \in \mathbb{R}^+$, then n has the following property: $\forall_{n \in [0,1]} : n_t - n_{t-1} = \epsilon$. The interpolation guarantees that m converges at a stable real value which scales the feature edited input.

We applied this layer to a model built to produce predictions for IDH mutant biomarker and the tumor id labels. The feature importance layer was attached to the input of the model with its parameters locked from gradient updates, the model was then trained to increase accuracy on both targets for one epoch. To decrease the batch effect, we then locked all model parameters and unlocked the parameters of the feature importance layer. We then trained again, this time forcing the model to keep increasing the IDH mutant accuracy while reducing the accuracy of the tumor id predictions. This forced the feature importance layer to adapt its parameters such that they transformed the spectral features to remove the batch effect while maintaining biomarker accuracy. The features which were deemed important by this layer, when trained using our adversarial strategy with 10-fold cross-validation, are shown in Figure 4.

According to our layer, the most important features for determining IDH mutant status are aligned with peaks found on the first half of the Raman spectrum. The layer also manages to recognize the insignificance of features with minimal variation around the silent region (the area around index 1000). The adversarial loop was repeated for 1000 epochs using learning rate decay to force model convergence. Our final results showed that we managed to reduce the batch effect from 68 $\%$ to 12% while maintaining high IDH mutant vs. wild-type accuracy. This endeavor, along

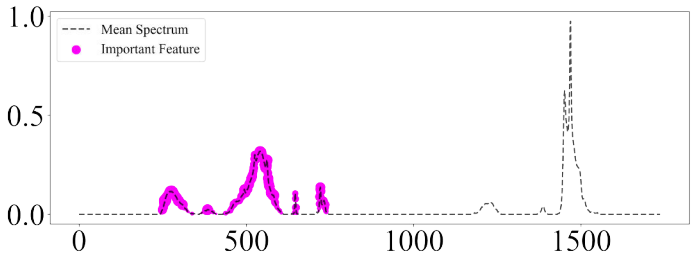

Figure 4. The most important features determined by our feature importance layer are displayed as magenta colored dots over the mean spectrum of the training set. The size of the dots correlates with the computed importance of the marked features.

with more detailed results, is covered in Paper III. One limitation of the study is the reliance on a single dataset. The next step for evaluating this strategy is to combine multiple datasets to evaluate the batch effect, and reduce it. By including more data from different acquisition sites, the method will be further optimized to enable learning tasks with larger combinations of datasets.

4 Glioma Predictions Based on Molecular Fingerprints

The classification pipeline developed during this project consists of a RF model, trained on the classification problems to produce feature importance scores for the spectral frequencies. The 20 most important features are then extracted based on the computed RF scores to train a support vector classifier (SVC) for the final spectrum classification. We found that this pipeline marginally outperformed DL models tested in parallel for the same classification problems. To evaluate common methodologies on our dataset we elected to utilize established methods for preprocessing the tumor spectra rather than applying RADAR. After training our models on the classification problems, we utilized RADAR and reduced the batch effect to retrain the classification pipeline. This was done to compare their performances and evaluate the benefits of our methods against established literature methods.

To avoid bias among the tumors originating from the same patients, we combined the spectra from tumors belonging to the same patient. This created a grouped dataset consisting of 46 patient samples, each containing thousands of Raman spectra. Our binary classification models were trained using a stratified 5-fold cross-validation setup. This setup guaranteed that the number of samples per category remained balanced in both the training and validation sets. Furthermore, cross-validation was essential to reduce the batch effect in the data, which we showed could be leveraged to predict patient ids given singular spectra in Paper III.

The dataset used in this section consists exclusively of the tumor spectra identified through the DBSCAN algorithm. We preprocessed the data by utilizing the airPLS algorithm [85] on the separated non-silent regions in the data. The silent regions were manually removed from the spectra by concatenating the non-silent regions, resulting in spectra containing 1167 frequencies. The remaining spectra contained 67% of the total number of spectral frequencies (1738). We then used the ReLU activation on the data to remove negative intensity values and normalized the spectra utilizing L2-normalization. We computed the F1-score, area under the receiver operating curve (AUROC), balanced accuracy and confusion matrices of each fold to evaluate model performance on the validation sets. The final metrics reported in this work, are the means of each metric across each validation fold. Of all samples included in this study, one sample (HF-1887) contained a baseline that airPLS failed to properly remove, meaning the sample contained patterns discernible from other

sample spectra. We elected to remove it from the training experiments to prevent the unique patterns from inducing bias during model training. Utilizing labels consistent with the methylation clusters from [71], we solved the following classification problems:

IDH-mutation Classification. We trained our model pipeline to predict the IDH mutant vs. wild-type status of measurement spots based on a single Raman spectrum. All spectra in our dataset are associated with a binary label indicating whether or not IDH mutation has occurred in their respective tumor. This classification setup is well balanced between the binary categories, which reduces the risk of bias due to label frequencies.

G-CIMP-low and G-CIMP-high Classification in IDH-mutant Tumors. We extracted a subset consisting of tumors with G-CIMP-low (LGm1) and G-CIMP-high (LGm2) characteristics from the IDH mutant cohort. Our ML pipeline was then trained to discriminate between G-CIMP-low and G-CIMP-high. To balance the dataset, we use the built-in data balancing functionality of both RF and SVC models provided by the scikit-learn library [106].

4.1 IDH Status Classification

IDH mutation or wild-type status is a label encompassing the entire dataset, allocating each spectrum into one of two categories. We elected to begin with this task due to the availability of all glioma samples in two balanced cohorts. We discovered that our models were capable of identifying the IDH-mutant and wild-type statuses of the samples with satisfactory consistency. The overall predictive performance of our pipeline was 0.81 AUROC taken as the average between the AUROC metrics of each fold. The average balanced accuracy from all folds was 78%. This showed that individual Raman spectra can be used to perform overall accurate predictions on entire tumors using single measurement spots despite the influence of FFPE slide influence. However, some samples proved difficult for the models to predict correctly when allocated to the validation set. We discovered that certain samples included heterogeneous spectra which caused their tumor surfaces to vary in certain areas. We hypothesized that this occurred due to genetic heterogeneity possessed by the tumor environment. Severe heterogeneity in our tumor samples could potentially invalidate our model results by reducing model prediction confidence and variability between learned functions across the validation folds.

To test our hypothesis and evaluate the severity of the heterogeneity, we inspected the predictions of all trained models on every sample across the validation folds. We displayed their surface predictions across each fold and visually determined the effect heterogeneity had on the training of each model with respect to each sample.

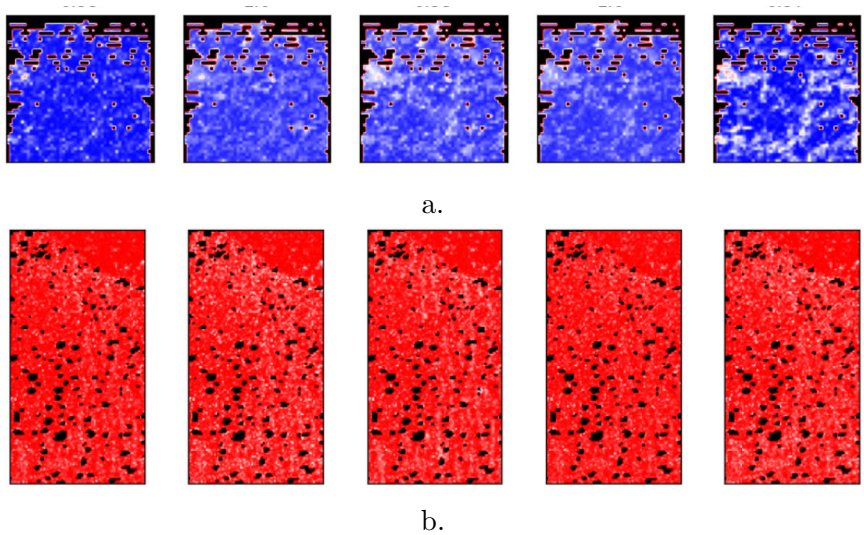

Figure 5. Surface predictions for IDH mutant vs. wild-type predictions in two independent patient tumor samples. The red color corresponds to mutant status while blue corresponds to wild-type. The color intensity indicates model certainty as predicted through the softmax output activation. White color indicates approximately 50% chance for either category and indicates model uncertainty. Black spots indicate the presence of non-tumor spectra, which are ignored by the model and evaluation process. Sample HF-442 is displayed in a. and sample HF-3271 is displayed in b.

For models with conservative numbers of parameters, we expected that the learned functions by the models were approximately similar and thus produce consistent predictions. In Figure 5, two tumor samples are shown with uniform and consistent prediction mappings; first row: HF-442 (IDH wild-type) and second row: HF-3271 (IDH mutant).

The black spots correspond to non-tumor spots, removed to focus the analysis exclusively on tumor spectra. Red color corresponds with IDH mutant predictions while the blue colors correspond with wild-type predictions. Model uncertainty is expressed by predictions with close to 50% probability of either class and is represented by the white color. The variation between model predictions is minimal, mainly spiking in cases where the sample is allocated to validation. We concluded that while heterogeneity is present in our dataset, it is not severe enough to prevent IDH mutant vs. wild-type predictions.

To explain the behavior of the SVC models producing the final outputs in our pipeline, we examined what features were used by each fold model. The feature importance scores are displayed in Figure 6

The 20 most important features appear within similar regions with few exceptions across all folds. This indicates consistency across all models, as the SVC mod-

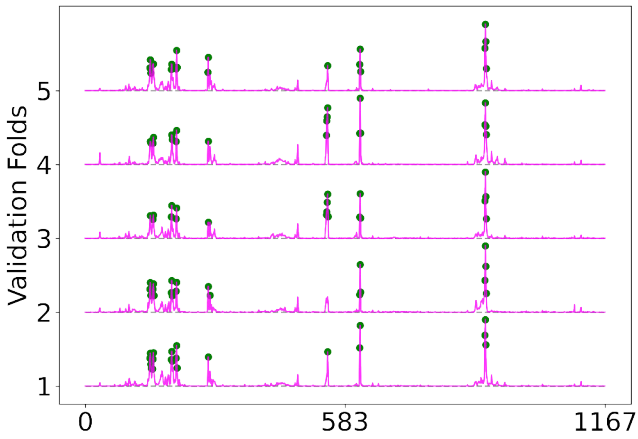

Figure 6. The feature importance vectors computed by our RF models sorted according to their respective validation folds. The vectors are represented by the magenta-colored lines. The 20 most important features, utilized by the SVC models are marked with green points.

els use similar features across the folds to produce their predictions. Based on this, we can pinpoint which frequencies are relevant for determining IDH mutant or IDH wild-type status of tumor spots.

4.2 G-CIMP-low vs. G-CIMP-high Classification

The subtypes G-CIMP-low and G-CIMP-high are subsets of the IDH mutant cohort. As such, the dataset consisting of spectra possessing these properties exclusively is significantly reduced from the previous classification problem. Training on unbalanced datasets poses two problems. The accuracy metric can become skewed, showing high accuracy by learning to predict the majority class exclusively. Additionally, the imbalanced frequency may lead to models rewriting changes made for the minority class due to frequent updates in favor of the majority class. To avoid this, we train our model pipeline by using balanced data batches.

The average model metrics of our pipeline reached AUROC 0.76 and a balanced accuracy of 75%. We attribute the lower AUROC and accuracy metrics to the reduced dataset size relative to the IDH mutant classification problem. Like in the previous section, we evaluated the presence of intra-tumor heterogeneity within the G-CIMP samples. We found that the heterogeneity was indeed expressed by the variance of our model surface predictions. Sample HF-3271, which was correctly predicted and displayed in Figure 5b, is again displayed in Figure 7. In this iteration, the surface appears harder to categorize as either uniquely G-CIMP-low or G-CIMP-high.

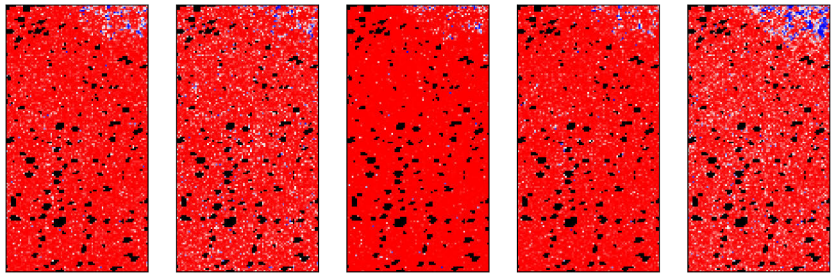

Figure 7. Surface predictions for sample HF-3271 for each of the 5 validation folds. Blue indicates G-CIMP-low predictions and red indicates G-CIMP-high predictions. Lower color intensity corresponds with higher prediction uncertainty with white indicating high uncertainty. The black spots indicate positions of non-tumor spectra removed from the analysis.

The appearance of blue spots in the sample's upper-right corner occurs consistently across all validation folds. This shows high model certainty despite it being trained to allocate all spectra to the same class. It is instead influenced by other tumor samples, leading to diverse surface predictions. This occurrence indicates a higher degree of heterogeneity than in the previous case. We derived a hypothesis that this could indicate the presence of both G-CIMP-low and G-CIMP-high status in a single tumor, which requires external validation for future experiments to be conducted. The entire process along with feature extraction for these classification tasks is outlined in Paper I.

As part of our modeling pipeline, we compute the 20 most important features in line with the IDH classification case. The feature vectors computed for each fold are shown in Figure 8.

Like in the previous section, we note that the area where the 20 most important features are concentrated appears consistent with few exceptions across all folds. Compared to the features computed for IDH classification, fewer points are located towards the left ends of the spectra. The majority of important features appear around the area of the biggest peak in the dataset and the middle peaks also appear relevant for determining G-CIMP status.

4.3 RADAR and Batch Effect Reduction Improves Classification

To further improve the accuracy of our classifiers we applied the methods introduced in Chapter 3 to the dataset and retrained our pipeline on the corrected dataset. For this approach, we replaced the airPLS algorithm meant to remove baseline signals from our dataset with the smaller of our two RADAR models and extracted the peak components from the spectra. To reduce the batch effect, we opted to use a stream-

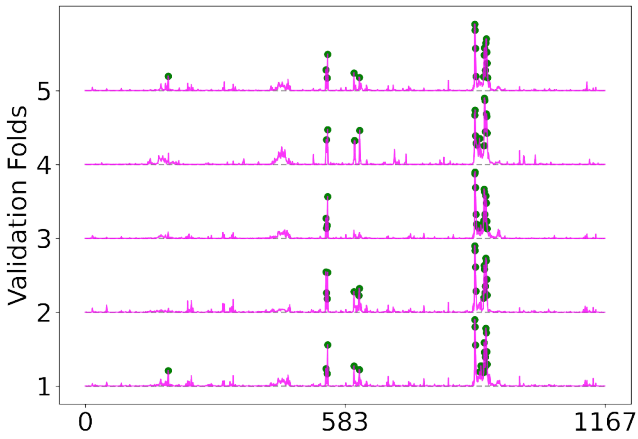

Figure 8. Feature importance vectors computed by the RF models for their respective folds are shown in purple. The 20 most important features utilized for training the SVC models are shown by the green dots.

lined approach for reducing data variability, making it harder for models to focus on simple patterns as categorization markers. This was done by utilizing a shallow CNN trained to predict the median spectrum of the training set given a single spectrum. We then used this model to encode both the training and validation sets before training our pipeline.

We found that the results from IDH mutant classification were consistent with those of our original setup. The accuracy of G-CIMP classifications drastically increased between the original approach and our novel methods. The balanced accuracy of the G-CIMP classification task increased from 75% to 82% and the AUROC metric increased from 0.76 to 0.90. The performance of each fold between the two methods are displayed in Figure 9.

The improvement in performance is mainly attributed to the second fold (displayed by the orange line in Figure 9a and Figure 9b.) while the other folds maintained consistent performance. The performance on fold 4 decreased using our methods as compared to the more conventional methods. It is possible that this is caused by the batch effect, present in the original approach in Figure 9a. This portrays the advantage of utilizing RADAR and minimizing the batch effect in training ML models for increased prediction accuracy.

The diversity of the tumors was also unveiled using our pipeline. Reducing the batch effect by moving each spectrum towards the global median made some samples harder to predict. Figure 10 shows the surface map of sample HF-3271 from each fold.

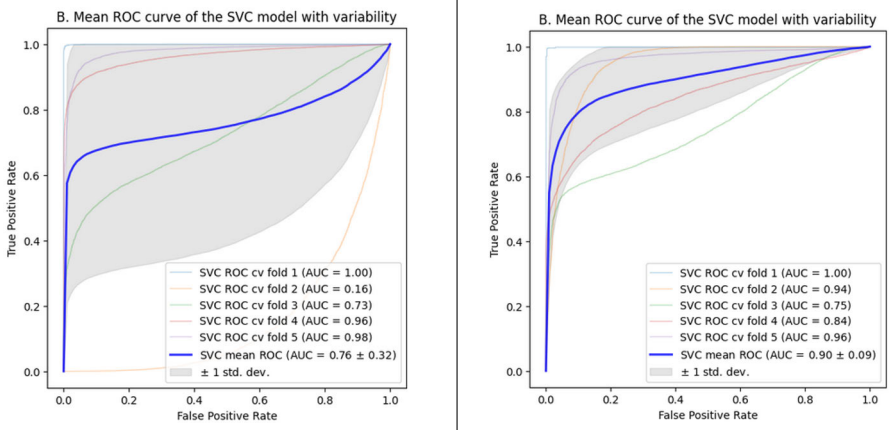

Figure 9. Comparison between our pipeline performance on G-CIMP classification using standard methods for preprocessing (displayed in a.) and RADAR along with reduction of the batch effect (displayed in b.)

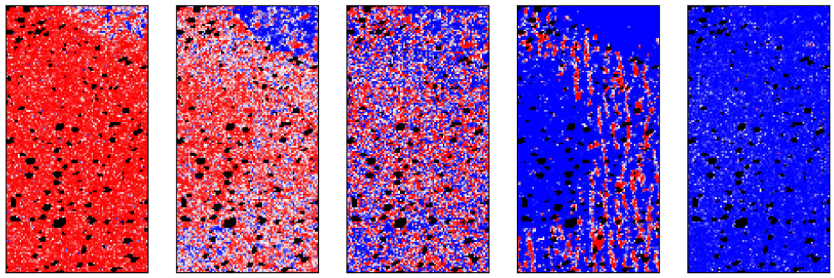

Figure 10. Surface predictions of the RADAR-processed sample HF-3271. The batch effect is reduced by utilizing a CNN designed to move the input spectra towards the global median spectrum of the training set.

The variance between surface predictions is drastically increased compared to our standard preprocessing pipeline. This is also true when analyzing the surface predictions in each fold. The high variance between areas inside the surface maps indicates tumor diversity which we hypothesize originates from genetic heterogeneity. By applying our methods, we have shown how our models can increase the accuracy of classification models on Raman spectroscopy data while also unveiling the diverse makeup of glioma tumors.

It is worth noting the limitations of the evaluation of our classification pipeline. We only had available a dataset based on 46 unique patient tumors and more would be needed for a more robust evaluation. Also, tumor samples from multiple different acquisition sites would be useful to evaluate model performance on potentially varying input signals. These are notoriously difficult problems in many ML-based Raman studies.

5 Discussion

In this thesis, we have presented ML models for preprocessing Raman spectra and glioma classification. We have demonstrated that our models can learn to link Raman spectra to cancer features characteristic of IDH mutant and G-CIMP status. For these classification problems, our models can efficiently learn to categorize the data distributions despite the heterogeneous nature of tumor-wide profiles. The heterogeneous patterns discovered in our work are captured by our model predictions, enabling surface analysis of glioma through model generalization. This exciting prospect has the potential to enable future projects for glioma categorization by utilizing RADAR and methods for reducing the batch effect. The pipeline simplifies the data patterns by reducing redundant signals through RADAR and boosting data uniformity by moving spectra towards the global median spectrum. Our models can indicate mutation within the IDH genes, utilizing one single Raman spectrum, providing an efficient method for determining glioma prognosis. Additionally, prognosis predictions can be specified further through our models designed for G-CIMP classifications. These results suggest Raman spectroscopy enables highly detailed classifications based on a single Raman spectrum. Furthermore, success on the classification tasks in this setup indicates the informative nature of Raman spectra, as they appear to encode the genetic properties of measurement spots in gliomas.

The heterogeneity of glioma genetics continues to complicate predictive modeling for tumor classifications. Our project has shown that ML can be applied to learn genetic patterns within Raman spectra given enough data. However, tumor-wide labels do not express the spot-by-spot variance required to identify the complete genetic profile of gliomas. To demonstrate this difficulty, we employ our pipeline to learn the differentiation between the three classes LGm1, LGm2 and LGm3. The prospect of recognizing 1p/19q-codeletion status (LGm3) based on the molecular fingerprint has tremendous value for glioma categorization. This would enable efficient recognition of oligodendrogliomas, a cancer possessing a more favorable prognosis compared to other gliomas lacking 1p/19q-codeletion. The three methylation classes are contained within the IDH mutant cohort. In contrast to other classification tasks presented in our work, this is a three-class classification problem with the LGm3 subtype containing fewer patient samples than the LGm1 and LGm2 classes. The heterogeneous spots present within the LGm3 tumors increase the complexity of the modeling tasks, which confuses the final predictions of our models within each fold.

This resulted in models with significantly reduced performance in a majority of validation folds compared to our previous experiments. In this trial, only two out of five folds achieved good validation accuracy (fold 1: 92% and fold 5: 83%). In the remaining folds, the models failed to generalize properly to the data (fold 2: 54%, fold 3: 45% and fold 4: 27%). Despite increased dataset size compared to the G-CIMP modeling problem, our models struggled to learn observable patterns. An example of the drastic increase in diversity among spot predictions was observed in sample HF-3271, shown in Figure 11.

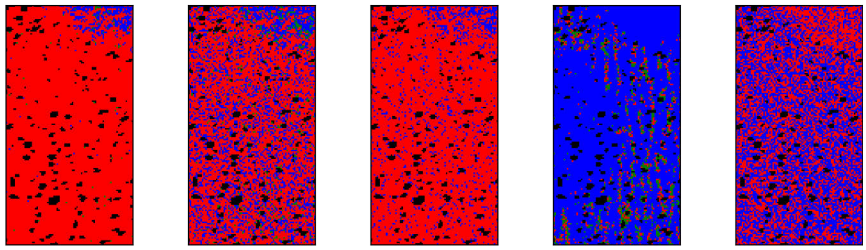

Figure 11. Surface predictions of sample HF-3271 for the classification problem concerning LGm1-3. Blue indicates G-CIMP-low, red indicates G-CIMP-high and green indicates 1p/19q-codeletion. Each column is represented by the corresponding models trained in the 5-fold cross-validation loop

While the diversity between folds is akin to those displayed in Figure 10, the performance in each fold is drastically decreased. This leads to uncertainty towards the prediction results. The surface prediction similarities to Figure 10 suggest that the models find consistent surface patterns. However, the true identity of each spot label is uncertain in this context. Folds fail to reach majority consensus among the spot predictions, limiting our ability to solidify proper spot labels. This problem persisted even in unsupervised clustering trials, where mini-batch K-means was applied to each cross-validation fold shown in Figure 12.

Establishing consensus based on mini-batch K-means clustering appears inefficient due to the lack of thorough consensus among all surface predictions. However, general areas consistent with our pipeline predictions appear to form, especially in the 2-cluster model. Our demonstration shows that while this problem is still relevant in our dataset despite multiple curation and preprocessing methods, general patterns are detected with varying clarity. This suggests ML can be applied to learn and utilize these patterns.

We hypothesize that the difficulty of model generalization on this problem lies in the genetic heterogeneity of gliomas. This hypothesis assumes that the heterogeneity is amplified when a third class (along with samples corresponding to it) is introduced. The dataset is either too small or the samples used during training are too diverse for the model to generalize its behavior for unseen patterns. Instead, the models forget learned behavior, preventing them from learning the dataset properly,

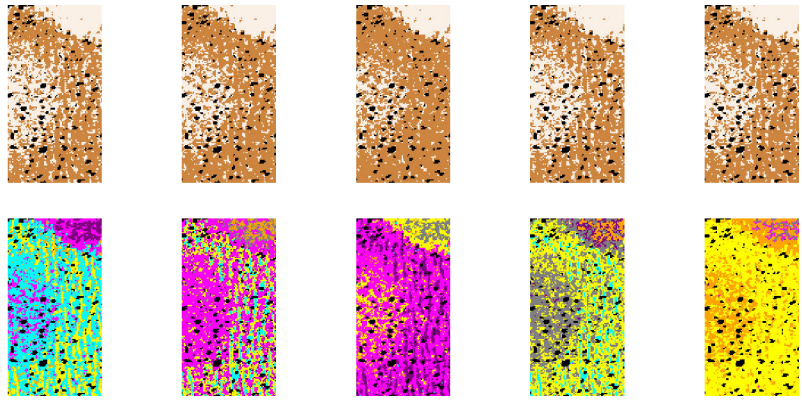

Figure 12. Unsupervised clustering via mini-batch K-means on sample HF-3271. The first row shows the 2-cluster model results of the tumor surface represented by the beige and brown colors. On the second row are the surface predictions of the 6-cluster model. Surface patterns become visible in both the 2-cluster and 6-cluster models. These patterns appear consistent with the patterns recognized by our classification model. The columns represent the cross-validation fold on which the clustering model was trained.

due to conflicting patterns. This problem is further exacerbated by including all six LGm classes in the dataset; despite a larger dataset size resulting from including all samples available, training resulted in validation folds with slightly above average accuracy (approximately 16% accuracy) for the six-class problem. As seen in Figure 10, this inconsistency among models appears despite promising model metrics. Therefore, it is recommended to train small models with a minimal number of learnable parameters to avoid overfitting.

The obvious solution to the uncertainty introduced by heterogeneity is to increase the dataset size by including more glioma samples. The drawback is that it requires further investment and efforts for data extraction and labeling. The second issue concerns the glioma heterogeneity, which implies the need for spot-wise labeling rather than tumor-wide categories. One potential solution to this is to deploy a learning strategy referred to as multiple-instance learning [107; 108]. This learning strategy could be adapted to the dataset by iteratively learning the data and identifying false positives with respect to the tumor label. Through this learning paradigm, training could progress until surface predictions stabilize, at which point the heterogeneity becomes apparent on each sample. In the future, we will look to apply multiple-instance learning to this dataset. Our aim is to identify consensus predictions on tumor spots between validation folds to derive tumor labels without renewed data extraction efforts.

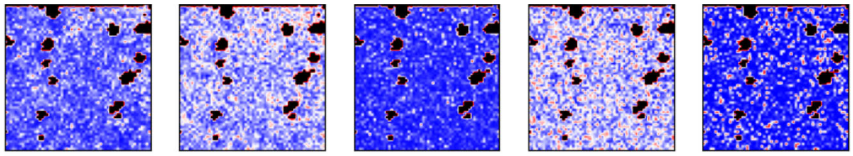

Figure 13. Surface predictions of the sample HF-1010 of the IDH mutant cohort. Each model trained across all folds in our pipeline predicts IDH wild-type states of the vast majority of the measurement spots.

To validate this approach, it is necessary to evaluate multiple models on the same sample. Consensus among model predictions on the sample surfaces could indicate that genetic patterns descriptive of gliomas are captured while variance among model predictions could indicate model overfitting. This assumes that the inclusion of more patient samples containing minimal heterogeneity will lead to stabilized model predictions. To identify which samples contain minimal heterogeneous patterns, we recommend utilizing leave-one-sample-out cross-validation which could show how well each sample performs given the training influence from all other samples. This could enable identification of samples containing minimal heterogeneity to include them in a dedicated training set. In our dataset, we recommend treating the more heterogeneous samples as uncertain and allocating them exclusively to the validation set. Sample heterogeneity can then be tested on these samples to determine if the surface patterns are consistent among model predictions.

Another anomaly within the dataset is sample HF-1010. Despite possessing IDH mutant status, sample HF-1010 had approximately all measurement spots categorized as IDH wild-type, regardless of whether it was included in the training or validation set. This could potentially indicate label noise within the dataset along with heterogeneity. The surface predictions in all folds are displayed in Figure 13.

The inconsistency between the assigned tumor label and the spot predictions for HF-1010 poses a problem that appears separate from tumor heterogeneity. Rather, it appears this sample has been erroneously labeled. Alternatively, the dataset is not large enough to teach the models to capture the appropriate patterns for glioma classification. In a test constructed using leave-one-sample-out cross-validation, we trained a model on all other samples and validated on HF-1010. Our results found that the loss was steadily increasing over the epochs, whereas the training loss decreased. The final predictive accuracy of the model was approximately 0%.

Deep learning is reliant on large datasets to properly learn patterns relevant for accurate predictions. Within the medical field, this is a task requiring significant resources and patient consent to use their data. Data remains the most substantial bottleneck within the field of machine learning; there is no upper limit on the desired number of data points. The size of 46 patient samples with a total number of over 300,000 spectra in our work is in fact quite large for a biological Raman dataset.

However, the need to separate samples into heterogeneous sub-regions to identify their genetic labels can quickly reduce dataset size. Despite collaborative efforts such as Kaggle competitions and The Cancer Genome Atlas (TCGA), amassing larger datasets still requires considerable time and resources. Instead of waiting for these datasets, innovative methods and learning strategies are necessary to enable training of deep and complex models using the limited data available today. Through our work, we have demonstrated that innovative methodology and thorough analysis can be performed with ML on limited data. Continued efforts building on this work has great potential for leveraging ML to unveil heterogeneous glioma characteristics, which can further our understanding and treatment strategies of the disease. The remaining uncertainties about tumor heterogeneity and how to properly capture it is an exciting avenue for future studies. Further directions of our projects include fine-tuning and improving the RADAR models, embedding methods for further reduction of the batch effect within Raman data and capturing heterogeneous patterns through semi-supervised learning. Raman spectra provide a descriptive lens through which ML can analyze biological data. This can be utilized to perform analysis in a rapid fashion provided efficient ML methods have been trained for the task. We believe further adoption of this technique can increase collaboration between mathematics and biology, promoting interdisciplinary studies for students within both fields.

List of References

[1] Jeroen T. J. M. van Dijck, Hilko Ardon et al. Survival prediction in glioblastoma: 10-year follow-up from the Dutch Neurosurgery Quality Registry. *J Neurooncol*, *volume 174*:753–764, 2025. ISSN 1573-7373.
URL https://doi.org/10.1007/s11060-025-05080-3

[2] Martin J. van den Bent, Marjolein Geurts et al. Primary brain tumours in adults. *The Lancet*, *volume 402*:1564–1579, 2023. ISSN 0140-6736, 1474-547X. Publisher: Elsevier.
URL https://doi.org/10.1016/S0140-6736(23)01054-1

[3] Yang Liu, Fei Zhou, Heba Ali, Justin D. Lathia and Peiwen Chen. Immunotherapy for glioblastoma: current state, challenges, and future perspectives. *Cell Mol Immunol*, *volume 21*:1354–1375, 2024. ISSN 2042-0226. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41423-024-01226-x

[4] J H Rees. Diagnosis and treatment in neuro-oncology: an oncological perspective. *British Journal of Radiology*, *volume 84*:S82–S89, 2011. ISSN 0007-1285.
URL https://doi.org/10.1259/bjr/18061999

[5] Mariana Afonso and Maria Alexandra Brito. Therapeutic Options in Neuro-Oncology. *International Journal of Molecular Sciences*, *volume 23*:5351, 2022. ISSN 1422-0067. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
URL https://doi.org/10.3390/ijms23105351

[6] Ben Kinnersley, Josephine Jung et al. Genomic landscape of diffuse glioma revealed by whole genome sequencing. *Nat Commun*, *volume 16*:4233, 2025. ISSN 2041-1723. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41467-025-59156-9

[7] Iman Karimi-Sani, Zahra Molavi et al. Personalized mRNA vaccines in glioblastoma therapy: from rational design to clinical trials. *Journal of Nanobiotechnology*, *volume 22*:601, 2024. ISSN 1477-3155.
URL https://doi.org/10.1186/s12951-024-02882-x

[8] Niclas Skarne, Rochelle C. J. D'Souza et al. Personalising glioblastoma medicine: explant organoid applications, challenges and future perspectives. *Acta Neuropathologica Communications*, *volume 13*:6, 2025. ISSN 2051-5960.
URL https://doi.org/10.1186/s40478-025-01928-x

[9] Breanna Mann, Nichole Artz et al. Opportunities and challenges for patient-derived models of brain tumors in functional precision medicine. *npj Precis. Onc.*, *volume 9*:47, 2025. ISSN 2397-768X. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41698-025-00832-w

[10] Alina Finch, Georgios Solomou et al. Advances in Research of Adult Gliomas. *International Journal of Molecular Sciences*, *volume 22*:924, 2021. ISSN 1422-0067. Publisher: Multidisciplinary Digital Publishing Institute.
URL https://doi.org/10.3390/ijms22020924

[11] Joel Wahl, Mikael Sjödahl and Kerstin Ramser. Single-Step Preprocessing of Raman Spectra Using Convolutional Neural Networks. *Applied Spectroscopy*, *volume 74*:427–438, 2020. ISSN 0003-7028. Publisher: SAGE Publications Ltd STM.
URL https://doi.org/10.1177/0003702819888949

[12] Mohammadrahim Kazemzadeh, Miguel Martinez-Calderon et al. Cascaded Deep Convolutional Neural Networks as Improved Methods of Preprocessing Raman Spectroscopy Data. *Analytical Chemistry*, *volume 94*:12907–12918, 2022. ISSN 0003-2700. Publisher: American Chemical Society.
URL https://doi.org/10.1021/acs.analchem.2c03082

[13] Qingliang Jiao, Xiuwen Guo et al. Deep learning baseline correction method via multi-scale analysis and regression. *Chemometrics and Intelligent Laboratory Systems*, *volume 235*:104779, 2023. ISSN 0169-7439.
URL https://doi.org/10.1016/j.chemolab.2023.104779

[14] Laurent James Livermore, Martin Isabelle et al. Rapid intraoperative molecular genetic classification of gliomas using Raman spectroscopy. *Neuro Oncol Adv*, *volume 1*:vdz008, 2019. ISSN 2632-2498.
URL https://doi.org/10.1093/noajnl/vdz008

[15] James G. Nicholson and Howard A. Fine. Diffuse Glioma Heterogeneity and Its Therapeutic Implications. *Cancer Discov*, *volume 11*:575–590, 2021. ISSN 2159-8274.
URL https://doi.org/10.1158/2159-8290.CD-20-1474

[16] Sabrina Xin Zi Quek and Khek Yu Ho. Artificial Intelligence in Upper Gastrointestinal Diagnosis. *Korean J Helicobacter Up Gastrointest Res*, *volume 25*:251–260, 2025. ISSN 1738-3331.
URL https://doi.org/10.7704/kjhugr.2025.0024

[17] Zozan Guleken, Paweł Jakubczyk et al. An application of raman spectroscopy in combination with machine learning to determine gastric cancer spectroscopy marker. *Computer Methods and Programs in Biomedicine*, *volume 234*:107523, 2023. ISSN 0169-2607.
URL https://doi.org/10.1016/j.cmpb.2023.107523

[18] Chenming Li, Shasha Liu et al. Combining Raman spectroscopy and machine learning to assist early diagnosis of gastric cancer. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *volume 287*:122049, 2023. ISSN 1386-1425.
URL https://doi.org/10.1016/j.saa.2022.122049

[19] Ya Zhang, Zheng Li et al. Employing Raman Spectroscopy and Machine Learning for the Identification of Breast Cancer. *Biol Proced Online*, *volume 26*:28, 2024. ISSN 1480-9222.
URL https://doi.org/10.1186/s12575-024-00255-0

[20] Sandryne David, Trang Tran et al. In situ Raman spectroscopy and machine learning unveil biomolecular alterations in invasive breast cancer. *JBO*, *volume 28*:036009, 2023. ISSN 1083-3668, 1560-2281.
URL https://doi.org/10.1117/1.JBO.28.3.036009

[21] Sanghwa Lee, Miyeon Jue et al. Early-stage diagnosis of bladder cancer using surface-enhanced Raman spectroscopy combined with machine learning algorithms in a rat model. *Biosensors and Bioelectronics*, *volume 246*:115915, 2024. ISSN 0956-5663.
URL https://doi.org/10.1016/j.bios.2023.115915

[22] Santosh Kumar Paidi, Joel Rodriguez Troncoso et al. Raman Spectroscopy and Machine Learning Reveals Early Tumor Microenvironmental Changes Induced by Immunotherapy. *Cancer Res*, *volume 81*:5745–5755, 2021. ISSN 0008-5472.
URL https://doi.org/10.1158/0008-5472.CAN-21-1438

[23] Xuecong Tian, Cheng Chen et al. Application of Raman spectroscopy technology based on deep learning algorithm in the rapid diagnosis of glioma. *Journal of Raman Spectroscopy*, *volume 53*:735–745, 2022. ISSN 1097-4555. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jrs.6302.
URL https://doi.org/10.1002/jrs.6302

[24] Jing Zhang, Yimeng Fan et al. Accuracy of Raman spectroscopy in differentiating brain tumor from normal brain tissue. *Oncotarget*, *volume 8*:36824–36831, 2017. ISSN 1949-2553.
URL https://doi.org/10.18632/oncotarget.15975

[25] Todd Hollon, Spencer Lewis, Christian W. Freudiger, X. Sunney Xie and Daniel A. Orringer. Improving the accuracy of brain tumor surgery via Raman-based technology. *Neurosurgical*

Focus, *volume 40*:E9, 2016. ISSN 1092-0684. Publisher: American Association of Neurological Surgeons Section: Neurosurgical Focus.
URL https://doi.org/10.3171/2015.12.FOCUS15557

[26] Gilbert Georg Klamminger, Karoline Klein et al. Differentiation of primary CNS lymphoma and glioblastoma using Raman spectroscopy and machine learning algorithms. *Free Neuropathol*, *volume 2*:26, 2021. ISSN 2699-4445.
URL https://doi.org/10.17879/freeneuropathology-2021-3458

[27] Leo Breiman. Random Forests. *Machine Learning*, *volume 45*:5–32, 2001. ISSN 1573-0565.
URL https://doi.org/10.1023/A:1010933404324

[28] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach Learn*, *volume 20*:273–297, 1995. ISSN 1573-0565.
URL https://doi.org/10.1007/BF00994018

[29] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. Deep learning. *Nature*, *volume 521*:436–444, 2015. ISSN 1476-4687. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/nature14539

[30] Mackenzie Price, Christine Ballard et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2017–2021. *Neuro-Oncology*, *volume 26*:vi1–vi85, 2024. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noae145

[31] Jakob V E Gerstl, Mackenzie Price et al. Years of life lost due to central nervous system tumor subtypes in the United States. *Neuro-Oncology*, page noaf142, 2025. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noaf142

[32] David N Louis, Arie Perry et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, *volume 23*:1231–1251, 2021. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noab106

[33] Adalberto M. Filho, Ariana Znaor et al. Cancers of the brain and central nervous system: global patterns and trends in incidence. *Journal of Neuro-Oncology*, *volume 172*:567–578, 2025. ISSN 1573-7373.
URL https://doi.org/10.1007/s11060-025-04944-y

[34] Tuomas Natukka, Jani Raitanen, Hannu Haapasalo and Anssi Auvinen. Incidence trends of adult malignant brain tumors in Finland, 1990–2016. *Acta Oncologica*, *volume 58*:990–996, 2019. ISSN 1651-226X.
URL https://doi.org/10.1080/0284186X.2019.1603396

[35] P. Wesseling and D. Capper. WHO 2016 Classification of gliomas. *Neuropathology and Applied Neurobiology*, *volume 44*:139–150, 2018. ISSN 1365-2990. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nan.12432.
URL https://doi.org/10.1111/nan.12432

[36] Patrick Y Wen, Michael Weller et al. Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro-Oncology*, *volume 22*:1073–1113, 2020. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noaa106

[37] Michael Weller, Stefan M. Pfister et al. Molecular neuro-oncology in clinical practice: a new horizon. *The Lancet Oncology*, *volume 14*:e370–e379, 2013. ISSN 1470-2045, 1474-5488. Publisher: Elsevier.
URL https://doi.org/10.1016/S1470-2045(13)70168-2

[38] Keyang Yang, Zhijing Wu et al. Glioma targeted therapy: insight into future of molecular approaches. *Molecular Cancer*, *volume 21*:39, 2022. ISSN 1476-4598.
URL https://doi.org/10.1186/s12943-022-01513-z

[39] Susan Costantini, Elena Di Gennaro et al. Glioblastoma metabolomics: uncovering biomarkers for diagnosis, prognosis and targeted therapy. *Journal of Experimental & Clinical Cancer Research*, *volume 44*:230, 2025. ISSN 1756-9966.
URL https://doi.org/10.1186/s13046-025-03497-2

[40] Aleksandr Shikalov, Igor Koman and Natalya M. Kogan. Targeted Glioma Therapy—Clinical Trials and Future Directions. *Pharmaceutics*, *volume 16*:100, 2024. ISSN 1999-4923. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
URL https://doi.org/10.3390/pharmaceutics16010100

[41] Alessia Pellerino, Mario Caccese, Marta Padovan, Giulia Cerretti and Giuseppe Lombardi. Epidemiology, risk factors, and prognostic factors of gliomas. *Clin Transl Imaging*, *volume 10*:467–475, 2022. ISSN 2281-7565.
URL https://doi.org/10.1007/s40336-022-00489-6

[42] Ruham Alshiekh Nasany and Macarena Ines de la Fuente. Therapies for IDH-Mutant Gliomas. *Curr Neurol Neurosci Rep*, *volume 23*:225–233, 2023. ISSN 1534-6293.
URL https://doi.org/10.1007/s11910-023-01265-3

[43] Hoon Kim, Siyuan Zheng et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.*, *volume 25*:316–327, 2015. ISSN 1088-9051, 1549-5469. Publisher: Cold Spring Harbor Lab.
URL https://doi.org/10.1101/gr.180612.114

[44] Karina Chornenka Martin, Crystal Ma and Stephen Yip. From Theory to Practice: Implementing the WHO 2021 Classification of Adult Diffuse Gliomas in Neuropathology Diagnosis. *Brain Sciences*, *volume 13*:817, 2023. ISSN 2076-3425. Publisher: Multidisciplinary Digital Publishing Institute.
URL https://doi.org/10.3390/brainsci13050817

[45] Andrew A. Hardigan, Joshua D. Jackson and Anoop P. Patel. Surgical Management and Advances in the Treatment of Glioma. *Seminars in Neurology*, *volume 43*:810–824, 2023. ISSN 0271-8235, 1098-9021. Publisher: Thieme Medical Publishers, Inc.
URL https://doi.org/10.1055/s-0043-1776766

[46] Xue Yang, Shibing Wang, Vedrana Montana, Xiangmin Tong and Vladimir Parpura. Status and Prospects of Glioblastoma Multiforme Treatments. *Journal of Neurochemistry*, *volume 169*:e70158, 2025. ISSN 1471-4159. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.70158.
URL https://doi.org/10.1111/jnc.70158

[47] Zhongxi Yang, Chen Zhao et al. A review on surgical treatment options in gliomas. *Frontiers in Oncology*, *volume 13*, 2023. ISSN 2234-943X. Publisher: Frontiers.
URL https://doi.org/10.3389/fonc.2023.1088484

[48] Nader Sanai and Mitchel S. Berger. Surgical oncology for gliomas: the state of the art. *Nature Reviews Clinical Oncology*, *volume 15*:112–125, 2018. ISSN 1759-4782. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/nrclinonc.2017.171

[49] Ilker Y. Eyüpoglu, Michael Buchfelder and Nic E. Savaskan. Surgical resection of malignant gliomas—role in optimizing patient outcome. *Nature Reviews Neurology*, *volume 9*:141–151, 2013. ISSN 1759-4766. Publisher: Nature Publishing Group.
URL http://doi.org/10.1038/nrneurol.2012.279

[50] Shawn L. Hervey-Jumper and Mitchel S. Berger. Maximizing safe resection of low- and high-grade glioma. *Journal of Neuro-Oncology*, *volume 130*:269–282, 2016. ISSN 1573-7373.
URL https://doi.org/10.1007/s11060-016-2110-4

[51] Elizabeth A. Maher, Frank B. Furnari et al. Malignant glioma: genetics and biology of a grave matter. *Genes & Development*, *volume 15*:1311–1333, 2001. ISSN 0890-9369, 1549-5477. Publisher: Cold Spring Harbor Lab.
URL https://doi.org/10.1101/gad.891601

[52] Pauline Latzer, Henning Zelba et al. A real-world observation of patients with glioblastoma treated with a personalized peptide vaccine. *Nature Communications*, *volume 15*:6870, 2024. ISSN 2041-1723. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41467-024-51315-8

[53] Alberto Delaidelli and Alessandro Moiraghi. Recent Advances in the Diagnosis and Treatment of Brain Tumors. *Brain Sciences*, *volume 14*:224, 2024. ISSN 2076-3425.
URL https://doi.org/10.3390/brainsci14030224

[54] Cameron W. Brennan, Roel G. W. Verhaak et al. The Somatic Genomic Landscape of Glioblastoma. *Cell*, *volume 155*:462–477, 2013. ISSN 0092-8674, 1097-4172. Publisher: Elsevier.
URL https://doi.org/10.1016/j.cell.2013.09.034

[55] Evgenia Bourkoula, Damiano Mangoni et al. Glioma-Associated Stem Cells: A Novel Class of Tumor-Supporting Cells Able to Predict Prognosis of Human Low-Grade Gliomas. *Stem Cells*, *volume 32*:1239–1253, 2014. ISSN 1066-5099.
URL https://doi.org/10.1002/stem.1605

[56] Tathiane M Malta, Camila F de Souza et al. Glioma CpG island methylator phenotype (G-CIMP): biological and clinical implications. *Neuro Oncol*, *volume 20*:608–620, 2018. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/nox183

[57] Houtan Noushmehr, Daniel J. Weisenberger et al. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, *volume 17*:510–522, 2010. ISSN 1535-6108.
URL https://doi.org/10.1016/j.ccr.2010.03.017

[58] Camila Ferreira de Souza, Thais S. Sabedot et al. A Distinct DNA Methylation Shift in a Subset of Glioma CpG Island Methylator Phenotypes during Tumor Recurrence. *Cell Reports*, *volume 23*:637–651, 2018. ISSN 2211-1247. Publisher: Elsevier.
URL https://doi.org/10.1016/j.celrep.2018.03.107

[59] Veronique G. LeBlanc and Marco A. Marra. DNA methylation in adult diffuse gliomas. *Brief Funct Genomics*, *volume 15*:491–500, 2016. ISSN 2041-2649.
URL https://doi.org/10.1093/bfgp/elw019

[60] Yingying Xu, Huashi Xiao et al. CIMP-positive glioma is associated with better prognosis: A systematic analysis. *Medicine (Baltimore)*, *volume 101*:e30635, 2022. ISSN 0025-7974.
URL https://doi.org/10.1097/MD.0000000000030635

[61] Sebastian Brandner, Alexandra McAleenan et al. Diagnostic accuracy of 1p/19q codeletion tests in oligodendroglioma: A comprehensive meta-analysis based on a Cochrane systematic review. *Neuropathology and Applied Neurobiology*, *volume 48*:e12790, 2022. ISSN 1365-2990. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nan.12790.
URL https://doi.org/10.1111/nan.12790

[62] Kurt A Jaeckle, Karla V Ballman et al. CODEL: phase III study of RT, RT + TMZ, or TMZ for newly diagnosed 1p/19q codeleted oligodendroglioma. Analysis from the initial study design. *Neuro Oncol*, *volume 23*:457–467, 2021. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noaa168

[63] Andrew B Lassman and Timothy F Cloughesy. Early results from the CODEL trial for anaplastic oligodendrogliomas: is temozolomide futile? *Neuro Oncol*, *volume 23*:347–349, 2021. ISSN 1522-8517.
URL https://doi.org/10.1093/neuonc/noab006

[64] E. Franceschi, A. Mura et al. Low grade glioma patients with IDH mutation and 1p19q codeletion: What to do after surgery? *Annals of Oncology*, *volume 28*:v110, 2017. ISSN 0923-7534, 1569-8041. Publisher: Elsevier.
URL https://doi.org/10.1093/annonc/mdx366.003

[65] Hikaru Sasaki, Yohei Kitamura, Masahiro Toda, Yuichi Hirose and Kazunari Yoshida. Oligodendroglioma, IDH-mutant and 1p/19q-codeleted-prognostic factors, standard of care and chemotherapy, and future perspectives with neoadjuvant strategy. *Brain Tumor Pathol*, *volume 41*:43–49, 2024. ISSN 1861-387X.
URL https://doi.org/10.1007/s10014-024-00480-1

[66] Roel G. W. Verhaak, Katherine A. Hoadley et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1,

EGFR, and NF1. *Cancer Cell*, *volume 17*:98–110, 2010. ISSN 1535-6108, 1878-3686. Publisher: Elsevier.
URL https://doi.org/10.1016/j.ccr.2009.12.020

[67] Qing-mei Kang, Jun Wang, Shi-man Chen, Si-rong Song and Shi-cang Yu. Glioma-associated mesenchymal stem cells. *Brain*, *volume 147*:755–765, 2024. ISSN 0006-8950.
URL https://doi.org/10.1093/brain/awad360

[68] Cho-Lea Tso, Peter Shintaku et al. Primary Glioblastomas Express Mesenchymal Stem-Like Properties. *Mol Cancer Res*, *volume 4*:607–619, 2006. ISSN 1541-7786.
URL https://doi.org/10.1158/1541-7786.MCR-06-0005

[69] Jinan Behnan, Pauline Isakson et al. Recruited Brain Tumor-Derived Mesenchymal Stem Cells Contribute to Brain Tumor Progression. *Stem Cells*, *volume 32*:1110–1123, 2014. ISSN 1066-5099.
URL https://doi.org/10.1002/stem.1614

[70] Jinan Behnan, Gaetano Finocchiaro and Gabi Hanna. The landscape of the mesenchymal signature in brain tumours. *Brain*, *volume 142*:847–866, 2019. ISSN 0006-8950.
URL https://doi.org/10.1093/brain/awz044

[71] Michele Ceccarelli, Floris P. Barthel et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, *volume 164*:550–563, 2016. ISSN 0092-8674.
URL https://doi.org/10.1016/j.cell.2015.12.028

[72] David N. Louis, Arie Perry et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*, *volume 131*:803–820, 2016. ISSN 1432-0533.
URL https://doi.org/10.1007/s00401-016-1545-1

[73] C. V. Raman and K. S. Krishnan. A New Type of Secondary Radiation. *Nature*, *volume 121*:501–502, 1928. ISSN 1476-4687. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/121501c0

[74] Marek Prochazka. Basics of Raman Scattering (RS) Spectroscopy. In Marek Prochazka, editor, *Surface-Enhanced Raman Spectroscopy: Bioanalytical, Biomolecular and Medical Applications*, Springer International Publishing, Cham, pages 7–19, 2016. ISBN 978-3-319-23992-7.
URL https://doi.org/10.1007/978-3-319-23992-7_2

[75] Gianmarco Lazzini, Daniela Massi et al. Raman spectroscopy diagnosis of melanoma. *Proceedings*, *volume 129*, 2025. ISSN 2504-3900.
URL https://doi.org/10.3390/proceedings2025129010

[76] Fengye Chen, Chen Sun et al. Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *volume 265*:120355, 2022. ISSN 1386-1425.
URL https://doi.org/10.1016/j.saa.2021.120355

[77] Fanghao Hu, Lixue Shi and Wei Min. Biological imaging of chemical bonds by stimulated Raman scattering microscopy. *Nature Methods*, *volume 16*:830–842, 2019. ISSN 1548-7105. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41592-019-0538-0

[78] Alen Ronč̆ević, Nenad Koruga et al. Personalized Treatment of Glioblastoma: Current State and Future Perspective. *Biomedicines*, *volume 11*:1579, 2023. ISSN 2227-9059. Publisher: Multidisciplinary Digital Publishing Institute.
URL https://doi.org/10.3390/biomedicines11061579

[79] Konstantinos Plakas, Lauren E. Rosch et al. Design and evaluation of Raman reporters for the Raman-silent region. *Nanotheranostics*, *volume 6*:1–9, 2022. ISSN 2206-7418. Publisher: Ivyspring International Publisher.
URL https://doi.org/10.7150/ntno.58965

[80] Louis-Michel Wong Kee Song and Norman E. Marcon. Fluorescence and Raman spectroscopy. *Gastrointestinal Endoscopy Clinics*, *volume 13*:279–296, 2003. ISSN 1052-5157, 1558-1950.

Publisher: Elsevier.
URL https://doi.org/10.1016/S1052-5157(03)00013-8

[81] S. Kouteva-Arguirova, Tz. Arguirov, D. Wolfframm and J. Reif. Influence of local heating on micro-Raman spectroscopy of silicon. *J. Appl. Phys.*, *volume 94*:4946–4949, 2003. ISSN 0021-8979.
URL https://doi.org/10.1063/1.1611282

[82] Weihua Zhang, Thomas Schmid, Boon-Siang Yeo and Renato Zenobi. Near-Field Heating, Annealing, and Signal Loss in Tip-Enhanced Raman Spectroscopy. *J. Phys. Chem. C*, *volume 112*:2104–2108, 2008. ISSN 1932-7447. Publisher: American Chemical Society.
URL https://doi.org/10.1021/jp077457g

[83] Chad A. Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Appl. Spectrosc.*, *volume 57*:1363–1367, 2003.
URL https://opg.optica.org/as/abstract.cfm?URI=as-57-11-1363

[84] Jianhua Zhao, Harvey Lui, David I. McLean and Haishan Zeng. Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy. *Applied Spectroscopy*, *volume 61*:1225–1232, 2007. ISSN 0003-7028. Publisher: SAGE Publications Ltd STM.
URL https://doi.org/10.1366/000370207782597003

[85] Zhi-Min Zhang, Shan Chen and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, *volume 135*:1138–1146, 2010. ISSN 1364-5528. Publisher: The Royal Society of Chemistry.
URL https://doi.org/10.1039/B922045C

[86] Guillaume Sheehy, Fabien Picot et al. Open-sourced Raman spectroscopy data processing package implementing a baseline removal algorithm validated from multiple datasets acquired in human tissue and biofluids. *Journal of Biomedical Optics*, *volume 28*:025002, 2023.
URL https://doi.org/10.1117/1.JBO.28.2.025002

[87] Darren A. Whitaker and Kevin Hayes. A simple algorithm for despiking Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, *volume 179*:82–84, 2018. ISSN 0169-7439. doi: 10.1016/j.chemolab.2018.06.009.
URL https://doi.org/10.1016/j.chemolab.2018.06.009

[88] Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.*, *volume 36*:1627–1639, 1964. ISSN 0003-2700. doi: 10.1021/ac60214a047. Publisher: American Chemical Society.
URL https://doi.org/10.1021/ac60214a047

[89] Eric Vittinghoff and Charles E. McCulloch. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, *volume 165*:710–718, 2007. ISSN 0002-9262.
URL https://doi.org/10.1093/aje/kwk052

[90] Yuanjie Liu. Adversarial nets for baseline correction in spectra processing. *Chemometrics and Intelligent Laboratory Systems*, *volume 213*:104317, 2021. ISSN 0169-7439.
URL https://doi.org/10.1016/j.chemolab.2021.104317

[91] Sinead Barton, Salaheddin Alakkari, Kevin O'Dwyer, Tomas Ward and Bryan Hennelly. Convolution network with custom loss function for the denoising of low snr raman spectra. *Sensors*, *volume 21*, 2021. ISSN 1424-8220.
URL https://doi.org/10.3390/s21144623

[92] Félix Lussier, Vincent Thibault, Benjamin Charron, Gregory Q. Wallace and Jean-Francois Masson. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends in Analytical Chemistry*, *volume 124*:115796, 2020. ISSN 0165-9936.
URL https://doi.org/10.1016/j.trac.2019.115796

[93] Chen Chen, Wei Wu et al. Rapid diagnosis of lung cancer and glioma based on serum Raman spectroscopy combined with deep learning. *Journal of Raman Spectroscopy*, *volume 52*:1798–1809, 2021. ISSN 1097-4555.
URL https://doi.org/10.1002/jrs.6224

[94] Francesco Conti, Mario D'Acunto et al. Raman spectroscopy and topological machine learning for cancer grading. *Sci Rep*, *volume 13*:7282, 2023. ISSN 2045-2322.
URL https://doi.org/10.1038/s41598-023-34457-5

[95] Chenxi Zhang, Ying Han et al. Label-free serum detection based on Raman spectroscopy for the diagnosis and classification of glioma. *Journal of Raman Spectroscopy*, *volume 51*:1977–1985, 2020. ISSN 1097-4555. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jrs.5931.
URL https://doi.org/10.1002/jrs.5931

[96] Tommaso Sciortino, Riccardo Secoli et al. Raman Spectroscopy and Machine Learning for IDH Genotyping of Unprocessed Glioma Biopsies. *Cancers*, *volume 13*:4196, 2021. ISSN 2072-6694.
URL https://doi.org/10.3390/cancers13164196

[97] Wilson Wen Bin Goh, Wei Wang and Limsoon Wong. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, *volume 35*:498–507, 2017. ISSN 0167-7799, 1879-3096. Publisher: Elsevier.
URL https://doi.org/10.1016/j.tibtech.2017.02.012

[98] Ser-Xian Phua, Kai-Peng Lim and Wilson Wen-Bin Goh. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Computational and Structural Biotechnology Journal*, *volume 20*:4369–4375, 2022. ISSN 2001-0370. Publisher: Elsevier.
URL https://doi.org/10.1016/j.csbj.2022.08.022

[99] Jelena Čuklina, Chloe H Lee et al. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Molecular Systems Biology*, *volume 17*:e10240, 2021. ISSN 1744-4292. Publisher: John Wiley & Sons, Ltd.
URL https://doi.org/10.15252/msb.202110240

[100] Jelena Čuklina, Patrick G. A. Pedrioli and Ruedi Aebersold. Review of Batch Effects Prevention, Diagnostics, and Correction Approaches. In Rune Matthiesen, editor, *Mass Spectrometry Data Analysis in Proteomics*, Springer New York, New York, NY, pages 373–387, 2020. ISBN 978-1-4939-9744-2.
URL https://doi.org/10.1007/978-1-4939-9744-2_16

[101] Farnaz Kheiri, Shahryar Rahnamayan, Masoud Makrehchi and Azam Asilian Bidgoli. Investigation on potential bias factors in histopathology datasets. *Sci Rep*, *volume 15*:11349, 2025. ISSN 2045-2322. Publisher: Nature Publishing Group.
URL https://doi.org/10.1038/s41598-025-89210-x

[102] Emily E. Storey and Amr S. Helmy. Optimized preprocessing and machine learning for quantitative Raman spectroscopy in biology. *Journal of Raman Spectroscopy*, *volume 50*:958–968, 2019. ISSN 1097-4555. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jrs.5608.
URL https://doi.org/10.1002/jrs.5608

[103] Yuping Liu, Junchi Wu, Yuqing Wang and Sicen Dong. Direct recognition of Raman spectra without baseline correction based on deep learning. *AIP Advances*, *volume 12*:085212, 2022. ISSN 2158-3226.
URL https://doi.org/10.1063/5.0100937

[104] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In D. W. Pfitzner and J. K. Salmon, editors, *Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Proceedings of a conference held August 2-4*, 1996, pages 226–331.

[105] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *volume 20*:53–65, 1987. ISSN 0377-0427.
URL https://doi.org/10.1016/0377-0427(87)90125-7

[106] F. Pedregosa, G. Varoquaux et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *volume 12*:2825–2830, 2011.

[107] Francisco Herrera, Sebastián Ventura et al. Multiple Instance Learning. In Francisco Herrera, Sebastián Ventura et al., editors, *Multiple Instance Learning: Foundations and Algorithms*, Springer International Publishing, Cham, pages 17–33, 2016. ISBN 978-3-319-47759-6. URL https://doi.org/10.1007/978-3-319-47759-6_2

[108] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, *volume 25*:1–25, 2010. ISSN 1469-8005, 0269-8889. URL https://doi.org/10.1017/S026988890999035X

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0549-2 (PRINT)
ISBN 978-952-02-0550-8 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)