Machine learning for predicting the glioma type of live cells using Raman spectroscopy

BSc Noor E. Zannat

MSc thesis
December 2025

DEPARTMENT OF MATHEMATICS AND STATISTICS

Supervisors:
Prof. Ion Petre
Dos. Yury Nikulin

UNIVERSITY OF TURKU, Department of Mathematics and Statistics

Master's Thesis
Major subject: Mathematics
Author: Noor E Zannat
Title: BSc
Supervisor: Prof. Ion Petre; Dos. Yury Nikulin
Number of pages: 49 pages + appendices 1 page
Date: December 2025

Abstract

Background: Gliomas are heterogeneous brain tumors, including aggressive glioblastomas, where rapid and accurate diagnosis is critical. Standard histopathology is invasive and slow. Raman spectroscopy provides a non-destructive alternative, probing molecular composition with high resolution. Coupled with machine learning, it offers potential for fast, precise glioma characterization and personalized care.

Methods: We collected Raman spectra from 284 live cell culture samples across four classes: astrocytoma (Astro), oligodendroglioma (Oligo), glioblastoma (GBM), and non-tumor controls. Preprocessing included cosmic rays removal, and baseline correction with airPLS. Spectra were normalized (min–max), silent regions and low-quality signals were excluded. A two-staged pipeline was developed: (1) cell contour detection from spectral profiles and (2) mutation status classification. Several classifiers were tested, and permutation feature importance was incorporated to identify discriminative Raman frequencies.

Results: Preliminary results demonstrate moderate separation of glioma IDH1 mutation status, with XGBoost classifier achieving moderate predictive performance across the four cell classes. Key discriminative features were observed in spectral regions associated with protein, lipid, and nucleic acid vibrations, consistent with known metabolic and structural differences between glioma types.

Conclusions and Outlook: Our workflow combines Raman spectroscopy with advanced preprocessing and machine learning for glioma classification. Beyond moderate accuracy, the identification of biologically meaningful spectral features enhances interpretability. This approach could accelerate diagnostics, reduce invasiveness, and support personalized strategies in neuro-oncology. Future work will extend the framework by increasing the performance of the model, classifying glioma subtypes, patient-derived samples, and advancing minimally invasive glioma diagnostics.

Keywords: Raman spectroscopy, Glioma live cells, Machine learning, XGBoost.

Contents

1 Introduction

Gliomas are a type of brain tumor, which can be aggressive and, is essential to identify the subtypes of it, including their mutation status, for providing the proper treatment. Traditional methods can be invasive, time-consuming, and labor-intensive. The combination of Raman spectroscopy (RS) and machine learning (ML) is a fast, non-destructive, and non-invasive method to classify glioma live cells. In this thesis, we have discussed how RS and ML can distinguish the subtypes of glioma live cells and their IDH1 mutation status.

1.1 Medical Background

Gliomas are one of the most common types of brain tumors among humans all over the world [1]. The source of this tumor is mostly in the glial cells of the brain and, as a result, is called a glioma [1]. The World Health Organization (WHO) has classified them into two categories, such as high-grade gliomas (HGGs) and low-grade gliomas (LGGs) [1]. LGGs have slow growth, while HGGs have faster growth, leading to more aggressive behavior [1]. Although LLGs grow slowly, they can form in an important part of the brain, causing a severe problem [2]. When the glial cells grow abnormally, it causes the glioma brain tumor [3].

Before the discovery of the IDH mutation [4] (before 2016), glioma was classified mainly by its morphologic features [5]. However, after the invention of the IDH mutation, it became one of the most significant biomarkers for the diagnosis, prediction of patient outcomes, and treatments [4] [6] [7]. In 2021, WHO classified adult diffuse gliomas into three major types by grading them within each tumor subtype, such as astrocytoma with IDH1mut, oligodendroglioma with IDH1mut, and glioblastoma with IDH1wt [8].

1.2 Diagnostic Practices and Limitations

In traditional methods, doctor's grade the glioma subtypes by collecting tissue samples through biopsy or surgery. These methods can be invasive, time-consuming, painful, and sometimes lead to error [9]. Apart from that, the risky behavior of these methods can bring more dangers than benefits to glioma patients. Besides, the number of radiologists to diagnose and evaluate brain tumors is not satisfactory [1]. Scientists have developed MRI in such a way that it can now predict important molecular features (IDH mutation [10], 1p/19q codeletion [11], and MGMT methylation [12]) of brain tumors before surgery with an accuracy level over 80%. However, this method still does not provide any biological composition to doctors, which has become a great concern for surgeons to completely rely on them [10] [12] [13]. Moreover, MRI can not detect other molecular markers, such as TERT, EGFR, chromosome 7/10 changes, and CDKN2A/B [14]. Because of all these factors, medical imaging and machine learning researchers are working hard to develop a non-invasive, fast, and non-destructive method to classify the glioma subtypes and their mutation status. So, it is essential to predict the glioma subtypes and identify the biochemical differences between the subtypes, so that doctors can provide better

and accurate treatment to the patient.

1.3 Motivation for Using Raman Spectroscopy and Machine Learning

Today, experts can measure how the tumor has spread by measuring the biomolecular features of tissue and sub-micron resolution. For measuring these, scientists have developed a label-free technique. Raman spectroscopy is one of these techniques that researchers have suggested to use for brain surgery since 1990 [15]. Raman spectroscopy is helpful because it can detect detailed chemical information about the tissue by producing a unique vibrational fingerprint for each sample. The application of this method is not limited to improving the performance of stereotactic brain tumor biopsies [16], but has spread to detect tumor infiltration in living patients [17] and help with the molecular classification of tumors [18].

Glioma classification also requires machine learning and DNA methylation profiling, minimizing the need for a histopathological technique, indicating the necessity of using molecular markers [19] [20] [21] [22] [23] [24] [25]. The combination of AI and Raman spectroscopy is more reliable in diagnosing tumors accurately [26]. Some of the previous studies have shown the capability of Raman spectroscopy to identify brain tumors during surgery. Livermore LJ, et al. built a predictive machine learning model utilizing linear discriminant analysis (LDA), principal component analysis (PCA), and Raman spectra to differentiate glioma tumors and normal tissues [27]. They also experimented in one of their studies on astrocytoma (IDH-wild-type), astrocytoma (IDH-mutant), and oligodendroglioma. Here, they used the same ML methods and found that the predictive model had a high sensitivity (79–94%) and specificity (90–100%), where IDH mutations had the highest sensitivity and specificity around 91% and 95%, respectively [28]. Uckermann O, et al. distinguished tumor and normal tissues using Raman spectroscopy (RS) and autofluorescence (AF). In this research, they trained the LDA model with ex vivo data to classify tumor and non-tumor samples with an accuracy level of approximately $83 - 84\%$ [29]. Jermyn M, et al. used a hand-held Raman probe to distinguish between normal brain tissues and tumor tissues with 93% sensitivity and 91% specificity [17]. Riva M, et al. analyzed 3450 Raman spectra from 63 samples using random forest and gradient boosting trees to identify the biological relevance of brain tumors, where they found 19 novel RS shifts with an accuracy of 83% and a precision of 82% [30]. Vrazhnov D, et al. investigated glioma tissue and blood serum biomarkers by applying RS and ML (support vector machine (SVM), RF, and XGboost) to observe changes in lactate, tryptophan, fatty acids, and lipids in serum [31]. Additionally, Zhang L, et al. demonstrated that the combination of PCA-SVM and RS has the ability to distinguish live glioma cells from normal tissues and different glioma grades with a performance rate of over 80% [32]. Ember K, et al, developed a ML model using RS data to detect glioblastoma and meningioma. They also performed the prior experiment using the Gaussian fitting technique to extract the position of the wavelengths [33]. In another study carried out by Klamminger GG, et al, they performed an analysis on FFPE brain tumors and metastases using RS with the combination

of RF classifier to identify tumor types and primary origins in metastatic cases [34]. In a recent study, Lita A, et al, introduced a computational workflow based on RS, which they named APOLLO. In this research, they used the spectra of FFPE tissue to classify between glioma and non-tumor tissue, IDH1 mutant and wild-type tumors, and their methylation subtypes [35].

Most of the current studies related to RS for classifying glioma emphasize tissue differentiation and IDH status. Nevertheless, they rarely did research for molecular marker information, stable cosmic rays or artifact correction. As a result, the diagnostic performance of these studies may sometimes be below the required routine clinical adoption. So, this highlights the necessity for developing a more advanced ML model using RS for high-quality pre-processing and molecularly informed classification.

1.4 Thesis Objectives

The main objectives of this thesis are to develop a ML model using RS data to classify live glioma cells and identify their IDH1 mutation status. This thesis seeks not only to build a predictive classifier model, but also to analyze the biologically meaningful insights that can be applicable to real-world clinical problems. In addition, our research mainly focuses on the following objectives:

- Developing a reliable preprocessing pipeline for Raman spectra by delineating the tumor cells from the water environment, removing cosmic rays, removing silent regions, correcting the baseline, normalizing the spectral intensities, and mitigating class imbalance to ensure that all the classes are balanced or given the same importance.
- Training clustering techniques to detect tumor and non-tumor cells. In our case, since we used the glioma live cells, so we tried to indicate the tumor and the water in our sample.
- Training supervised ML classifiers to predict IDH1 mutation status(IDH1mut and IDH1wt). This is the key task of our thesis.
- Evaluating the model performance of the ML models by calculating the prediction accuracy, precision, and other related matrics.
- Identifying the feature ranking to validate the biological interpretation, and to check the wavelengths where the Raman spectra model relies most.

All of these objectives aim to advance a ML technique using RS data to reveal the biological interpretation by distinguishing tumors and their mutation status.

1.5 Thesis Outcomes

This thesis illustrates that the combination of ML and RS is a non-invasive and fast method to classify the tumor mutation status. The outcomes address the objectives of our thesis. Although the justification for some of the objectives might be poor

because our project is still ongoing. In our study, we developed our preprocessed workflow, which can successfully mitigate the cosmic rays, remove non-informative or meaningless regions, correct the baseline, and normalize the RS data by preserving the biologically significant spectra. Apart from that, our enhanced clustering method has successfully separated the tumor and the water by identical comparison between the optical image and the cluster map, as well as labeling them. In addition, we have improved our ML IDH1 mutation status classifier model by making the accuracy level balanced. However, we need to develop it further, which has been discussed in the conclusions section. To measure the performance level of our model, we have used several matrices, which have revealed an almost balanced accuracy among the minority and majority classes. Finally, we have chosen the top 23 features to recognize the mechanism between the spectral data and their underlying molecular properties. To get the corresponding results of our objectives, we are continuing our analysis.

2 Raman Spectroscopy and Tumor Samples

The functionality of Raman spectroscopy is based on the inelastic scattering of light, with the advantages of specificity [36] [37]. This technique analyzes the biochemical characteristics of substances depending on the positions and intensities of Raman peaks [38]. Raman spectroscopy is widely used in tumor diagnosis. In our thesis, we analyzed glioma tumor cells using Raman spectroscopy.

2.1 Fundamentals of Raman Spectroscopy

In 1923, Smekal first predicted Raman scattering by molecules using classical quantum theory [39]. Raman and Krishnan experimentally observed them in 1928 [40] [41]. Today, Raman spectroscopy techniques have more than 25 types [42]. In our study, we used spontaneous Raman spectroscopy. It became prominent after fifty years of its first observation in the presence of water and other useful polar solvents because of its ability to absorb light in the infrared region [42].

In Raman spectroscopy, most photons are elastically scattered when photons from the laser interact with the sample or molecule(Figure 1). The molecule absorbs an incident photon while interacting. This results in Rayleigh scattering with the same wavelength or Anti-Stokes and Stokes Raman scattering, which are proportional to the vibration of the molecule. When a molecule comes into contact with light, it absorbs a little bit of light energy, which causes a Raman shift by slightly changing the energy of the scattered light. Since different molecules vibrate in their own pattern, they have their own Raman shifts. After we measure these changes or Raman shifts, we get a spectrum or fingerprint, which helps us to detect which molecule is present [43].

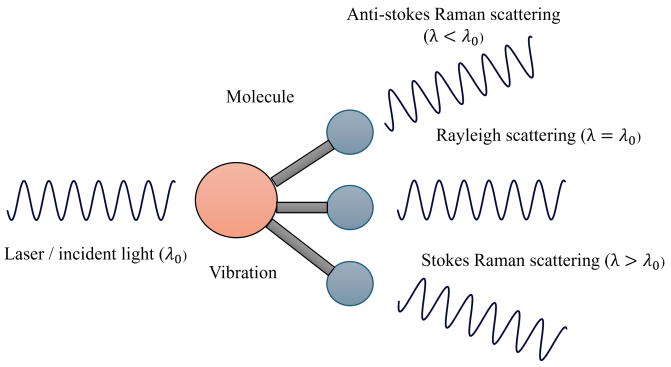


Figure 1: Photon scattering by a photon interacting with a molecule [43]. The molecule absorbs the incident laser light (photon). This photon either scatters with the same wavelength i.e., Rayleigh scattering, or with a wavelength shift proportional to the vibration of the molecule i.e., Anti-Stokes or Stokes Raman scattering.

Raman spectroscopy is a label-free imaging technique [44], which does not require any fluorescent marker molecules [45] [46]. This special feature has made this technique significant in the biological and medical field research [44]. Spontaneous Ra-

man spectroscopy can detect the molecular fingerprints of tissue without preparing the samples using a vibrational technique, which assists in measuring lipids with subcellular resolution [47] [48] [49]. For instance, to identify and differentiate benign and malignant breast cancer lesions, this method investigates their unique chemical composition. In addition, it helps detect the margin of the brain tumor and diffuse tumor cells [17] [50] [51] [52], predict brain tumor grades [53], and genetic subtypes [28].

2.2 Brain Tumor Subtypes and Sample Groups

The brain tumor has the same features as the tumors of the other parts of the body, but causes some special problems because the brain is a very delicate organ of the human body [54]. There are different types of brain tumors, such as gliomas, meningiomas, and pituitary adenomas. In our thesis, we focus only on gliomas because removing them entirely from the brain is challenging. When removing gliomas, the surgeons give importance to remove them as much as possible without damaging other vital areas of the brain. But, glioma often infiltrates healthy cells, which makes it difficult to distinguish between tumor and non-tumor cells. Since patient survival is mostly dependent on how safely gliomas are removed from the brain, so it is significant to identify them properly [55] [56].

Adult gliomas [57] are classified according to their features. If the gliomas have IDH-mutant (IDH-mut), they are classified as astrocytoma (astro) or oligodendroglioma (oligo). To distinguish between these two types, the changes of 1p/19q codeletion are observed. If any change is noticed in 1p/19q codeletion, then it is classified as an oligo type; otherwise, an astro type. The most aggressive type of glioma, named glioblastoma (GBM) is diagnosed when the brain tumor has IDH-wildtype (IDH-wt). The grade (2, 3, or 4) depends on how aggressive the tumors look under the microscope.

2.2.1 IDH1-mutant

Glioma, a common malignant brain tumor, can be grouped by changing a gene called IDH. This change helps the tumor grow [58]. More than 95 percent of IDH1-mut in glioma change the position of the amino acid from arginine 132 (R132) to something else [59] [60] [61]. During the IDH1 mutation, tumors may grow due to the 2-hydroxyglutarate (2-HG) molecule [59] [60] [61]. A large meta-analysis has found that IDH1mut gliomas have a higher survival rate than non-mutated gliomas with a hazard ratio for overall survival of 0.39 (95 percent CI: 0.34–0.45) [62]. This type of tumor is often diagnosed among young people compared to IDH1wt tumors [63].

2.2.2 GBM IDH1wt

Glioblastoma multiform (GBMs) is the most aggressive type of brain tumor among adults [8]. The central nervous system (CNS) classified it as a grade 4 astrocytoma (astro) in 2021 [8]. Among all glioblastoma, IDH1 wildtype (wt) grows fast and responds poorly to treatment, so most patients die within 15 months, although

providing the best available care [64]. The classification between IDH1 mutant and wildtype is crucial, since it helps doctors choose the correct treatment and plays a significant role in how the World Health Organization (WHO) classifies diffuse gliomas [65].

2.2.3 GBM IDH1mut

GBM IDH1mut is a type of GBM tumor in which IDH1 is mutated [4]. In secondary GBM, mutation of IDH1 often causes it to tend to obtain the biological and clinical features of its lower-grade counterparts [66]. GBM IDH1mut [67] suppresses the immune system by having excessive molecules and fewer cancer-killing T cells. As a result, these types of tumors don't respond well to immune-based treatments.

2.3 Histological and Molecular Subgroups of Glioma

Based on IDH1mut and IDH1wt, the gliomas are classified into some histological subtypes. In this section, we discuss them.

2.3.1 Oligodendroglioma (Oligo)

Oligodendroglioma [68] is a rare tumor that responds well to treatment compared to other tumors. It is generally found in the white matter of the cerebral hemispheres. This type of tumor can be differentiated as low-grade or high-grade tumors. As the first symptom, the low-grade oligo causes seizures, and the high-grade oligo causes headache or loss of focus. Patients who have low-grade oligo can delay their treatment until tumors start growing significantly, whereas patients with high-grade oligo should receive immediate treatment.

2.3.2 Astrocytoma (Astro)

Astrocytoma [69] is one of the most common types of brain tumors in children. WHO divided astro into low-grade tumors (grade II), high-grade tumors (grade III, includes anaplastic astrocytomas (ANAs)), and glioblastomas (grade IV). Although the symptoms of ANAs and GBMs are poor, there are some significant differences between them [70]. Supratentorial pilocytic astrocytomas (PSTs), a subtype of astro, are rare and uncommon in both children and adults [71] [72]. Removing them completely through surgery is easier compared to the others, because of their growth [73]. Astrocytic tumors come under the heterogeneous group, which is theoretically divided into four grades, but actually, they have distinct subgroups [73].

2.3.3 Glioblastoma (GBM)

The most aggressive malignant type of brain tumors is glioblastoma [74], whose risk factor is widely elusive. Before, traditional diagnosis was used to identify them utilizing a microscope; however, today, doctors also give importance to molecular markers. These markers help them taking decisions regarding personalized treatment. Currently, doctors use surgery, radiotherapy, and chemotherapy to treat

GBM tumors. But, researchers are looking for new strategies to treat it, focusing on specific molecular changes and immunological approaches.

2.4 Normal Tissue

Normal brain tissues [75] are healthy tissues in our brain that help researchers identify changes in tumors and metabolic activities. Surprisingly, some IDHwt gliomas behave as healthy normal tissue because they are less aggressive and have a better patient outcome compared to IDHwt GBMs.

2.5 Our Dataset

In this thesis, we used a diverse dataset from cell-culture grown in a water-based medium. Since we derived the cultures from primary brain tumors, it is common to obtain very few cells from the derived cultures. These cells don't look the same. These are some crucial features to consider in our thesis, as this may affect the accuracy of the machine learning classification model.

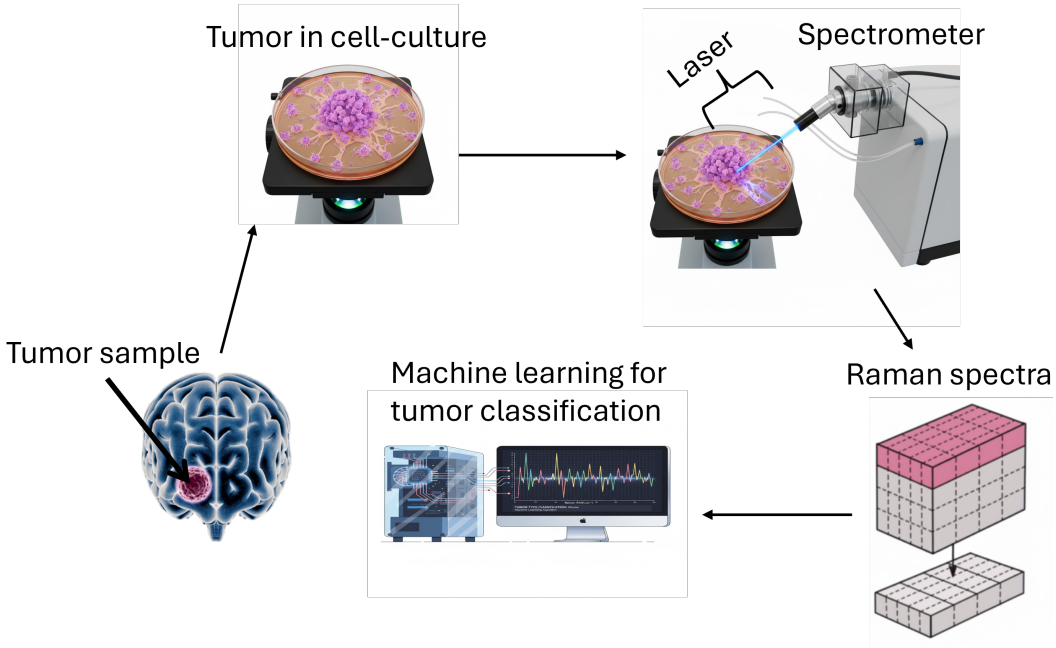


Figure 2: Study design involves collecting glioma types of tumors from the brain, growing them in cell-culture, getting the spectra from them with the help of Raman spectroscopy, and analyzing these spectra using machine learning. We generated this image through Google Gemini.

We used a total of 284 samples from both tumor and non-tumor types. The dataset includes four primary categories, such as oligo (29 samples), astro (18 samples), GBM (231 samples), and normal brain tissue (6 samples). Additionally, these tumor subtypes were grouped according to IDH1mut status, such as IDH1mut (51 samples), GBM IDH1wt (211 samples), GBM IDH1mut (16 samples) and normal

brain tissue (6 samples). This diversity in our data plays a significant role in comparing the molecular and spectral differences between tumor and non-tumor cells. The collaborators (Mioara Larion and Adrian Lita) at the National Institutes of Health (NIH), National Cancer Institute (NCI), Bethesda, USA, collected these data from cell-culture using Raman spectroscopy as shown in Figure 2.

We got samples from the water medium because they were growing there, so it is essential to characterize the spectral contributions of the tumor cells and the water. For this reason, we used Raman spectroscopy in our analysis. Figure 3 shows the significant differences between water and cell cultures in Raman spectroscopy frequency.

Water produces a very high Raman signal with intensities over 5000-6000 in the region above $3000cm^{-1}$, approximately. Cells give a high signal for the Raman shift around $2800cm^{-1}$, where we have lipids and proteins. Water gives almost no signal before frequency $2800cm^{-1}$. So, if we sum up the intensities up to frequency $2800cm^{-1}$, the water spot will give a very small sum compared to the cells. In Figure 3, we have in red a spectrum of a water region, in blue a cell region. On the left of the plot, the difference between the frequencies is visible as the highest and the difference in the other direction, just before frequency $2800cm^{-1}$ is much more subtle.

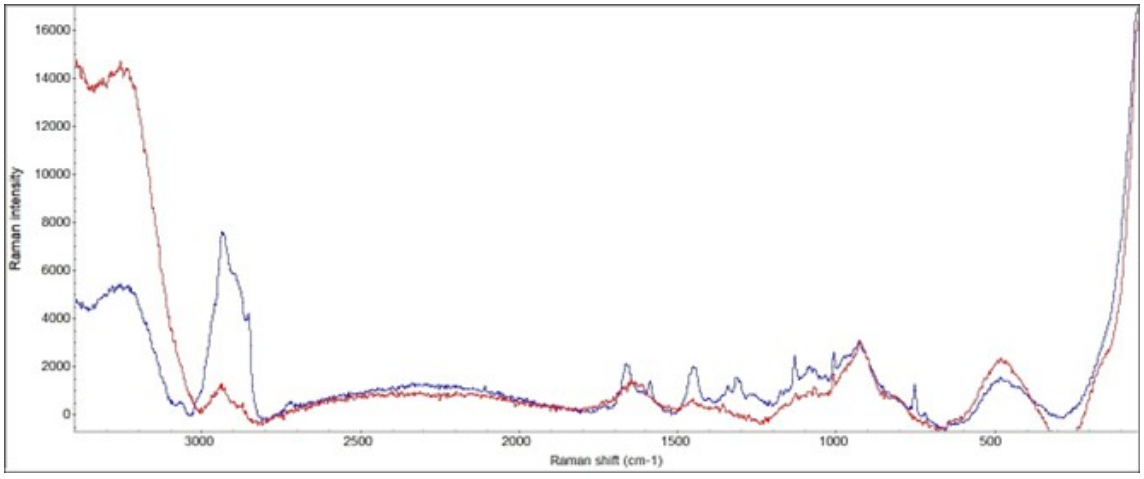


Figure 3: Raman spectral differences between water and cell-containing regions. The blue color indicates the intensity of glioma tumor cells, and the red color represents the intensity of the water from the cell-culture.

3 Spectra Processing

Spectra preprocessing is an essential step in removing the artifacts and noise from our Raman spectra data before classification. The preprocessed steps enable us to utilize clean data in our machine learning model, resulting in fewer errors. The preprocessing of the Raman spectra contains some steps, as shown in Figure 4. In this chapter, we discuss these steps.

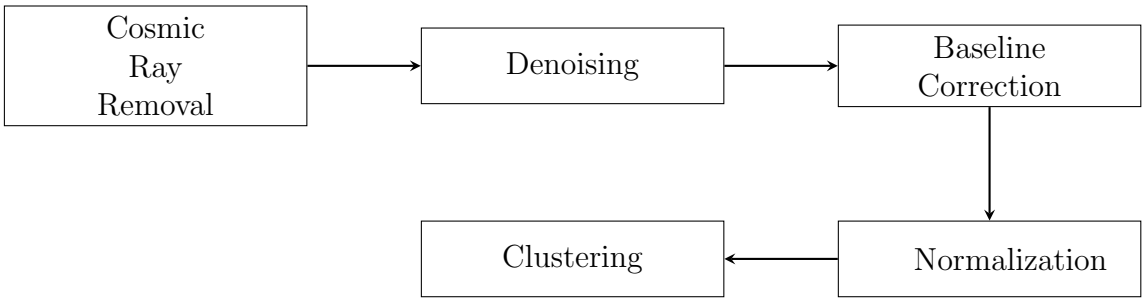


Figure 4: Data preprocessing workflow.

3.1 Cosmic Rays Removal

Cosmic rays produce very large random spikes with a very narrow bandwidth in the spectrum, which do not have a good correlation with the rest of the data [76]. Those spikes mainly occur when a cosmic ray hits the detector while taking the measurements [77]. This is a widespread occurrence for live glioma cells. To address this issue, we performed some methods to despike the data.

3.1.1 Modified Z-Score

The modified z-score is mainly used to detect outliers. In our case, we considered cosmic rays as outliers, since they do not have a good correlation with the actual Raman spectra data. It is calculated using the following equation,

$$\text{Modified } z\text{-score} = 0.6745\frac{(x_i - \tilde{x})}{\text{MAD}} \tag{1}$$

where,

- x_i is a variable encoding a single data,
- $\tilde{x}$ represents the median of the dataset,
- $MAD = \text{median}(|x_i - \text{median}(x)|)$

To detect the cosmic rays, we set a threshold. Iglewicz and Hoaglin proposed a threshold of 3.5 for detecting outliers as a guideline [78]. Whitaker and Hayes used 6 as a threshold to despike the Raman spectra [79]. Those threshold values didn't work with our data.

For the threshold, we tried both the standard deviation (std) of our dataset and the std of the modified z-score. If $std >$ modified z-score, or std(modified z-score) $>$ modified z-score, then we considered it a cosmic ray. For replacing the cosmic rays, we used a window size of 11 for calculating the median.

3.1.2 Optimized Cosmic Rays Removal Method

We developed an optimized cosmic rays detection and reduction pipeline for 3D Raman spectra data of glioma cells. Each Raman data had a shape of (X, Y, Z), stored as a 3D numpy array. To replace the cosmic rays, we tried to optimize two parameters: cosmic rays detection threshold ($p1$) and window size for replacing the cosmic rays ($p2$) by the median value. We estimated the median absolute deviation (MAD) for each cell, and tried to optimize the $p1$ value, which ranges from $3 * median(MAD)$ to $8 * median(MAD)$.

We divided the range from $3 * median(MAD)$ to $8 * median(MAD)$ into 10 equal lengths to estimate the optimal $p1$ value. Each time we calculated the $p1$ value with these 10 numbers and estimated which $p1$ value, it detected the maximum cosmic rays.

Counting the total number of spikes relied on two different parameters: the height multiplier and the maximum width of the spikes. We tested our Raman spectra data using height multipliers from 1 to 5 and maximum widths from 1 to 5. The height multiplier and the maximum width helped us to detect how long and narrow spikes can be considered as cosmic rays, respectively. The high multiplier for calculating the threshold during spike detection helps us to avoid false positives, and the lower multiplier for calculating the threshold during replacement ensures that the detected spikes are handled gently.

For correcting the detected cosmic rays, we applied the median replacement technique. We replaced the intensity of each spike with its' neighborhood values. To select the optimized number of neighborhood values, we took a window size $p2$ ranging from 3 to 11, with increments of 2. When $p1 < |x_i - median(x)|$, we replaced those cosmic rays with the median with the optimized $p2$ value.

We applied this procedure to the entire dataset for each 3D numpy array cell file. This procedure continues until we reach the optimal solution. Figure 5 explained this whole process with the help of a flowchart.

3.2 Savitzky-Golay Filter

Data smoothing is necessary to increase the precision of data without distorting the signal's tendency. To fulfill this purpose, the Savitzky-Golay (SG) [80] filter is one of the known methods, which is a simple least square fit convolution for smoothing. It works with two parameters: polynomial order and window size. The window size decides how many data points the SG filter is going to smooth, and the polynomial order helps to smooth those points, replacing them with the polynomial

curve. However, selecting the window size and polynomial order is very crucial for data analysis. Chen et al. proposed a window size of 7 and polynomial order from 2 to 4 for the SG filter [80]. Liu et al. used polynomial order 6 for the SG filter in their work [81]. In our work, we used polynomial orders from 3 to 7, and window sizes from 3 to 11 with an increment of 2.

3.3 Baseline Correction

Baseline correction plays a significant role in data preprocessing in Raman spectroscopy by removing background effects and separating accurate spectroscopic signals from interference effects [82]. One of the most familiar methods is modified polynomial fitting, which requires user intervention and is sensitive to the noisy data [83]. To overcome this issue, we used the adaptive iteratively reweighted Penalized Least Squares (airPLS) algorithm [84] in our thesis. This algorithm does not require user intervention or peak detection. This algorithm is basically based on the penalized least squares (PLS), which requires minimizing the following objective function:

$$S = \sum_{i=1}^{m} w_i (y_i - z_i)^2 + \lambda \sum_{i=d+1}^{m} (\Delta^d z_i)^2 \tag{2}$$

where,

- y_i is the original signal,
- z_i is the estimated baseline,
- w_i are weights,
- λ is the smoothing parameter,
- Δ^d represents the d-th order difference operator,
- m is the length of the signal.

For the high weights, the first term helps to fit the baseline of the data, and the second term helps us to smooth the baseline.

3.3.1 Whittaker Smoother

The Whittaker smoother [84] facilitates the implementation of the penalized least squares method for the data. This smoothing function is as follows:

$$(\mathbf{W} + \lambda \mathbf{D}^T \mathbf{D})\mathbf{z} = \mathbf{W}\mathbf{y} \tag{3}$$

where,

- $\mathbf{W}$ is a diagonal matrix of weights,
- $\mathbf{D}$ is a difference matrix of specified order,
- λ controls the smoothness of the resulting baseline.

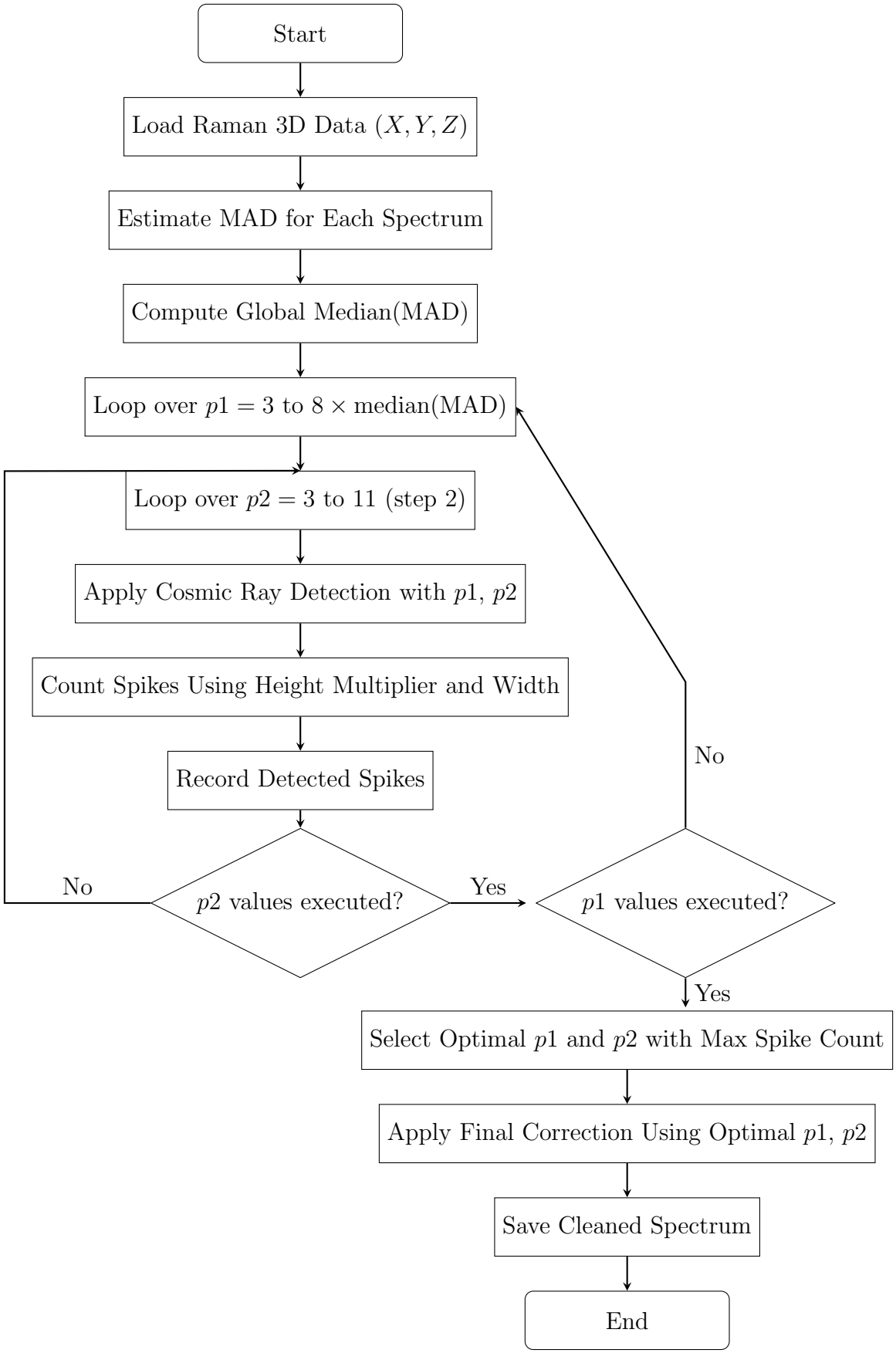


Figure 5: Flowchart for optimized cosmic rays removal process.

3.3.2 Adaptive Iterative Algorithm

In this step, we need to calculate the weights iteratively based on the residue of the measured signal and the current baseline estimate. The points above the estimated baseline have zero weights, whereas the points below the estimated baseline increase exponentially.

After initializing the weights to 1 and estimating the baseline z_i using the Whittaker smoother, we need to calculate the residual $d_i = y_i - z_i$. We need to update the weights adaptively [84] in the following way, after calculating the sum of the absolute values of the negative differences ($dssn^i = \sum_{j=1}^{n} max(0, -d_j^i)$):

$$w_j^i = \begin{cases} 0, & d_j^i \geq 0 \\ \exp\left(\dfrac{p \cdot i \cdot |d_j^i|}{\text{dssn}^i}\right), & d_j^i < 0 \end{cases} \tag{4}$$

where,

- i is the iteration number,
- p is the power factor that scales how aggressively the weights grow for negative deviations.

3.4 Normalization

Normalization is a technique used to scale the data after removing outliers, typically within a range of 0 to 1 or -1 to 1 [85]. To avoid multiplicative effects and to maintain data consistency and simplicity, normalization is necessary [86]. In our thesis, we tried two types of normalization on the Raman spectra data.

3.4.1 Min-Max Normalization

Min-max normalization [87] is a common data preprocessing technique that helps to scale different features of the data between 0 and 1 [88]. This method is easy to implement because of its simplicity. These days, this method has become a good choice among data scientists because it mitigates the effect of outliers by compressing them. The transformation formula of this method is,

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

where,

- x' is the normalized value,
- x is the original data,
- x_{min} is the minimum value of the original data,
- x_{max} is the maximum value of the original data.

3.4.2 L2 Normalization

L2 normalization [89] modifies the dataset in a way that the sum of the squares of each row will be up to 1. This method primarily works with vectors. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where $\mathbf{x}$ is a vector, then the L2 norm or the euclidean length is,

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{6}$$

To get the normalized value, we need to divide each data point by its norm in the following way,

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} = \left(\frac{x_1}{\|\mathbf{x}\|_2}, \frac{x_2}{\|\mathbf{x}\|_2}, \ldots, \frac{x_n}{\|\mathbf{x}\|_2} \right) \tag{7}$$

where,

- $\mathbf{x}'$ is the normalized value,
- $\mathbf{x}$ is the vector,
- $\|\mathbf{x}\|_\mathbf{2}$ is the L2 norm.

3.5 Clustering

Clustering is widely used as an unsupervised learning method, making it a significant part of data mining [90]. This method groups similar instances, while other instances with different features belong to other groups [90]. We employed this technique in our thesis to separate tumor cells, water, and other substances based on their distinct features and positions.

3.5.1 K-Means Clustering

K-means [91] clustering is one of the most commonly used and straightforward algorithms [92]. Its main goal is to enhance the similarities between the data points and their associated centroids. That means this algorithm partitions n observations into k clusters, making each observation fall into the same cluster with the closest mean or cluster centroid by minimizing the inertia or within-cluster sum-of-squares in the following way:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \|x_i - \mu_j\|^2 \tag{8}$$

where,

- x_i is the i-th data point,
- μ_j is the possible cluster centroid,
- C is the set of cluster centroids.

3.5.2 Agglomerative Clustering

Agglomerative clustering [93] falls into one of the hierarchical strategy categories. It is widely used for small to medium-sized data sets because of its simplicity. This clustering method works as a bottom-up approach. Initially, it considers each data point as an individual cluster. In the next step, Agglomerative emphasizes the distance metrics (e.g., Euclidean distance) to measure the similarities between two data points.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{9}$$

where,

- x_i and x_j are two data points

We merge similar data points or clusters using the linkage criterion (e.g., ward linkage) [94] [95].

$$D(A, B) = \frac{|A|\,|B|}{|A \cup B|} \|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|^2 \tag{10}$$

where,

- A and B are two clusters.
- $|A|$ and $|B|$ represent the data points of each cluster,
- $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ represent the centroids of clusters A and B, respectively.

This process continues until we reach the stopping criteria.

3.5.3 BIRCH

For very large datasets, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [96] is an appropriate choice for separating data. Instead of scanning all data points, this method measures the natural closeness of the points. Instead of treating all data points uniformly, BIRCH treats a dense region of points as a single cluster by building the Clustering Feature Tree (CFT). This clustering technique removes the points outside of this region, considering them as outliers.

Let, a subcluster contain N data points $X_1, X_2, ..., X_N$. Then the centroid (X_0), radius(R), and diameter(D) of the subcluster are defined as,

$$X_0 = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{11}$$

$$R = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|X_i - X_0\|^2} \tag{12}$$

$$D = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|X_i - X_j\|^2} \tag{13}$$

In BIRCH, two disjoint subclusters merge using the additive property of CFT. This clustering method has two parameters, such as the threshold and the branching factor. The branching factor helps to limit the number of subclusters, and the threshold assists in keeping the distance between the entering sample and the existing subclusters within a limit [97].

3.6 Metrics and Scoring

To evaluate the clustering performance, we estimate some metrics and scores. Those matrices help us to get the optimal number of clusters from the clustering algorithm [98]. Which metric to choose depends on the goal and the application.

3.6.1 Davies-Bouldin Score

The Davies-Bouldin (DB) [99] score evaluates the quality of clustering by measuring the compactness and separation. Compactness helps to measure how close the data points are within a single cluster. On the other hand, separation measures the distance between different clusters. Since most similar clusters are identified by the intra and inter-cluster distances, the DB score is the average of these maximum similarity values of all clusters. This value ranges from zero (0) and gives better clustering quality when it gives lower values. The equation for the Davies-Bouldin score is as follows:

$$\mathrm{DB} = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left(\frac{S_i + S_j}{d_{ij}} \right) \tag{14}$$

where,

- n represents the number of clusters
- S_i represents the intra-cluster scatter
- d_{ij} represents the inter-cluster distance

3.6.2 Calinski Harabasz Score

Calinski Harabasz (CH) [100] score, or the Variance Ratio Criterion, is another method of evaluating the clustering quality. It depends on two factors: the sum of between-cluster dispersion and of within-cluster dispersion. The CH score ranges from zero (0), and the higher value represents good quality clustering. We can calculate the CH score as follows:

$$\mathrm{CH} = \frac{B_k \, (N-k)}{W_k \, (k-1)} \tag{15}$$

where,

- B_k represents the between-cluster dispersion
- w_k represents the within-cluster dispersion
- N represents the number of samples
- k represents the number of clusters

3.6.3 Silhouette Score

Silhouette (SH) score [101] is a clustering quality measure, which depends on the mean intra-cluster distance and the mean-nearest cluster distance for each sample. This score ranges between -1 and 1. Negative value indicates doing wrong clustering. The Silhouette score measuring equation is as follows:

$$SH = \frac{(b-a)}{max(a,b)} \tag{16}$$

where,

- b represents the mean nearest-cluster distance
- a represents the mean intra-cluster distance

3.7 Otsu's Method

Otsu's method [102] is a thresholding technique that separates the data depending on a threshold (t). We choose the optimal value of t by minimizing the within-class variance, or maximizing the between-class variance. Generally, this method is used for unbalanced Raman spectra. Apart from that, people widely use it to separate the background and the foreground of an image. The Otsu's threshold can be calculated from the following equation:

$$t = \arg\max_{t \in \{t_j\}} \sigma_b^2(t) \tag{17}$$

where,

- $\sigma_b^2(t)$ represents the between-class variance

3.8 K-Means + Otsu's Method

In this technique, Otsu's method is applied to the K-Means clustered result. First, the method optimizes the number of clusters (k) between 2 and 5 using the Davies-Bouldin score. From the optimized cluster labels, we need to calculate the centroids to get the cluster intensities (average of each centroid spectrum across bands). Furthermore, this value is used to measure the pixel intensity from the K-Means clustered labels. Since K-Means may fail to separate the tumor and non-tumor cells accurately, Otsu's thresholding method on the calculated pixel intensity helps to separate the tumor and non-tumor cells almost perfectly. If the pixel intensity is greater than or equal to the threshold (t), then this hybrid method identifies the pixel as a tumor cell; otherwise, a non-tumor cell.

3.8.1 Morphological Closing

In the Raman spectra data, some small artifacts still remain, which behave like artifacts or outliers. Some small holes or gaps appear even after applying the combination of K-means and Otsu's method, which can cause an issue when applying the machine learning algorithms. So, using morphological closure [103] can be a better option to solve this problem. This technique helps fill in small gaps in an image while preserving the shape and sizes of other objects. This method also assists in smoothing the object boundaries by connecting nearby objects. This technique can be mathematically expressed as,

$$A \bullet B = (A \oplus B) \ominus B \tag{18}$$

where,

- A represents a binary image
- B represents structuring element
- $\oplus$ represents dilation
- $\ominus$ represents erosion

4 Classifiers to Predict the Glioma Types

To diagnose and treat patients with brain tumors, an accurate classification is essential. In this chapter, we focus on machine learning algorithms which we used to classify the categories of glioma using Raman spectral data. Besides, we added model architectures, data splits, and training strategies used to predict the type of glioma.

4.1 Machine learning models

We used various types of machine learning models-both linear and non-linear to distinguish the spectral structures.

4.1.1 Random Forest (RF) Classifier

Ho developed the idea [104] of combining multiple trees using maximum votes in 1995 using oblique splits. In this method, the forest becomes more accurate when more trees are added without overfitting because each tree only looks at a random subset of features. Later, Leo Breiman properly introduced random forests in one of his papers [105], which has become one of the most cited papers in the world [106]. In his paper, he explained the process of building decision tree forests, where trees are not highly correlated, combining randomization and a CART-like method at each split and bagging [105]. He got this idea from a paper on character recognition, written by Amit and Geman in 1997, where they randomly selected a large number of geometric features for the best split at each node [105].

The general idea behind RF [105] is that when we train the data, it works with multiple small decision trees and classifies the data in each tree. After all trees are classified, bagging is performed, where the prediction or classification is made based on the majority vote of the decision trees. Figure 6 displays the working process or workflow of the RF classifier.

Suppose that θ_k is an independent random vector for the k-th tree, having the same distribution as $\theta_1, \ldots, \theta_{k-1}$. Each tree grows using the training set and θ_k to obtain a classifier $h(\boldsymbol{x}, \theta_k)$, where $\boldsymbol{x}$ is an input vector. Assume that N is the number of examples in the training set. In the bagging process, RF creates a random vector in such a way that we throw N darts into N boxes. In the random split selection process, the randomness depends on how we choose the random numbers between 1 and k. When we have a large number of trees after these procedures, they vote for the most popular class, which is then called the RF classifier.

The margin of the ensemble on a sample $(\boldsymbol{X}, Y)$ is defined as:

$$mg(\boldsymbol{X}, Y) = av_k I(h_k(\boldsymbol{X}) = Y) max_{j=Y} av_k I(h_{k(\boldsymbol{X})} = j) \tag{19}$$

where,

- $I()$ is the indicator function.

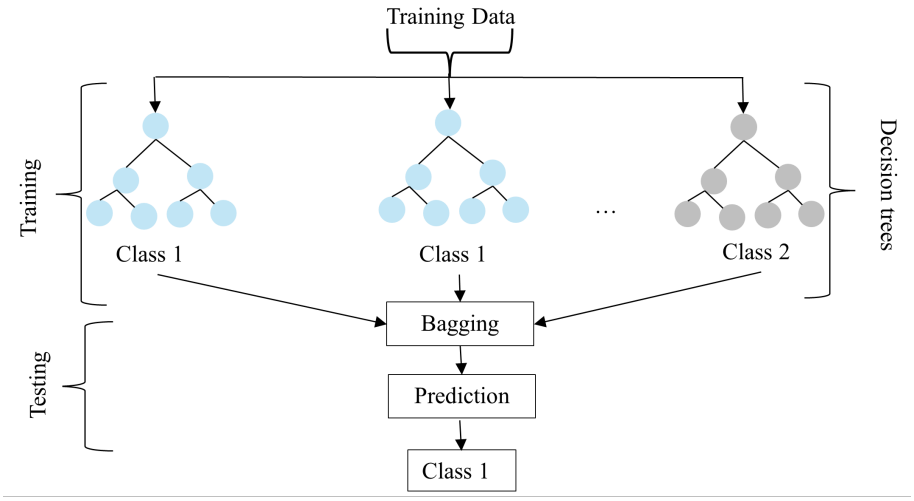


Figure 6: Random Forest workflow [107]. The blue color indicates the trees voting for class 1, and the gray color represents the trees voting for class 2.

This margin measures how many votes the correct class Y exceeds compared to the other class. The confidence of the classification depends on the margin in such a way that if the margin is considerable, then the classification will have more margin. The generalization error is as follows:

$$PE^* = P_{\boldsymbol{X},Y}\big(mg(\boldsymbol{X},Y) < 0\big) \tag{20}$$

where,

- The subscript represents that the probability is over the $\boldsymbol{X}, Y$ space.

The margin function for a random forest is

$$mr(\boldsymbol{X},Y) = P_\theta(h(\boldsymbol{X},\theta) = Y) max_{j \neq Y} P_\theta(h(\boldsymbol{X},\theta) = j) \tag{21}$$

The strength of the set of classifiers $h(\boldsymbol{x},\theta)$ is

$$s = E_{\boldsymbol{X},Y} mr(\boldsymbol{X},Y) \tag{22}$$

Let $s \geq 0$, then Chebyshev's inequality gives

$$PE \leq var(mr)/s^2 \tag{23}$$

where,

- s represents strength, which measures how good an individual tree is on average.

Let the raw margin function for the tree with randomness θ be;

$$rmg(\theta,\boldsymbol{X},Y) = I(h(\boldsymbol{X},\theta) = Y) I(h(\boldsymbol{X},\theta) = \hat{j}(\boldsymbol{X},Y)) \tag{24}$$

where,

- rmg expects to become mg with respect to θ.

Let us assume ρ is the correlation between the rmg of two different trees. Then the variance of the margin becomes

$$var(mr) = E_{\theta,\theta'}(cov_{\boldsymbol{X},Y} rmg(\theta, \boldsymbol{X}, Y) rmg(\theta', \boldsymbol{X}, Y)) = E_{\theta,\theta'}((\theta, \theta') sd(\theta) sd(\theta')) \tag{25}$$

where,

- $sd(\theta)$ represents the standard deviation of $rmg(\theta, \boldsymbol{X}, Y)$ holding θ fixed.

So,

$$var(mr) = \bar{\rho}(E_\theta sd(\theta))^2 \leq \bar{\rho} E_\theta var(\theta) \tag{26}$$

where,

- $\bar{\rho}$ is the average value of the correlation.

So, we can write

$$E_\theta var(\theta) \leq E_\theta(E_{\boldsymbol{X},Y} rmg(\theta, \boldsymbol{X}, Y))^2 s^2 = 1 - s^2 \tag{27}$$

Using equations (23), (26), and (27), Leo Breiman found an upper bound for the generalization error [105] as

$$PE^* \leq \frac{\hat{\rho}(1 - s^2)}{s^2}. \tag{28}$$

During bagging, RF averages the variances of the trees using majority voting, since each tree has a different variance. In addition, the RF algorithm randomly selects a subset of features in each split of a decision tree node. This feature selection procedure reduces the correlation between trees, thereby improving the model's performance. This model tells us that the error stabilizes as the number of trees escalates.

4.1.2 XGBoost Classifier

XGBoost is a fast, accurate, improved, and optimized distributed gradient boosting system [108]. The goal of XGBoost is to use modern computer hardware to provide a scalable and accurate boosting method that can be applied to real-world problems [109]. This technique increases training speed when building trees in parallel by improving traditional gradient boosting [110]. Although it uses smart approximation and practical techniques, it maintains high prediction accuracy and speeds up learning. XGBoost increases the quality of the model by iteratively minimizing the loss function while fitting the new model to the residual of the previous model [109].

XGBoost [111] starts with an initial prediction on the training dataset using the decision trees. After obtaining the initial prediction, it calculates the residuals between the actual and predicted data. To reduce residuals, this method adds decision trees in proportion to the learning rate to adjust the loss function. This process continues until it reaches a point where no further improvement is possible. Figure 7 displays the working process or workflow of the XGBoost classifier.

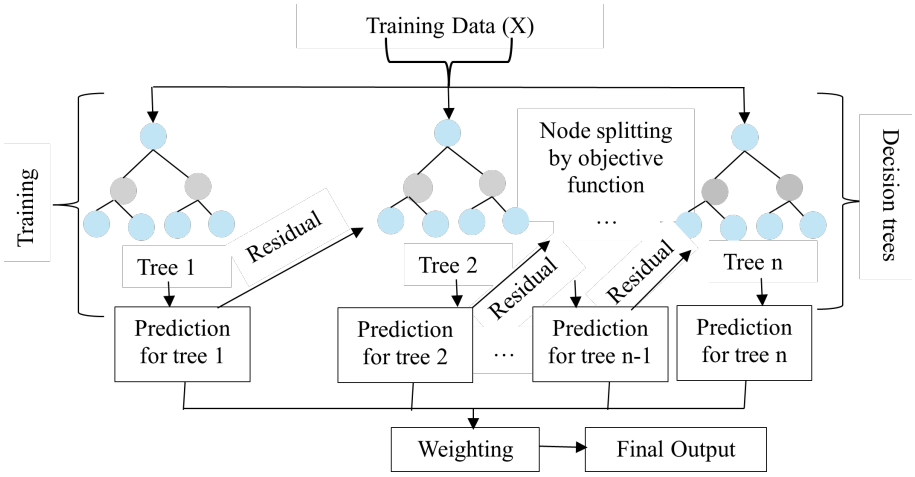


Figure 7: XGBoost classifier workflow [112]. The blue color in the bottom nodes is the leaf nodes containing the weights that are summed to take the final decision. The gray middle nodes are decision nodes, where splitting takes place. The top blue nodes are the root nodes of the tree.

The objective function for classification using log loss or binary entropy is expressed as

$$\theta = \sum_{i=1}^{M} \left[y_i \log\left(1 + e^{\hat{y}_i}\right) \left(1 + y_i\right) \log\left(1 + e^{\hat{y}_i}\right) \right] + \sum_{j=1}^{J} \Omega(f_j) \tag{29}$$

where,

- M represents the samples number in the dataset
- y_i represents the true class label for the i-th sample
- $\hat{y}_i$ represents the positive class predicted probability for the i-th sample
- $\Omega(f_j)$ represents the regularization parameter

In this objective function, for $y_i \log\left(1 + e^{\hat{y}_i}\right)$, when $y_i = 1$, the XGBoost model is penalized whenever the class 0 is predicted. When y_i is large or negative, $e^{\hat{y}_i}$ drops and $\log\left(1 + e^{\hat{y}_i}\right)$ rises. Additionally, for $(1 + y_i) \log\left(1 + e^{\hat{y}_i}\right)$, when $y_i = 0$, the model is penalized if class 1 is predicted. When y_i is large and positive, the model is penalized for classification errors in class 1.

Hence, the predictive probability of XGBoost is

$$p(y = i \mid x_i) = \frac{1}{1 + \exp\left(\sum_{j=1}^{J} f_j(x_i)\right)} \tag{30}$$

where,

- x_i designs the input parameters required for the i-th sample
- $f_j(x_i)$ designs the prediction of the j-th tree of the i-th sample

In Python, XGBoost has several parameters as follows, which help to improve the model performance:

- `n_estimators` represents the number of boosting rounds.
- `learning_rate` controls the contributions of each tree in the final prediction.
- `max_depth` controls the maximum depth of each decision tree.
- `subsample` tells us the fraction of training samples for growing each tree.
- `colsample_bytree` represents the fraction of features when constructing each tree.
- `gamma` controls the tree complexity.
- `reg_alpha` (L1 regularization) and `reg_lambda` (L2 regularization) help to prevent overfitting, where $L1$ selects features and $L2$ reduces weight size.
- `scale_pos_weight` helps to control imbalanced data by giving more importance to the minority.
- `early_stopping_rounds` help to stop training when no improvement is noticed.

4.1.3 Linear Support Vector Classifier (LinearSVC)

Support Vector Machine (SVM) [113] is one of the classical machine learning approaches that can help to classify big data, used especially in multidominal applications. Due to its complexity and expensive behavior, in our thesis, we used the Linear Support Vector Classifier (LinearSVC).

For a two-class or binary classification problem, the classifier predicts using the following function:

$$f(x) = (\boldsymbol{w}\boldsymbol{x}' + \gamma) \tag{31}$$

where,

- $\boldsymbol{x}$ represents the feature vector
- $\boldsymbol{w}$ represents the weight vector
- γ represents bias

The optimization problem for linearly separable data is [114]

$$\begin{aligned} \min_{\boldsymbol{w},\gamma} \quad & \frac{\|\boldsymbol{w}\|^2}{2} \\ \text{subject to} \quad & y(\boldsymbol{w}\boldsymbol{x}' + \gamma) \geq 1 \end{aligned} \tag{32}$$

For multi-class classification, the optimization problem is defined as [115]

$$\begin{aligned} \min_{\boldsymbol{w},\gamma} \quad & \frac{\|\boldsymbol{w}\|^2}{2} \\ \text{subject to} \quad & s(\boldsymbol{w}\boldsymbol{x'} + \gamma \boldsymbol{I}) \geq \boldsymbol{I} \end{aligned} \tag{33}$$

where,

- $\boldsymbol{x}$ represents the n data points
- s represents the response variables of x
- $\boldsymbol{I}$ represents the identity matrix
- γ represents the intercept of the straight line or the hyperplane

In the above equation, the "min" and "subject to" terms indicate the Support Vector Machine measure (SVM-measure) and the label error, respectively.

In case of non-separable classes, a new slack variable is introduced into the previous equation, leading to the following equation [114]:

$$\begin{aligned} \min_{\boldsymbol{w},\gamma} \quad & \frac{\|\boldsymbol{w}\|^2}{2} + \epsilon(\zeta) \\ \text{subject to} \quad & s(\boldsymbol{w}\boldsymbol{x'} + \gamma \boldsymbol{I}) + \zeta \geq \boldsymbol{I} \end{aligned} \tag{34}$$

where,

- ζ represents the slack variable, which describes the acceptance of false positives in the optimization problem.

LinearSVC minimizes the regularized hinge loss to avoid overfitting as follows [116]:

$$\min_{\boldsymbol{w},\gamma} \quad \frac{\|\boldsymbol{w}\|^2}{2} + C\ \max\big(0,\ 1 - s\big(\boldsymbol{w}\boldsymbol{x'} + \gamma \boldsymbol{I}\big)\big) \tag{35}$$

where, $\epsilon(\zeta) = C\zeta$

4.1.4 The Nyström Method

The SVM can also be non-linear. Nyström [117] uses the kernel approximation method to capture the non-linear structure to retain the computational efficiency. Let X be a dataset where m is the number of some samples $(\hat{\boldsymbol{x}_1}, ..., \hat{\boldsymbol{x}_m})$ belonging to X. K be the kernel matrix with these m samples. The Nyström method approximates the matrix K and constructs a low-rank matrix

$$\hat{K_r} = K_b \hat{K}^{\dagger} K_b^T \tag{36}$$

where,

- $K_b = [\kappa(\boldsymbol{x}_i, \hat{\boldsymbol{x}_j})]_{N \times m}$

- $\hat{K} = [\kappa(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)_{m\times m}]$
- $\hat{K}^{\dagger}$ represents the pseudo inverse of $\hat{K}$
- r represents the rank of $\hat{K}$

To train the data, we need to drive the vector representation of the data by

$$\boldsymbol{z}_n(x) = \hat{D_r}^{-1/2}\hat{V_r}^T(\kappa(\boldsymbol{x}, \hat{\boldsymbol{x}}_1), ..., \kappa(\boldsymbol{x}, \hat{\boldsymbol{x}_m}))^T \tag{37}$$

where,

- $\hat{D}_r = diag(\lambda_1, ..., \lambda_r)$
- $\hat{V_r} = (v_1, \hat{...}, \hat{v_r})$

So, it is verified that

$$\boldsymbol{z}_n(x_i)^T\boldsymbol{z}_n(x_j) = [\hat{K_r}]_{ij} \tag{38}$$

After replacing $\boldsymbol{x}$ in the LinearSVC optimization problem by $\boldsymbol{z}_n(x)$ to learn the linear machine $f(x) = \boldsymbol{w}^T\boldsymbol{z}_n(x)$, we get the following optimization problem:

$$\min_{\boldsymbol{w}\in R^r} \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{i=1}^{N}\ell\big(\boldsymbol{w}^{\top}\boldsymbol{z}_n(x_i),\, y_i\big) \tag{39}$$

4.2 Data Splitting Strategy

In our study, we partitioned our Raman spectra data using k-fold cross-validation to evaluate the unbiased model and ensure reproducibility.

4.2.1 K-Fold Cross-Validation(CV)

Cross-validation (CV) [118] is one of the most popular model selection methods after the introduction of leave-one-out cross-validation (LOOCV). The K-fold CV uses (K1) folds to train the data or build the model, and the remaining fold to validate the model. This process continues until the model uses all the folds for validation.

Let us partition the dataset D into K disjoint subsets [119] in such a way that all subsets are of the same size $m \triangleq \frac{n}{K}$. If T_k is the k-th fold (or subset) for validation, then D_k is the training set after removing the elements of T_k from D. In the CV technique, we estimate the error for the validation fold (T_k). After calculating the error from all the validation folds in a repetitive process, we average them to get the final error.

$$\mathrm{CV}(\mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{\boldsymbol{z}_i\in T_k} L\big(\mathcal{A}(D_k), \boldsymbol{z}_i\big) \tag{40}$$

where,

- A represents the learning algorithm
- $\boldsymbol{z}_i$ represents the individual data point in the test set T_k

4.3 Handling Class Imbalance

In the Raman spectroscopy data, we have class imbalance, i.e., the number of tumor samples in each class is unequal, and there is a noticeable difference in sample sizes. This problem can lower the accuracy of our classification model. To get rid of this issue, we performed oversampling and adjusted the class weight.

4.3.1 Oversampling

Random oversampling is a non-heuristic algorithm, which repeats the samples from the minority class randomly to balance it [120]. However, this technique increases the risk of overfitting, as it reproduces the same samples from the minority class [121]. Figure 8 shows how oversampling works by increasing the number of samples in the minority class. Besides, this method makes the learning process overwhelming, although it is a good choice for imbalanced data [120].

Figure 8: Oversampling technique [122]. The orange color highlights the majority class, and the blue color denotes the minority class.

4.3.2 Class-Weight Adjustments

The models we used in our thesis have a class weight parameter. If we set the parameter to "balanced," the model will penalize misclassifications of minority classes more heavily, thereby improving its sensitivity.

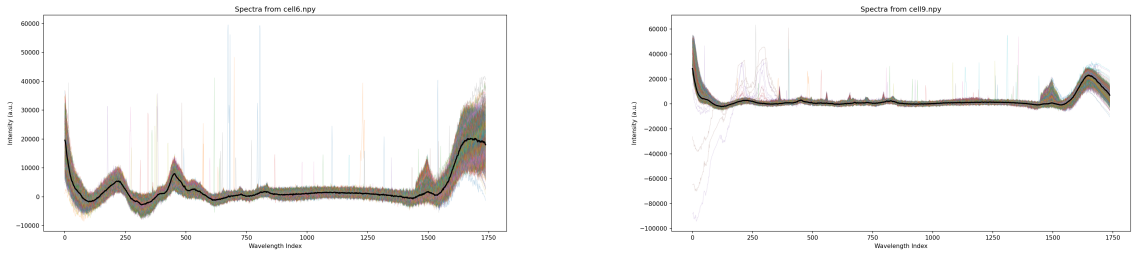
5 Results and Discussions

This chapter illustrates what we found after our analysis. First, the outcomes of each of the preprocessing steps are shown, and then how we tried to balance our data are discussed. Also, we demonstrate the classifiers of the glioma IDH1 mutation status.

5.1 Data Preprocessing

We obtained Raman spectral data from live glioma cells in two experimental phases conducted in January 2025 and May 2025 by NIH. In the first phase, we obtained spectral data from 85 different cells, which were then extended to 286 individual cells in the second phase. During data preprocessing, we identified and removed unusable data files due to acquisition errors from our dataset. Specifically, the deletion of files cell43 and cell248 resulted in a final Raman spectra dataset of 284 live glioma cell files.

According to the initial inspection, our raw data has a significant level of cosmic rays and noise across many spectra. To gain a clear understanding of this issue, we have included 2 reference pictures of the raw data in Figure 9. We have uploaded the rest of the figures to this link:`https://seafile.utu.fi/d/46d34af2160e438097e8/`. The Raman spectra data for each sample(Figure 9) exhibit a negative direction and showcase sudden peaks in the fingerprint region (0-3400cm^{-1}) due to artifacts and water or the slowly varying background signal (baseline drift). In each sample, the median Raman spectra indicate that each of the Raman spectra should be close to the median. Otherwise, they can have artifacts, cosmic rays, and noise, which are essential to remove from the data for further analysis.

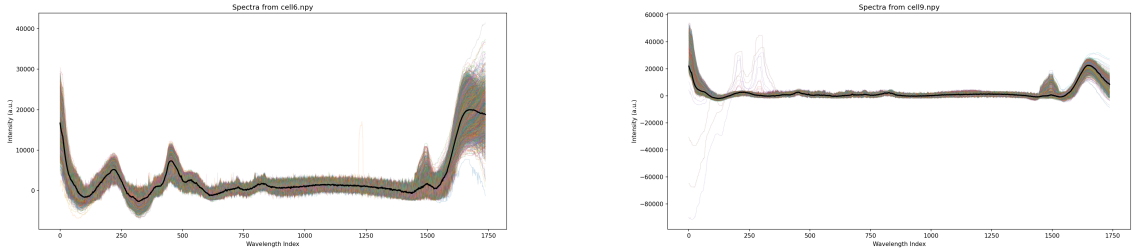


(a) (b)

Figure 9: Per spectra in raw data with their median. The random spikes in graphs (a: cell6) and (b: cell9) are the cosmic rays. In both graphs, the data includes noise and meaningless regions (silent regions). Each color of graphs (a) and (b) represents the spectra of each sample/cell.

To resolve this issue, we developed an optimized pipeline for cosmic rays detection and correction. According to this process, we selected the optimized values for the height multiplier and the maximum width as 5 and 1, respectively. Additionally, we estimated the median absolute deviation (MAD) for each cell and selected $p1 = 3 * median(MAD)$ to replace cosmic rays with the median values, and $p2 = 11$

for the window size. Figure 10 indicates that our pipeline has successfully removed most of the cosmic rays from our data because very few spikes are visible in some spectra. We have uploaded all the figures after removing the cosmic rays in the Seafile (Link:`https://seafile.utu.fi/d/1c22585dd0cc4d8a83bc/`).



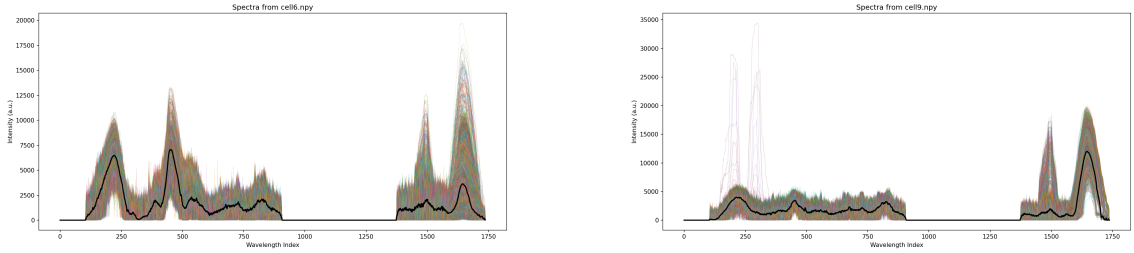
(a) (b)

Figure 10: Per spectra data after removing cosmic rays with their median. Almost most of the cosmic rays from graphs (a: cell6) and (b: cell9) are removed. Only some baseline noise are still present.

We tried to reduce the noise in the reshaped data using the Savitzky-Golay (SG) filter with a polynomial order of 3 and a window size of 11. However, the average root-mean-square error (RMSE) across the dataset (total files=284) revealed that the SG-filter modifies the cosmic rays removed data by approximately 1.62% relative to the maximum spectral intensity. Visual inspection also did not make any significant difference in our dataset. Hence, we dropped the use of this filtering method from our analysis.

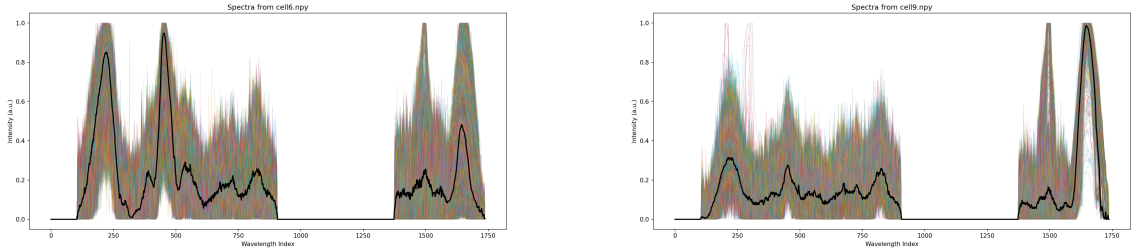
Following denoising, we tried to fix the non-informative regions of the spectra using baseline correction. For that, we clipped the remaining negative intensity values at 0. Additionally, we removed the non-informative silent regions, between $0 - 248.651529$ cm^{-1}, and $1797.210 - 2697.804$ cm^{-1}, and replaced these regions with 0 intensity values. This silent region removal technique divided the spectra into two chemically informative regions, and the zero padding technique preserved the matrix structure to continue machine learning applications. We used the airPLS algorithm with parameters $\lambda = 4$, and $p - value = 1$ to approximate the non-silent regions. Figure 11 showcases that the negative spectra have been clipped to 0, and the silent regions were padded with 0 to avoid any arising problem during the application of the machine learning approach. The rest of the figure of the sample cells can be found in this link:`https://seafile.utu.fi/d/088818a32a284bff9a69/`.

To prepare the data for further computational analysis, we applied min-max normalization and L2-normalization to each spectrum to obtain all intensity values between 0 and 1. But, we chose min-max normalization for our further analysis to maintain the simplicity of our analysis. Figure 12 displays that all spectra of each sample cell have ranged between 0 and 1. The rest of the figure of the sample cells can be found in the following Seafile link:`https://seafile.utu.fi/d/f9170f3f18a84e8a8261/`.



(a) (b)

Figure 11: Raman spectra data after removing the silent region and baseline correction. The base for both of the samples is corrected, and the non-informative regions are removed. The intensities of graphs (a: cell6) and (b: cell9) do not fall under the same range.

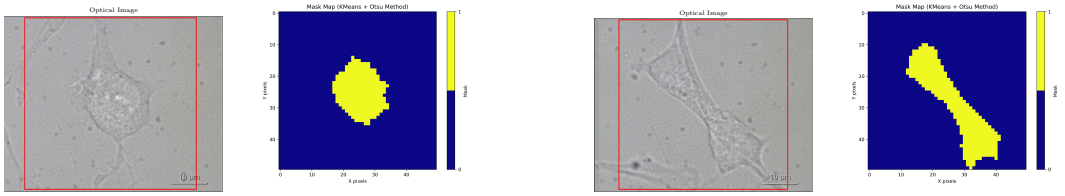


(a) (b)

Figure 12: Raman spectra data after normalization (min-max). All the intensities of graphs (a: cell6) and (b: cell9) fall between the range 0 and 1.

After normalization, we performed a clustering analysis to find group similarities among the spectra. For this, we tried several clustering algorithms, such as k-means, agglomerative, birch, and Otsu's thresholding method on the whole Raman shift. However, the cluster map was very poor throughout the range ($50.019449 - 3399.76628\,cm^{-1}$). As a result, we applied our clustering algorithm to the Raman shift from $2800.0131\,cm^{-1}$ to $3025.64372\,cm^{-1}$, because in this region the intensity of the tumor is higher than that of the non-tumor. To optimize the cluster number, we evaluated three cluster performance indices named the DB score, CH score, and SH score. Among all of these indices, the DB score performed better compared to the other two. Although when optimizing the number of clusters for k-means clustering, the DB score had a suboptimal performance, which led to a poorer clustering result. Among all the clustering algorithms that we used, the combination of k-means clustering and Otsu's method separated the similar spectra better. To get rid of the small holes or gaps, we applied morphological closing using a minimum object size of 7, and a disk of 3. This combined method successfully separated the spectra into two groups (cell and water). A summary of Otsu's thresholds and the corresponding Davies–Bouldin scores to optimize the number of clusters for k-means clustering for each file is presented in the chart titled `Otsu_DBI_Cluster_Summary.xlsx` (link: `https://seafile.utu.fi/f/39ac686d3d0143c3a7ac/`). This approach was successful for most of the samples, as shown in Figure 13, where we compared

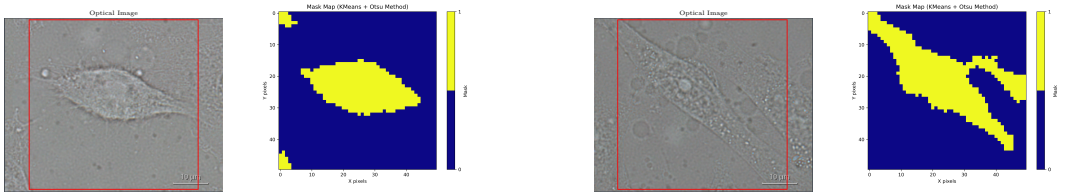
the optical images of the cells and the clustering images of some samples. (link: `https://seafile.utu.fi/d/958141f43ffc4108ad91/`)



(a) (b)

Figure 13: Comparison between optical image and clustering image (a: cell6, b: cell9). The yellow region in both of the graphs indicates the tumor cells (labeled as 1), and the blue region represents water (labeled as 0) from the cell-culture.

Some samples (14 samples) showed poor separation with this method with normalized data, such as cell77, cell79, cell82, cell255, cell256, cell257, cell258, cell259, cell260, cell261, cell266, cell267, cell268, and cell269. However, when we analyzed the data without normalization, we successfully recovered them. The reason behind this is that normalization helps to correct the experimental differences, such as focus drift, which improves classification. However, for some samples, the absolute intensity of the spectra plays a significant role. Normalization blends these intensities with the main intensity by reducing the intensity information. So, the cluster map on the non-normalized data for these samples showed better images compared to the cluster map of the normalized data. Figure 14 showcases some of the samples in which we recovered the clustered data. Here, in the 2D spatial representation, the yellow pixels denote tumor spectra, and the blue pixels denote non-tumor spectra or water. The other sample examples can be found on Seafile: link:`https://seafile.utu.fi/d/074bb4d4b2464f00a736/`. The clustering results identified 300184 spectra (42.54% of the data) as tumor spots and 405509 spectra (57.46% of the data) as non-tumor spots.



(a) (b)

Figure 14: Comparison of the original optical images and the clustering results obtained from non-normalized data. The yellow regions in graphs (a: cell77) and (b: cell79) represent the tumor cells, and the blue regions are the water.

5.2 Discrimination between IDH1mut and IDH1wt

We used random forests and XGBoost to identify the most important Raman frequencies for discriminating IDH1mut and IDH1wt tumor types across all 284 Raman

samples. For that, we split our data into a 10-fold cross-validation. The samples were distributed equally across each fold based on IDH1mut and IDH1wt. We excluded the samples (16 samples: cell61-cell76) that are GBM_IDH1mut because these are cells that have the chemistry of the GBM (which normally comes with IDH1WT), but with the mutation IDH1mut inserted in them. Since, IDH1mut is the minority class (51 samples) compared to IDH1wt (211 samples), we applied oversampling on the minority class in such a way that the minority samples in each fold repeat them until they become the same number of samples of the majority class in that fold. However, the XGBoost classifier performs better compared to the RF in this case. To measure the accuracy of our classifier model, we calculated balanced accuracy, ROC-AUC, precision, recall, F1, and a confusion matrix for each fold used as it behaved like a validation fold. After obtaining the results from each fold, we measured the mean accuracy for all the metrics that we have performed in this study. Each fold accuracy with its final mean results and confusion matrix is shown in Table 1, and Table 2.

Table 1: Cross-validation performance metrics for XGBoost classifier across 10 folds.

Fold	Accuracy	ROC-AUC	Precision		Recall		F1-score	
			Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
1	0.551	0.551	0.164	0.887	0.681	0.421	0.265	0.571
2	0.632	0.632	0.132	0.951	0.658	0.607	0.220	0.741
3	0.472	0.472	0.198	0.763	0.390	0.553	0.263	0.641
4	0.588	0.588	0.227	0.896	0.355	0.821	0.277	0.857
5	0.739	0.739	0.581	0.895	0.583	0.894	0.582	0.895
6	0.727	0.727	0.311	0.940	0.737	0.717	0.437	0.813
7	0.603	0.603	0.269	0.882	0.402	0.803	0.323	0.841
8	0.536	0.536	0.155	0.893	0.263	0.810	0.195	0.849
9	0.530	0.530	0.455	0.724	0.116	0.943	0.185	0.819
10	0.502	0.502	0.215	0.794	0.089	0.916	0.125	0.851
Mean	**0.588**	**0.588**	**0.271**	**0.862**	**0.427**	**0.749**	**0.287**	**0.788**

The overall mean accuracy (0.588) of the XGBoost classifier across the 10 folds indicates moderate classification performance. ROC-AUC scores indicate that the model's ability to distinguish between two classes is moderate (0.588). The significant discrepancy in the ROC-AUC values suggests a class imbalance or a model bias towards class 1 (IDH1wt), which is reflected in the confusion matrices. The mean precision values for class 0 (0.271) and class 1 (0.862) tell that our XGBoost model is more precise when classifying IDH1wt. Similarly, the recall for IDH1wt is higher (0.749) than for IDH1mut (0.427). That means a considerable number of IDH1mut are misclassified as IDH1wt. The F1-score helps to balance the precision and recall scores. This score (IDH1mut: 0.287, IDH1wt: 0.788) also justifies the result of the precision and recall score.

The confusion matrices (Table 2) assist to measure the class-specific performance of our XGBoost model. This matrix demonstrates that only 42.7% of IDH1mut is classified correctly, while 57.3% of it is misclassified as IDH1wt. On the other

Table 2: Confusion matrices of the XGBoost classifier for each fold.

Fold	True Class 0		True Class 1	
	Pred 0	Pred 1	Pred 0	Pred 1
1	0.681	0.319	0.579	0.421
2	0.658	0.342	0.393	0.607
3	0.390	0.610	0.447	0.553
4	0.355	0.645	0.179	0.821
5	0.583	0.417	0.106	0.894
6	0.737	0.263	0.283	0.717
7	0.402	0.598	0.197	0.803
8	0.263	0.737	0.190	0.810
9	0.116	0.884	0.057	0.943
10	0.089	0.911	0.084	0.916
Mean	**0.427**	**0.573**	**0.251**	**0.749**

hand, the XGBoost classifier perfectly classified around 74.9% of IDH1wt, as well as misclassified it for approximately 25.1%. This confirms that our XGBoost model gives more importance to class 1 (IDH1wt) while classifying.

Although the XGBoost performs better than the other models, the accuracy of this model is still poor, and it also has an imbalance among the mutation statuses. To overcome this problem, we chose the top 23 features and applied the Nyström approximation and Linear SVM on top of these specific features. As a result, we got that the model performance accuracy with the confusion matrix (class 0 (IDH1mut): 60.7%, class 1 (IDH1wt): 58.5%) has become balanced and improved.

6 Conclusions

Raman spectroscopy is a method in which the laser light from the spectrometer tells us about the details of tissue composition at the biochemical level by changing the color of the light. From this change, we obtain the spectra of the cells. However, these resulting spectra are tough to understand just by looking at them. In this thesis, we developed a ML system which can remove the cosmic rays, correct the baseline after removing the non-informative regions, normalize the Raman spectra data, and can separate the tumor and non-tumor cells accurately. This ML approach to classify the glioma live cells can be non-invasive, fast, and non-destructive. In this section, we have discussed some key findings and contributions of our ML approach to classify the live cells of glioma.

In this work, we analyzed Raman spectra from 284 glioma live cells by fixing our goal to improve the data preprocessing pipeline, classifying glioma types and their mutation status. Our data includes various types of gliomas that allowed us to have a detailed experiment of them. Because of the features of capturing a mixture of chemical signals, Raman spectroscopy data have noise, artifacts, and baseline drift. As a result, we first established and validated a proper machine learning data preprocessing pipeline for glioma live cells.

One of the most important components of this preprocessing workflow is our custom cosmic rays removal technique. This custom method is an optimized way to detect random spikes in our RS data and remove them entirely from our data. These spikes or cosmic rays can mislead the ML algorithm if they are not handled properly. Furthermore, we developed our preprocessing pipeline by removing the silent regions. The silent region is a region that does not contain any biologically meaningful information, which can affect baseline estimation. Following this, we corrected the baseline using the airPLS algorithm in such a way that the spectra can only reflect the original fingerprints of the cells. Besides, we also applied the min-max normalization on our baseline-corrected Raman data, so that we could compare different cells. Moreover, this normalization helped us to preserve the chemically significant features even after processing our data. This whole preprocessing framework prepared our RS data for the application of clustering and classification.

The SG filter is a smoothing method that can be applied to RS data. During our preprocessing operation flow, we applied this filter to our data. However, we found out visually and quantitatively that it did not have any specific value in smoothing our RS data. The average RMSE demonstrated that the SG-filter modified the cosmic rays removed data by approximately 1.62% relative to the maximum spectral intensity. As a result, we excluded the SG-filter from our final RS data preprocessing process, so that we did not lose the important features.

To separate tumor and non-tumor cells, we applied a combined clustering method (K-means + Otsu's thresholding method). In this procedure, K-means first groups spectra, depending on their biochemical similarities, and Otsu's method determines

an optimal threshold for each sample to separate the preprocessed data into tumor and non-tumor cells. To avoid the noisy cluster, we applied morphological closing to group the tumor and non-tumor cells properly and smoothly. Finally, this technique gave us 42.54% tumor spectra and 57.46% non-tumor spectra, indicating the necessity of performing a good clustering.

An unusual finding occurred when we performed our clustering method on both normalized and non-normalized data. For some samples, the clustering method worked better on non-normalized data, compared to normalized data. This is an important case, which tells us that although normalization is crucial, but sometimes it can distort natural biochemical differences.

Next, we tried to develop a ML model that can classify the IDH1 mutation status (IDH1mut vs IDH1wt) of glioma live cells. Among all the algorithms that we applied, XGBoost performs better, although the performance is still poor. This is mainly because of having the imbalanced properties in our dataset. To overcome this issue, we tried to duplicate the samples in the minority class in the same cross-validation fold, and then chose the top 23 features after applying the XGBoost classifier. This feature selection technique helps to obtain a ML classifier model based on the most important features in our data set. When we applied the Nyström kernel approximation and a Linear SVM on these 23 features, it allowed us to capture the non-linear spectral features. This procedure improved and balanced the classification accuracy, indicating that careful selection of important features can reduce the class imbalance issue.

Our thesis establishes a developed preprocessing and clustering pipeline of Raman spectra data. At the same time, it has opened new directions to advance the research by enhancing classification performance and so on. We plan to continue our research in the following directions:

Firstly, since we duplicated the samples in each cross-validation fold until it had the same number of samples as the majority class, so the XGBoost model had the probability to overfit. This is because, while doing this, it memorizes the pattern of the spectra instead of learning the features. We plan to check the model whether it fits perfectly or not.

Secondly, we plan to enhance the number of selected features in our model from $50 - 100$. Increasing the number of selected features may help the classifier model obtain more detailed spectral information that we lost in our analysis.

Thirdly, we plan to enhance our classification from 2-class to 3-class (GBM vs. Astro vs. Oligo). Although it will be challenging, but it will assist us to get a more detailed understanding of glioma subtypes.

Finally, we plan to look for the answer to the question "How do we separate Astro and GBM when they are spectrally very similar?"

References

[1] Kirti Raj Bhatele and Sarita Singh Bhadauria. Machine learning application in glioma classification: review and comparison analysis. *Archives of Computational Methods in Engineering*, 29:247–274, 2022. DOI: https://doi.org/10.1007/s11831-021-09572-z.

[2] Benedikt Wiestler, Anne Kluge, Mathias Lukas, Jens Gempt, Florian Ringel, Jürgen Schlegel, Bernhard Meyer, Claus Zimmer, Stefan Förster, Thomas Pyka, et al. Multiparametric mri-based differentiation of who grade ii/iii glioma and who grade iv glioblastoma. *Scientific Reports*, 6(1):35142, 2016. DOI: https://doi.org/10.1038/srep35142.

[3] Quinn T Ostrom, Luc Bauchet, Faith G Davis, Isabelle Deltour, James L Fisher, Chelsea Eastman Langer, Melike Pekmezci, Judith A Schwartzbaum, Michelle C Turner, Kyle M Walsh, et al. The epidemiology of glioma in adults: a "state of the science" review. *Neuro-Oncology*, 16(7):896–913, 2014. DOI: https://doi.org/10.1093/neuonc/nou087.

[4] Hai Yan, D. Williams Parsons, Genglin Jin, Roger McLendon, B. Ahmed Rasheed, Weishi Yuan, Ivan Kos, Ines Batinic-Haberle, Siân Jones, Gregory J. Riggins, Henry Friedman, Allan Friedman, David Reardon, James Herndon, Kenneth W. Kinzler, Victor E. Velculescu, Bert Vogelstein, and Darell D. Bigner. Idh1 and idh2 mutations in gliomas. *New England Journal of Medicine*, 360(8):765–773, 2009. DOI: https://doi.org/10.1056/NEJMoa0808710.

[5] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvet, Bernd W Scheithauer, and Paul Kleihues. The 2007 who classification of tumours of the central nervous system. *Acta Neuropathologica*, 114(2):97–109, 2007. DOI: https://doi.org/10.1007/s00401-007-0243-4.

[6] Adam L Cohen, Sheri L Holmen, and Howard Colman. Idh1 and idh2 mutations in gliomas. *Current Neurology and Neuroscience Reports*, 13(5):345, 2013. DOI: https://doi.org/10.1007/s11910-013-0345-4.

[7] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820, 2016. DOI: https://doi.org/10.1007/s00401-016-1545-1.

[8] David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro Oncology*, 23(8):1231–1251, 2021. DOI: https://doi.org/10.1093/neuonc/noab106.

[9] OncoLink Team. All about pediatric gliomas (low and high grade). https://www.oncolink.org/cancers/brain-tumors/all-about-pediatric-gliomas-low-and-high-grade", 2020. Accessed: 2025-11-30.

[10] Yan Tan, Shuai-tong Zhang, Jing-wei Wei, Di Dong, Xiao-chun Wang, Guoqiang Yang, Jie Tian, and Hui Zhang. A radiomics nomogram may improve the prediction of idh genotype for astrocytoma before surgery. *European Radiology*, 29(7):3325–3337, 2019. DOI: https://doi.org/10.1007/s00330-019-06056-4.

[11] Claire L MacIver, Ayisha Al Busaidi, Balaji Ganeshan, John A Maynard, Stephen Wastling, Harpreet Hyare, Sebastian Brandner, Julia E Markus, Martin A Lewis, Ashley M Groves, et al. Filtration-histogram based magnetic resonance texture analysis (mrta) for the distinction of primary central nervous system lymphoma and glioblastoma. *Journal of Personalized Medicine*, 11(9):876, 2021. DOI: https://doi.org/10.3390/jpm11090876.

[12] Chendan Jiang, Ziren Kong, Sirui Liu, Shi Feng, Yiwei Zhang, Ruizhe Zhu, Wenlin Chen, Yuekun Wang, Yuelei Lyu, Hui You, et al. Fusion radiomics features from conventional mri predict mgmt promoter methylation status in lower grade gliomas. *European Journal of Radiology*, 121:108714, 2019. DOI: https://doi.org/10.1016/j.ejrad.2019.108714.

[13] Chia-Feng Lu, Fei-Ting Hsu, Kevin Li-Chun Hsieh, Yu-Chieh Jill Kao, Sho-Jen Cheng, Justin Bo-Kai Hsu, Ping-Huei Tsai, Ray-Jade Chen, Chao-Ching Huang, Yun Yen, et al. Machine learning–based radiomics for molecular subtyping of gliomas. *Clinical Cancer Research*, 24(18):4429–4436, 2018. DOI: https://doi.org/10.1158/1078-0432.CCR-17-3445.

[14] MHT Reinges, H-H Nguyen, T Krings, B-O Hütter, V Rohde, and JM Gilsbach. Course of brain shift during microsurgical resection of supratentorial cerebral lesions: limits of conventional neuronavigation. *Acta Neurochirurgica*, 146(4):369–377, 2004. DOI: https://doi.org/10.1007/s00701-003-0204-1.

[15] Todd Hollon and Daniel A Orringer. Label-free brain tumor imaging using raman-based methods. *Journal of Neuro-Oncology*, 151(3):393–402, 2021. DOI: https://doi.org/10.1007/s11060-019-03380-z.

[16] Steven N Kalkanis, Rachel E Kast, Mark L Rosenblum, Tom Mikkelsen, Sally M Yurgelevic, Katrina M Nelson, Aditya Raghunathan, Laila M Poisson, and Gregory W Auner. Raman spectroscopy to distinguish grey matter, necrosis, and glioblastoma multiforme in frozen tissue sections. *Journal of Neuro-Oncology*, 116(3):477–485, 2014. DOI: https://doi.org/10.1007/s11060-013-1326-9.

[17] Michael Jermyn, Kelvin Mok, Jeanne Mercier, Joannie Desroches, Julien Pichette, Karl Saint-Arnaud, Liane Bernstein, Marie-Christine Guiot, Kevin

Petrecca, and Frederic Leblond. Intraoperative brain cancer detection with raman spectroscopy in humans. *Science Translational Medicine*, 7(274):274ra19, 2015. DOI: https://doi.org/10.1126/scitranslmed.aaa2384.

[18] Ortrud Uckermann, Wenmin Yao, Tareq A Juratli, Roberta Galli, Elke Leipnitz, Matthias Meinhardt, Edmund Koch, Gabriele Schackert, Gerald Steiner, and Matthias Kirsch. Idh1 mutation in human glioma induces chemical alterations that are amenable to optical raman spectroscopy. *Journal of Neuro-Oncology*, 139(2):261–268, 2018. DOI: https://doi.org/10.1007/s11060-018-2883-8.

[19] Michele Ceccarelli, Floris P Barthel, Tathiane M Malta, Thais S Sabedot, Sofie R Salama, Bradley A Murray, Olena Morozova, Yulia Newton, Amie Radenbaugh, Stefano M Pagnotta, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563, 2016. DOI: https://doi.org/10.1016/j.cell.2015.12.028.

[20] David Capper, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, et al. Dna methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018. DOI: https://doi.org/10.1038/nature26000.

[21] Zhichao Wu, Zied Abdullaev, Drew Pratt, Hye-Jung Chung, Shannon Skarshaug, Valerie Zgonc, Candice Perry, Svetlana Pack, Lola Saidkhodjaeva, Sushma Nagaraj, et al. Impact of the methylation classifier and ancillary methods on cns tumor diagnostics. *Neuro-Oncology*, 24(4):571–581, 2022. DOI: https://doi.org/10.1093/neuonc/noab227.

[22] Zane Jaunmuktane, David Capper, David TW Jones, Daniel Schrimpf, Martin Sill, Monika Dutt, Nirosha Suraweera, Stefan M Pfister, Andreas von Deimling, and Sebastian Brandner. Methylation array profiling of adult brain tumours: diagnostic outcomes in a large, single centre. *Acta Neuropathologica Communications*, 7(1):24, 2019. DOI: https://doi.org/10.1186/s40478-019-0668-8.

[23] Sevin Turcan, Daniel Rohle, Anuj Goenka, Logan A Walsh, Fang Fang, Emrullah Yilmaz, Carl Campos, Armida WM Fabius, Chao Lu, Patrick S Ward, et al. Idh1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483:479–483, 2012. DOI: https://doi.org/10.1038/nature10866.

[24] Houtan Noushmehr, Daniel J Weisenberger, Kristin Diefes, Heidi S Phillips, Kanan Pujara, Benjamin P Berman, Fei Pan, Christopher E Pelloski, Erik P Sulman, Krishna P Bhat, et al. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–522, 2010. DOI: https://doi.org/10.1016/j.ccr.2010.03.017.

[25] Camila Ferreira de Souza, Thais S Sabedot, Tathiane M Malta, Lindsay Stetson, Olena Morozova, Artem Sokolov, Peter W Laird, Maciej Wiznerowicz, Antonio Iavarone, James Snyder, et al. A distinct dna methylation shift in a subset of glioma cpg island methylator phenotypes during tumor recurrence. *Cell Reports*, 23(2):637–651, 2018. DOI: https://doi.org/10.1016/j.celrep.2018.03.107.

[26] Yafeng Qi, Guochao Zhang, Lin Yang, Bangxu Liu, Hui Zeng, Qi Xue, Dameng Liu, Qingfeng Zheng, and Yuhong Liu. High-precision intelligent cancer diagnosis method: 2d raman figures combined with deep learning. *Analytical Chemistry*, 94(17):6491–6501, 2022. DOI: https://doi.org/10.1021/acs.analchem.1c05098.

[27] Laurent J Livermore, Martin Isabelle, Ian M Bell, Oliver Edgar, Natalie L Voets, Richard Stacey, Olaf Ansorge, Claire Vallance, and Puneet Plaha. Raman spectroscopy to differentiate between fresh tissue samples of glioma and normal brain: a comparison with 5-ala–induced fluorescence-guided surgery. *Journal of Neurosurgery*, 135(2):469–479, 2020. DOI: https://doi.org/10.3171/2020.5.JNS20376.

[28] Laurent James Livermore, Martin Isabelle, Ian Mac Bell, Connor Scott, John Walsby-Tickle, Joan Gannon, Puneet Plaha, Claire Vallance, and Olaf Ansorge. Rapid intraoperative molecular genetic classification of gliomas using raman spectroscopy. *Neuro-Oncology Advances*, 1(1):vdz008, 2019. DOI: https://doi.org/10.1093/noajnl/vdz008.

[29] Ortrud Uckermann, Jonathan Ziegler, Matthias Meinhardt, Sven Richter, Gabriele Schackert, Ilker Y Eyüpoglu, Mido M Hijazi, Dietmar Krex, Tareq A Juratli, Stephan B Sobottka, et al. Raman and autofluorescence spectroscopy for in situ identification of neoplastic tissue during surgical treatment of brain tumors. *Journal of Neuro-Oncology*, 170(3):543–553, 2024. DOI: https://doi.org/10.1007/s11060-024-04809-w.

[30] Marco Riva, Tommaso Sciortino, Riccardo Secoli, Ester D'amico, Sara Moccia, Bethania Fernandes, Marco Conti Nibali, Lorenzo Gay, Marco Rossi, Elena De Momi, et al. Glioma biopsies classification using raman spectroscopy and machine learning models on fresh tissue samples. *Cancers*, 13(5):1073, 2021. DOI: https://doi.org/10.3390/cancers13051073.

[31] Denis Vrazhnov, Anna Mankova, Evgeny Stupak, Yury Kistenev, Alexander Shkurinov, and Olga Cherkasova. Discovering glioma tissue through its biomarkers' detection in blood by raman spectroscopy and machine learning. *Pharmaceutics*, 15(1):203, 2023. DOI: https://doi.org/10.3390/pharmaceutics15010203.

[32] Liang Zhang, Yan Zhou, Binlin Wu, Shengjia Zhang, Ke Zhu, Cheng-Hui Liu, Xinguang Yu, and Robert R Alfano. A handheld visible resonance raman analyzer used in intraoperative detection of human glioma. *Cancers*, 15(6):1752, 2023. DOI: https://doi.org/10.3390/cancers15061752.

[33] Katherine Ember, Frédérick Dallaire, Arthur Plante, Guillaume Sheehy, Marie-Christine Guiot, Rajeev Agarwal, Rajeev Yadav, Alice Douet, Juliette Selb, Jean Philippe Tremblay, et al. In situ brain tumor detection using a raman spectroscopy system—results of a multicenter study. *Scientific Reports*, 14(1):13309, 2024. DOI: https://doi.org/10.1038/s41598-024-62543-9.

[34] Gilbert Georg Klamminger, Laurent Mombaerts, Françoise Kemp, Finn Jelke, Karoline Klein, Rédouane Slimani, Giulia Mirizzi, Andreas Husch, Frank Hertel, Michel Mittelbronn, et al. Machine learning-assisted classification of paraffin-embedded brain tumors with raman spectroscopy. *Brain Sciences*, 14(4):301, 2024. DOI: https://doi.org/10.3390/brainsci14040301.

[35] Adrian Lita, Joel Sjöberg, David Păcioianu, Nicoleta Siminea, Orieta Celiku, Tyrone Dowdy, Andrei Păun, Mark R Gilbert, Houtan Noushmehr, Ion Petre, et al. Raman-based machine-learning platform reveals unique metabolic differences between idhmut and idhwt glioma. *Neuro-Oncology*, 26(11):1994–2009, 2024. DOI: https://doi.org/10.1093/neuonc/noae101.

[36] John R Ferraro, Kazuo Nakamoto, and Chris W. and Brown. *Introductory raman spectroscopy*. Elsevier, 2003. DOI: https://doi.org/10.1016/B978-0-12-254105-6.X5000-8.

[37] Yafeng Qi, Yuhong Liu, and Jianbin Luo. Recent application of raman spectroscopy in tumor diagnosis: from conventional methods to artificial intelligence fusion. *PhotoniX*, 4(1):22, 2023. DOI: https://doi.org/10.1186/s43074-023-00098-0.

[38] Nicole M. Ralbovsky and Igor K. Lednev. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chemical Society Reviews*, 49(20):7428–7453, 2020. DOI: https://doi.org/10.1039/d0cs01019g.

[39] Adolf Smekal. Zur quantentheorie der dispersion. *Naturwissenschaften*, 11:873–875, 1923. DOI: https://doi.org/10.1007/BF01576902.

[40] Chandrasekhara Venkata Raman and Kariamanikkam Srinivasa Krishnan. A new type of secondary radiation. *Nature*, 121:501–502, 1928. DOI: https://doi.org/10.1038/121501c0.

[41] Ruchita S Das and YK Agrawal. Raman spectroscopy: Recent advancements, techniques and applications. *Vibrational Spectroscopy*, 57(2):163–176, 2011. DOI: https://doi.org/10.1016/j.vibspec.2011.08.003.

[42] Robin R Jones, David C Hooper, Liwu Zhang, Daniel Wolverson, and Ventsislav K Valev. Raman techniques: fundamentals and frontiers. *Nanoscale Research Letters*, 14(1):231, 2019. DOI: https://doi.org/10.1186/s11671-019-3039-2.

[43] Saga Bergqvist. Raman spectroscopy in neurosurgery. Master's Thesis, Department of Computer Science, Electrical and Space Engineering, 2020. Available at: https://ltu.diva-portal.org/smash/get/diva2:1426395/FULLTEXT01.pdf.

[44] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, et al. Using raman spectroscopy to characterize biological materials. *Nature Protocols*, 11:664–687, 2016. DOI: https://doi.org/10.1038/nprot.2016.036.

[45] Thomas Hellerer, Claes Axäng, Christian Brackmann, Per Hillertz, Marc Pilon, and Annika Enejder. Monitoring of lipid storage in caenorhabditis elegans using coherent anti-stokes raman scattering (cars) microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37):14658–14663, 2007. Available at: http://www.jstor.org/stable/25449009.

[46] Olaf Maier, Volker Oberle, and Dick Hoekstra. Fluorescent lipid probes: some properties and applications (a review). *Chemistry and Physics of Lipids*, 116(1-2):3–18, 2002. DOI: https://doi.org/10.1016/S0009-3084(02)00017-8.

[47] Artem Pliss, Andrey N Kuzmin, Adrian Lita, Rahul Kumar, Orieta Celiku, G Ekin Atilla-Gokcumen, Omer Gokcumen, Dhyan Chandra, Mioara Larion, and Paras N Prasad. A single-organelle optical omics platform for cell science and biomarker discovery. *Analytical Chemistry*, 93(23):8281–8290, 2021. DOI: https://doi.org/10.1021/acs.analchem.1c01131.

[48] Adrian Lita, Andrey N Kuzmin, Artem Pliss, Alexander Baev, Alexander Rzhevskii, Mark R Gilbert, Mioara Larion, and Paras N Prasad. Toward single-organelle lipidomics in live cells. *Analytical Chemistry*, 91(17):11380–11387, 2019. DOI: https://doi.org/10.1021/acs.analchem.9b02663.

[49] Andrey N Kuzmin, Artem Pliss, Alex Rzhevskii, Adrian Lita, and Mioara Larion. Bcabox algorithm expands capabilities of raman microscope for single organelles assessment. *Biosensors*, 8(4):106, 2018. DOI: https://doi.org/10.3390/bios8040106.

[50] Yan Zhou, Cheng-Hui Liu, Yi Sun, Yang Pu, Susie Boydston-White, Yulong Liu, and Robert R Alfano. Human brain cancer studied by resonance raman spectroscopy. *Journal of Biomedical Optics*, 17(11):116021, 2012. DOI: https://doi.org/10.1117/1.JBO.17.11.116021.

[51] Michael Jermyn, Joannie Desroches, Jeanne Mercier, Karl St-Arnaud, Marie-Christine Guiot, Frederic Leblond, and Kevin Petrecca. Raman spectroscopy detects distant invasive brain cancer cells centimeters beyond mri capability in humans. *Biomedical Optics Express*, 7(12):5129–5137, 2016. DOI: https://doi.org/10.1364/BOE.7.005129.

[52] Nadia Amharref, Abdelilah Beljebbar, Sylvain Dukic, Lydie Venteo, Laurence Schneider, Michel Pluot, and Michel Manfait. Discriminating healthy from tumor and necrosis tissue in rat brain tissue samples by raman spectral imaging. *Biochimica et Biophysica Acta*, 1768(10):2605–2615, 2007. DOI: https://doi.org/10.1016/j.bbamem.2007.06.032.

[53] Yan Zhou, Cheng-Hui Liu, Binlin Wu, Xinguang Yu, Gangge Cheng, Ke Zhu, Kai Wang, Chunyuan Zhang, Mingyue Zhao, Rui Zong, et al. Optical biopsy identification and grading of gliomas using label-free visible resonance raman spectroscopy. *Journal of Biomedical Optics*, 24(9):1–12, 2019. DOI: https://doi.org/10.1117/1.jbo.24.9.095001.

[54] Karl Herholz, Karl-Josef Langen, Christiaan Schiepers, and James M Mountz. Brain tumors. In *Seminars in Nuclear Medicine*, volume 42, pages 356–370. Elsevier, 2012. DOI: https://doi.org/10.1053/j.semnuclmed.2012.06.001.

[55] Aaron C Tan, David M Ashley, Giselle Y López, Michael Malinzak, Henry S Friedman, and Mustafa Khasraw. Management of glioblastoma: State of the art and future directions. *CA: A Cancer Journal for Clinicians*, 70(4):299–312, 2020. DOI: https://doi.org/10.3322/caac.21613.

[56] Philip C De Witt Hamer, Martin Klein, Shawn L Hervey-Jumper, Jeffrey S Wefel, and Mitchel S Berger. Functional outcomes and health-related quality of life following glioma surgery. *Neurosurgery*, 88(4):720–732, 2021. DOI: https://doi.org/10.1093/neuros/nyaa365.

[57] Yoon Hwan Byun and Chul-Kee Park. Classification and diagnosis of adult glioma: a scoping review. *Brain & Neurorehabilitation*, 15(3):e23, 2022. DOI: https://doi.org/10.12786/bn.2022.15.e23.

[58] Mathew D Lin, Alexander C-Y Tsai, Kalil G Abdullah, Samuel K McBrayer, and Diana D Shi. Treatment of idh-mutant glioma in the indigo era. *NPJ Precision Oncology*, 8(1):149, 2024. DOI: https://doi.org/10.1038/s41698-024-00646-2.

[59] Lenny Dang, David W White, Stefan Gross, Bryson D Bennett, Mark A Bittinger, Edward M Driggers, Valeria R Fantin, Hyun Gyung Jang, Shengfang Jin, Marie C Keenan, et al. Cancer-associated idh1 mutations produce 2-hydroxyglutarate. *Nature*, 462(7274):739–744, 2009. DOI: https://doi.org/10.1038/nature08617.

[60] Maria E Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F Fernandez, et al. Leukemic idh1 and idh2 mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer Cell*, 18(6):553–567, 2010. DOI: https://doi.org/10.1016/j.ccr.2010.11.015.

[61] Patrick S Ward, Jay Patel, David R Wise, Omar Abdel-Wahab, Bryson D Bennett, Hilary A Coller, Justin R Cross, Valeria R Fantin, Cyrus V Hedvat, Alexander E Perl, et al. The common feature of leukemia-associated idh1 and idh2 mutations is a neomorphic enzyme activity converting α-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell*, 17(3):225–234, 2010. DOI: https://doi.org/10.1016/j.ccr.2010.01.020.

[62] Liang Xia, Bin Wu, Zhiquan Fu, Fang Feng, Enqi Qiao, Qinglin Li, Caixing Sun, and Minghua Ge. Prognostic role of idh mutations in gliomas: a meta-analysis of 55 observational studies. *Oncotarget*, 6(19):17354–17365, 2015. DOI: https://doi.org/10.18632/oncotarget.4008.

[63] Hao-Wen Sim, Romina Nejad, Wenjiang Zhang, Farshad Nassiri, Warren Mason, Kenneth D Aldape, Gelareh Zadeh, and Eric X Chen. Tissue 2-hydroxyglutarate as a biomarker for isocitrate dehydrogenase mutations in gliomas. *Clinical Cancer Research*, 25(11):3366–3373, 2019. DOI: https://doi.org/10.1158/1078-0432.CCR-18-3205.

[64] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005. DOI: https://doi.org/10.1056/nejmoa043330.

[65] Yalan Zhang, Calixto-Hope G Lucas, Jacob S Young, Ramin A Morshed, Lucie McCoy, Nancy Ann Oberheim Bush, Jennie W Taylor, Mariza Daras, Nicholas A Butowski, Javier E Villanueva-Meyer, et al. Prospective genomically guided identification of "early/evolving" and "undersampled" idh-wildtype glioblastoma leads to improved clinical outcomes. *Neuro Oncology*, 24(10):1749–1762, 2022. DOI: https://doi.org/10.1093/neuonc/noac089.

[66] Hanwen Bai, Akdes Serin Harmancı, E Zeynep Erson-Omay, Jie Li, Süleyman Coşkun, Matthias Simon, Boris Krischek, Koray Özduman, S Buelent Omay, Eric A Sorensen, et al. Integrated genomic characterization of idh1-mutant glioma malignant progression. *Nature Genetics*, 48:59–66, 2016. DOI: https://doi.org/10.1038/ng.3457.

[67] Hong Jiang, Samarth Hegde, Brett L Knolhoff, Yu Zhu, John M Herndon, Melissa A Meyer, Timothy M Nywening, William G Hawkins, Irina M Shapiro, David T Weaver, et al. Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy. *Nature Medicine*, 22:851–860, 2016. DOI: https://doi.org/10.1038/nm.4123.

[68] Martin J Van den Bent, Michele Reni, Gemma Gatta, and Charles Vecht. Oligodendroglioma. *Critical Reviews in Oncology/Hematology*, 66(3):262–272, 2008. DOI: https://doi.org/10.1016/j.critrevonc.2007.11.007.

[69] Caterina Giannini and Bernd W Scheithauer. Classification and grading of low-grade astrocytic tumors in children. *Brain Pathology*, 7:785–798, 1997. DOI: https://doi.org/10.1111/j.1750-3639.1997.tb01064.x.

[70] Peter C Burger, F Stephen Vogel, Sylvan B Green, and Thomas A Strike. Glioblastoma multiforme and anaplastic astrocytoma pathologic criteria and prognostic implications. *Cancer*, 56(5):1106–1111, 1985. DOI: https://doi.org/10.1002/1097-0142(19850901)56:5%3C1106::aid-cncr2820560525%3E3.0.co;2-2.

[71] Peter A Forsyth, Edward G Shaw, Bernd W Scheithauer, Judith R O'Fallon, Donald D Layton Jr, and Jerry A Katzmann. Supratentorial pilocytic astrocytomas. a clinicopathologic, prognostic, and flow cytometric study of 51 patients. *Cancer*, 72(4):1335–1342, 1993. DOI: https://doi.org/10.1002/1097-0142(19930815)72:4%3C1335::aid-cncr2820720431%3E3.0.co;2-e.

[72] Donald M Ho, Tai-Tong Wong, Chih-Yi Hsu, Ling-Tan Ting, and Hung Chiang. Mib-1 labeling index in nonpilocytic astrocytoma of childhood: a study of 101 cases. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 82(12):2459–2466, 1998. DOI: https://doi.org/10.1002/(sici)1097-0142(19980615)82:12%3C2459::aid-cncr21%3E3.0.co;2-n.

[73] Isabelle Camby, Nathalie Nagy, Maria-Beatriz Lopes, Beat W Schäfer, Claude-Alain Maurage, Marie-Magdeleine Ruchoux, Petra Murmann, Roland Pochet, Claus W Heizmann, Jacques Brotchi, et al. Supratentorial pilocytic astrocytomas, astrocytomas, anaplastic astrocytomas and glioblastomas are characterized by a differential expression of s100 proteins. *Brain Pathology*, 9:1–19, 1999. DOI: https://doi.org/10.1111/j.1750-3639.1999.tb00205.x.

[74] Hans-Georg Wirsching and Michael Weller. Glioblastoma. *Malignant brain tumors: state-of-the-art treatment*, pages 265–288, 2017. DOI: https://doi.org/10.1007/978-3-319-49864-5_18.

[75] HD Nguyen, A Allaire, P Diamandis, M Bisaillon, MS Scott, and M Richer. A machine learning analysis of a "normal-like" idh-wt diffuse glioma transcriptomic subgroup associated with prolonged survival reveals novel immune and neurotransmitter-related actionable targets. *BMC Medicine*, 18:280, 2020. DOI: https://doi.org/10.1186/s12916-020-01748-x.

[76] Augusto Maury and Reynier I. Revilla. Autocorrelation analysis combined with a wavelet transform method to detect and remove cosmic rays in a single raman spectrum. *Applied Spectroscopy*, 69(8):984–992, 2015. DOI: https://doi.org/10.1366/14-07834.

[77] Filip Peška. Raman microspectroscopy data processing. Bachelor's Thesis, Department of Theoretical Computer Science and Mathematical Logic, 2022. Available at: https://dspace.cuni.cz/bitstream/handle/20.500.11956/174211/130333450.pdf?sequence=1.

[78] David C. Hoaglin. Volume 16: How to detect and handle outliers. In *Proceedings of the American Statistical Association*, 2013. Available at: https://api.semanticscholar.org/CorpusID:208231456.

[79] Darren A Whitaker and Kevin Hayes. A simple algorithm for despiking raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 179:82–84, 2018. DOI: https://doi.org/10.1016/j.chemolab.2018.06.009.

[80] Jin Chen, Per. Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter. *Remote Sensing of Environment*, 91(3-4):332–344, 2004. DOI: https://doi.org/10.1016/j.rse.2004.03.014.

[81] Y. Liu, B. Dang, Y. Li, et al. Applications of savitzky-golay filter for seismic random noise reduction. *Acta Geophysica*, 64:101–124, 2016. DOI: https://doi.org/10.1515/acgeo-2015-0062.

[82] Kristian Hovde Liland, Elling-Olav Rukke, Elisabeth Fjærvoll Olsen, and Tomas Isaksson. Customized baseline correction. *Chemometrics and Intelligent Laboratory Systems*, 109(1):51–56, 2011. DOI: https://doi.org/10.1016/j.chemolab.2011.07.005.

[83] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010. DOI: https://doi.org/10.1039/b922045c.

[84] Carlos Cobas. Applications of the whittaker smoother in nmr spectroscopy. *Magnetic Resonance in Chemistry*, 56(12):1140–1148, 2018. DOI: https://doi.org/10.1002/mrc.4747.

[85] Peshawa Jamal Muhammad Ali, Rezhna Hassan Faraj, Erbil Koya, Peshawa J Muhammad Ali, and Rezhna H Faraj. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1):1–6, 2014. DOI: https://doi.org/10.13140/RG.2.2.28948.04489.

[86] Nils Kristian Afseth, Vegard Herman Segtnan, and Jens Petter Wold. Raman spectra of biological samples: A study of preprocessing methods. *Applied Spectroscopy*, 60(12):1358–1367, 2006. DOI: https://doi.org/10.1366/000370206779321454.

[87] GeeksforGeeks, Sanchhaya Education Private Limited. Min-max normalization. https://www.geeksforgeeks.org/machine-learning/data-normalization-in-data-mining/. Accessed: 2025-11-30.

[88] Saichon Sinsomboonthong. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *International Journal of Mathematics and Mathematical Sciences*, 2022(1):1–9, 2022. DOI: https://doi.org/10.1155/2022/3584406.

[89] GeeksforGeeks, Sanchhaya Education Private Limited. L2 normalization. https://www.geeksforgeeks.org/machine-learning/l1-l2-norms-in-sparse-modeling/. Accessed: 2025-11-30.

[90] Lior Rokach and Oded Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005. Available at: https://link.springer.com/book/10.1007/b107408.

[91] Noam Slonim, Ehud Aharoni, and Koby Crammer. Hartigan's k-means vs. lloyd's k means–is it time for a change? In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1677–1684. AAAI Press, 2013. Available at: https://www.ijcai.org/Proceedings/13/Papers/249.pdf#:~:text=Hartigan%E2%80%99s%20method%20for%20k-means%20clustering%20holds%20several%20potential,is%20a%20subset%20of%20those%20of%20Lloyd%E2%80%99s%20method.

[92] Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023. DOI: https://doi.org/10.1016/j.ins.2022.11.139.

[93] Richard Mojena. Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal*, 20(4):359–363, 1977. DOI: https://doi.org/10.1093/comjnl/20.4.359.

[94] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *WIREs DataMining and Knowledge Discovery*, 2(1):86–97, 2011. DOI: https://doi.org/10.1002/widm.53.

[95] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. DOI: https://doi.org/10.1080/01621459.1963.10500845.

[96] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM SIGMOD Record*, 25(2):103–114, 1996. DOI: https://doi.org/10.1145/235968.233324.

[97] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997. DOI: https://doi.org/10.1023/A:1009783824328.

[98] Elly Muningsih, Chandra Kesuma, Sunanto, Suripah, and Aprih Widayanto. Combination of k-means method with davies bouldin index and decision tree method with parameter optimization for best performance. In *2nd International Conference on Advanced Information Scientific Development (ICAISD) 2021*, volume 2714, pages 020021–1–020021–7. AIP Conference Proceedings, 2023. DOI: https://doi.org/10.1063/5.0129119.

[99] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227, 1979. DOI: https://doi.org/10.1109/TPAMI.1979.4766909.

[100] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. DOI: https://doi.org/10.1080/03610927408827101.

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

[102] Dongju Liu and Jian Yu. Otsu method and k-means. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, volume 1, pages 344–349. IEEE, 2009. DOI: https://doi.org/10.1109/HIS.2009.74.

[103] Jean Serra and Luc Vincent. An overview of morphological filtering. *Circuits, Systems and Signal Processing*, 11(1):47–108, 1992. DOI: https://doi.org/10.1007/BF01189221.

[104] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada*, volume 1, pages 278–282. IEEE, 1995. DOI: https://doi.org/10.1109/ICDAR.1995.598994.

[105] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

[106] Helen Pearson, Heidi Ledford, Matthew Hutson, and Richard Van Noorden. The most-cited papers of the twenty-first century. *Nature*, 640:588–592, 2025. Available at: https://www.nature.com/articles/d41586-025-01125-9.

[107] GeeksforGeeks, Sanchhaya Education Private Limited. Random forest. https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/. Accessed: 2025-11-30.

[108] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. DOI: https://doi.org/10.48550/arXiv.1603.02754.

[109] Soukaina Hakkal and Ayoub Ait Lahcen. Xgboost to enhance learner performance prediction. *Computers and Education: Artificial Intelligence*, 7:100254, 2024. DOI: https://doi.org/10.1016/j.caeai.2024.100254.

[110] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, (5):1189–1232, 2001. DOI: https://doi.org/10.1214/aos/1013203451.

[111] Robin Schmucker, Jingbo Wang, Shijia Hu, and Tom M Mitchell. Assessing the performance of online students–new data, new approaches, improved accuracy. *Journal of Educational Data Mining*, (1), 2022. DOI: https://doi.org/10.5281/zenodo.6450190.

[112] GeeksforGeeks, Sanchhaya Education Private Limited. Xgboost. https://www.geeksforgeeks.org/machine-learning/xgbclassifier/. Accessed: 2025-11-30.

[113] Shan Suthaharan. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, pages 207–235. Springer, 2016. DOI: https://doi.org/10.1007/978-1-4899-7641-3_9.

[114] Michelle Dunbar, John M Murray, Lucette A Cysique, Bruce J Brew, and Vaithilingam Jeyakumar. Simultaneous classification and feature selection via convex quadratic programming with application to hiv-associated neurocognitive disorder assessment. *European Journal of Operational Research*, 206(2):470–478, 2010. DOI: https://doi.org/10.1016/j.ejor.2010.03.017.

[115] V Jeyakumar, G Li, and S Suthaharan. Support vector machine classifiers with uncertain knowledge sets via robust optimization. *Optimization*, 63:1099–1116, 2014. DOI: https://doi.org/10.1080/02331934.2012.703667.

[116] Ingo Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003. Available at: https://dl.acm.org/doi/10.5555/945365.964289.

[117] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*, volume 25, pages 476–484, 2012. 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012 ; Conference date: 03-12-2012 Through 06-12-2012.

[118] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. DOI: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[119] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004. Available at: https://jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf.

[120] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248. IEEE, 2020. DOI: https://doi.org/10.1109/ICICS49469.2020.239556.

[121] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal*

of Artificial Intelligence Research, 16:321–357, 2002. DOI: https://doi.org/10.1613/jair.953.

[122] TURING. Oversampling. https://www.turing.com/kb/how-data-collection-and-data-preprocessing-in-python-help-in-machine-learning. Accessed: 2025-11-30.

7 Appendices

Appendix A: Use of AI in the Thesis

Figure 2 was generated using Google Gemini to show the study design from data collection to analyze the RS data using ML. Grammar was checked and modified using Grammarly and ChatGPT.