Analytical and Bioanalytical Chemistry manuscript No.
(will be inserted by the editor)

Raman Spectroscopic Grading of Astrocytoma Tissues: Using Soft Reference Information

Claudia Beleites · Kathrin Geiger · Matthias Kirsch · Stephan B. Sobottka · Gabriele Schackert · Reiner Salzer

Received: date / Accepted: date

Abstract Gliomas are the most frequent primary brain tumours. During neurosurgical treatment, locating the exact tumour border is often difficult. This study assesses grading of astrocytomas based on Raman spectroscopy for a future application in intra-surgical guidance. Our predictive classification models distinguish the surgically relevant classes "normal tissue", and "low" and "high grade astrocytoma" in Raman maps of moist bulk samples (80 patients) acquired with a fibre-optic probe.

We introduce partial class memberships as a strategy to utilize borderline cases for classification. Borderline cases supply most valuable training and test data for our application. They are (a) examples of the sought boundary and (b) the cases for which new diagnostics are needed. Besides, the number of suitable training samples increases considerably: *Soft* logistic regression (LR) utilises 85 % more spectra and 50 % additional patients than LDA. The predictive soft LR models achieve ca. 85, 67, and 84 % (normal, low, and high grade) sensitivity and specificity. We discuss the different heuristics of LR and LDA in the light of borderline samples.

While we focus on prediction, the spectroscopic interpretation of the predictive models agrees with previous descriptive studies. Unsaturated lipids are used to differentiate between normal and tumour tissues, while the total lipid content prominently contributes to the determination of the tumour grade. The high wavenumber region above 2800 cm^{-1} alone did not allow successful grading.

We give a proof of concept for Raman spectroscopic grading of moist astrocytoma tissues and propose to include borderline samples into classifier training and testing.

Keywords Gliomas · Astrocytomas · Grading · Classification · Tumour · Raman Spectroscopy · Linear Discriminant Analysis · Logistic Regression · Soft Classification

1 Introduction

1.1 Gliomas

Gliomas are the most frequent primary brain tumours. Among them, astrocytomas are the largest subgroup. The world health organisation distinguishes four grades of astrocytomas according to their histology and behaviour [1–3]. In adults, only °II to IV are found. Some astrocytomas °II (A °II) do not appear to be malignant, but most de-differentiate in time and gain in malignancy. Astrocytomas °III (A °III) are malignant, and glioblastomas (°IV; GBM) are the most undifferentiated gliomas. While some A °III and GBM are formed by de-differentiation of lower grade tumours, others appear *de novo* [1, 4].

Glioma treatment comprises surgery if possible, and the complete removal of the tumour is one of the most important factors for the prediction of the recurrence-free survival time of the patient [4, 5]. In tumour surgery outside the brain, ample safety margins around the tumour are often applied.

C. Beleites (✉)
CENMAT, Materials and Natural Resources Dept.
University of Trieste
Via Alfonso Valerio 6/a
34127 Trieste/Italy
E-mail: cbeleites@units.it

C. Beleites · R. Salzer
Analytical Chemistry,
Dresden University of Technology,
Dresden/Germany

K. Geiger · M. Kirsch · S. B. Sobottka · G. Schackert
University Hospital Carl-Gustav Carus,
Dresden University of Technology,
Dresden/Germany

This is not possible in brain surgery as the normal brain tissue *must* be preserved. An additional difficulty arises from the infiltrative growth of the astrocytomas: the tumour border is hardly visible. Thus, although complete removal of the tumour is asked for, the surgical decision is often to remove the malignant part of the tumour only. Neurosurgeons work with a precision up to 1 mm but so far no diagnostic technique delineating the proper excision border for the astrocytomas is available at this resolution.

The pre-operative diagnostic involves standard imaging techniques such as magnet resonance tomography (MRT) and computed tomography (CT) as well as less standard methods like ^{18}F or ^{11}C positron emission tomography (PET), or single photon emission computed tomography (SPECT). During surgery, however, the opening of the skull, incision, and dislocation by the surgical tools as well as swelling of the cut tissue and the movement due to the heart beat lead to substancial displacement of the tissues. This limits the spatial resolution of stereo-navigation based on pre-surgery images to several millimeters, in particular situations even to 1 cm.

During surgery, imaging techniques such as angiography, intra-operative MRT, or fluorescence guidance by 5-aminolevulinic acid may be used. Intra-operative MRT requires not only extremely costly instrumentation but also considerable amounts of time (at least 1 h) and, while less affected by brain shift than pre-surgery MRT, frequently the image quality is considerably lower. Fluorescence guided surgery is recommended for malignant tumours [6], but like MRT and CT with contrast agents, the enhancement is restricted to areas where the blood brain barrier is compromised [6–8]. Angiography is also useful for the diagnosis of malignant tumours, but not even all A °III show enhancement [6]. The latter two techniques have thus inherent difficulties delineating the border between low grade and normal tissue. Histopathological diagnosis yields details about the tumour biology and grade down to cell level. For the finding of the tumour border, however, it is of limited aid as it is a purely *ex-vivo* technique and even with rapid staining protocols at least 20 min are needed to arrive at a diagnosis.

In contrast to histopathologic (*ex-vivo*) questions, the total time of a the surgical procedure is critical, as longer narcosis imposes considerable physical strain on the patient. Therefore, not only the time to arrive at a diagnosis for a piece of tissue but also the time to treat this tissue is critical, and surgical treatment of single cells is not feasible. The surgeon should be presented with information at the desired level of (spatial) detail which may even even be lower than the maximal working precision.

In other words, tools to help surgeons finding the proper excision border *in-vivo* with a spatial resolution adapted to the surgeons' working precision of up to 300 µm are badly needed.

1.2 Raman Spectroscopy and Grading of Astrocytomas

Raman spectroscopy has shown its potential to identify tumours (see e.g. the reviews [9–11]). It has been applied to differentiate tissues in human GBM [12–14]. In animal models, GBM and C6 gliomas were studied *ex-vivo* using thin sections [15] and moist bulk samples [16, 17], and recently first *in-vivo* measurements were conducted [17, 18]. All these studies, however, focus on descriptive models of particular subsets of the tissues encountered in gliomas. In contrast, we present here dedicated predictive models with larger patient numbers and including patients of all relevant tumour grades: A °II, A °III, and GBM. Predictive as well as descriptive grading of astrocytomas has been studied by mid-infrared spectroscopy [8, 19, 20] including A °II, A °III, and GBM. The first two [19, 20] model the patient's tumour grade (maximal de-differentiated morphology), the third [8] models the predominant morphology, too.

For intra-operative *in-vivo* diagnostics, Raman spectroscopy offers significant advantages over mid-infrared spectroscopy. Water does not disturb the analysis of Raman spectra. Thus, native tissue is easily analysed. Fibre optic probes give spatial flexibility, and the spatial resolution can be chosen according to the needs of the surgeon. In general, larger focus diameters allow higher total excitation power without damaging the tissue and thus shorter acquistion times per spectrum. Therefore, the spatial resolution of the probe should match the surgeon's needs also in order to allow speedy collection of the spectra. Raman probes may be configured to measure deeper into the tissue than mid-infrared techniques which can access the first few µm only. Last but not least, fibre optic Raman probes allow non-destructive non-invasive (in the sense that the tissue is already exposed during surgery, and it is not compromised by the measurements) collection of the spectra: in contrast to *ex-vivo* techniques like histology or infrared spectroscopy, there is no need to remove the brain tissue for Raman based diagnosis.

Utzinger and Richards-Kortum [21] discuss fibre optic probes for biomedical applications, and Santos et al [22] studied the background signal of commercially available optical fibres. Optical fibres conducting light emit a Raman spectrum due to their excitation by the conducted light. However, the silica signals appear only in the fingerprint region of the Raman spectrum. If the spectral region above 2000 cm^{-1} Raman shift suffices for the data analysis, unfiltered probes with combined excitation and detection fibre can be used. This considerably eases miniaturisation. A proof-of-concept study distinguished vital GBM from necrosis in the high wavenumber region of the Raman spectra [13].

2 Hard and Soft Classification

2.1 Hard and Soft Classification for Tumour Grading

Like qualitative analysis, classification addresses alternative questions. Each sample is assigned to one of a set of preknown categories, the classes. The classes are, *e. g.*, presence or absence of certain analytes, or of a disease. Classification is a supervised technique: the chemometric models are built (trained) using spectra together with a reference information stating to which class each spectrum belongs. Such information connecting sample and class is called label or membership. The model can then classify (predict) new samples. So

far, the class labels must be hard (crisp): each spectrum must belong exactly and completely to one class.

In tumour grading, several issues are not described appropriately by hard labels:

1. Astrocytomas may de-differentiate, *i. e.* areas with cells currently undergoing the progression from one WHO grade to the next may be encountered.
2. The tumours frequently are spatially heterogeneous: one tumour can contain cells of varying differentiation. As astrocytomas grow infiltratively, transition zones of gradually changing amounts of cells result.
3. Like measurements, diagnoses are subject to random and systematic uncertainty, and histology and/or different methods' diagnoses may disagree (see *e. g.* [23]).
4. Sometimes, like in our study, the diagnosis is given for parallel cryo-sections while the spectra are taken of moist bulk tissue. Further uncertainty arises with regard to the exact location of the different tissues in the bulk sample.

The first of these problems is inherent to tumour grading. Gene and protein expression of morphologically similar astrocytoma tissues can vary depending on the patient's tumour grade. The biochemical changes during de-differentiation of astrocytomas are quite continuous [1, 24]. In addition, any therapy the patient has received previously may induce selection ($\frac{1}{4}$ of our samples are from recurrent tumours). Raman spectroscopy probes the biochemical composition, while the histological grading uses morphology. The borders due to morphological changes thus may not coincide with the most prominent changes in the spectra.

Considering the second problem, it may be possible to arrive at a diagnosis at cell level. However, single cell treatment is not feasible in open surgery. *In-vivo* tools for intra-surgical guidance should deliver the spatial resolution requested by the surgeon. This is particularly important for our application as higher spatial resolution implies disproportionally longer measurement time.

Regarding the (dis)agreement among different histologists, it should be noted that the agreement in astrocytoma grading among neuropathologists is much higher than among surgical pathologists [23]. Wrensch et al [25] reached consensus diagnosis in $\frac{886}{900} = 98.5\,\%$ of the patients. In constrast to grading a patient's tumour, our aim is the distinction of tissues *within* the tumour. This implies a fundamentally different "detailed" concept of reference diagnosis which is explained in detail below (section 4.1).

Transferring the neuropathologist's findings onto the measurement grid can raise substantial difficulties (see the experimental section).

Samples for which a hard reference label cannot be obtained are frequently excluded from the classification training data. This is a rather unfortunate decision:

- In biomedical spectroscopy samples are rare, but the quality of classification models depends crucially on the number of samples (patients) per variate (data points per spectrum) [26–28]. Many spectroscopic studies aiming at medical diagnosis comprise patient numbers 10 to 100 times below the recommended 5 samples per class and variate [28–31].

  As samples are so scarce, *every* sample should be used.
- More importantly, spectra of the transition zone are actually examples of the sought decision border, and are thus most useful training samples.
- Diagnostic tools are preferentially used for difficult cases. Borderline cases should therefore be included into the training (or at least test) data as early as possible.
- Excluding borderline samples from classifier training creates a serious risk of overestimating the class separability. Moreover, this can only be detected by borderline cases in the test set. Such test data is suitable for training as well.

Soft classification denotes two different and independent generalization strategies of traditional, hard classification. Firstly, one-class classifiers (like soft independent modelling of class analogies, SIMCA) allow that samples belong to more than one class each. Secondly, samples may belong to each class only partially: partial memberships state a degree of belonging, and take values between 0 and 1 (100 %). Here, we use partial memberships while requiring all memberships to sum to 1. One-class classfication is not meaningful in our application as the classes are mutually exclusive.

Throughout the paper, we refer to samples or spectra belonging completely to one class as *crisp*, and to samples (or spectra) with partial memberships in more classes as *soft*. Likewise, classifiers built using soft (and crisp) samples are soft classifiers.

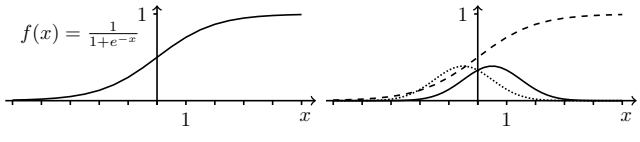
Many classifiers produce partial memberships as primary output. Fewer methods allow partial membership in the reference labels. This explains why soft labelled data is usually excluded from training.

Partial class memberships are interpreted in two ways. Firstly, the sample can be a mixture of the underlying classes, *e. g.* cells undergoing de-differentiation or mixtures of different cell types. Thus the qualitative analysis becomes quantitative: the analysis of a mixture. Non-chemical disciplines frequently discuss this in the light of fuzzy set theory (*e. g.* remote sensing [32]). Secondly, soft memberships may express probabilities or degrees of (un)certainty, *e. g.* uncertainty of histological diagnosis.

From a chemometric point of view, the mixture interpretation leads to calibration techniques such as PLS. In fact, partial least squares regression with a threshold (PLS-DA) has been used for hard classification in the context of biomedical Raman spectroscopy (*e. g.* [33, 34]). The probability interpretation calls for regression techniques that model a probability, *e. g.* logistic regression.

2.2 Logistic Regression (LR).

Logistic regression linearly models the log odds of class memberships [26, 35]. *Hard* LR is a standard method particularly in the medical and social sciences, and has been applied for hard vibrational spectroscopic classification with binary [36], multinomial [37, 38], and hierarchical multiclass setups [39–42]. To our knowledge, however, the LR's capability to handle *soft* reference labels has not yet been used for vibrational spectroscopic classification.



(a) logistic regression (b) linear discriminant analysis

Fig. 1: (a) The logistic function models posterior probabilities. (b) LDA with $m = \pm 0.5$ and $s = 1$ (solid and dotted) yields the logistic function as posterior probability (dashed), too.

Link to linear discriminant analysis (LDA) . LDA models classes as multivariate normal distributions with a common covariance matrix [10, 26, 35].

Formally, LR and LDA can be shown to be closely linked: the log odds of the LDA's posterior probability have the same form as the LR model. LR, however, does not assume any particular distribution of the samples. As usual for parametric models, LDA is more powerful if its assumptions are met. But it reacts sensitively to distant samples [26].

LR concentrates on the boundaries *between* the classes, while LDA describes the class boundaries towards the outside of LD space as well. In other words, the LR in figure 1a will hardly be influenced by a new sample at $x = 5$ belonging to the class with positive x as the probability is already very close to 1. The LDA in figure 1b will change much more as the new sample is outside the distribution of the class (black).

Practical considerations. LR classifiers should not be fitted by least squares as the residuals are not normally distributed [26, 35]. Fitting software for artificial neural networks (ANN) conveniently avoids this issue, and can be used as the LR model is equivalent to an ANN without hidden layer using the logistic function as sigmoid. Such LR models can easily be extended into fully-featured ANN.

2.3 Validation of Soft Classification Models

Validation measures the performance of a chemometric model. General guidelines can be found elsewhere [27, 43, 44].

For medical diagnostic tests, *sensitivity* and *specificity* are widely used performance measures [45]. The sensitivity is the number of correctly recognised samples (spectra) of a class divided by the number of samples truly belonging to the class. It answers the question how well the class is recognised. In contrast, specificity asks how well the model recognises that a sample does *not* belong to the respective class. It is calculated as the fraction of samples correctly not assigned to the class among all samples that truly do not belong to the class.

These two performance measures can be depicted in the *specificity-sensitivity diagram* (*e. g.* fig. 7), a flipped receiver operating curve. Sensitivity and specificity are defined solely for crisp data, *i. e.* data with crisp reference *and* crisp prediction. Soft predictions are "hardened" into crisp predictions by threshold values. Varying the threshold trades off between sensitivity and specificity, yielding a curve in the specificity-sensitivity diagram that characterises the overall performance of the models.

The concepts of sensitivity and specificity can be extended to soft reference data, but this is beyond the scope of this paper. Instead, we use performance measures appropriate for regression models, the mean absolute error MAE and the root mean squared error RMSE.

$$MAE_j = \frac{1}{n}\sum_{i=1}^{n} |\hat{p}_{i,j} - p_{i,j}|$$

$$RMSE_j = \sqrt{\frac{1}{n}\sum_{j=1}^{g} (\hat{p}_{i,j} - p_{i,j})^2}$$

with the respective class j, the number of spectra n, the number of classes g, the reference labels p, and the model's predictions $\hat{p}$. Compared to MAE, RMSE emphasises larger deviations from the reference, whereas small deviations are downweighted. The comparison of MAE and RMSE shows whether the predictions have small deviations for many samples (low RMSE compared to MAE), or whether fewer spectra are grossly misclassified (high RMSE).

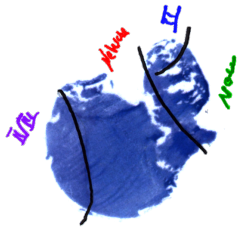
For models that allow soft prediction, MAE and RMSE are much more sensitive to deviations between the test data's label and the prediction than crisp measures like sensitivity and specificity. Hardending blurs slight deviations, the information of the soft prediction is partially lost [46]. Consider a classifier test where always 60 % posterior probability for the correct class are predicted. For threshold values between 40 and 60 %, the model reaches 100 % sensitivity and 100 % specificity, *i. e.* perfect performance. Nevertheless, the MAE will not be zero but 40 %. This accounts for thresholds outside the range 40 to 60 % which cause misclassification. Predicting $\hat{p} = 60\,\%$ instead of 100 % indicates a substantial risk that samples exist that are misclassified even though no such sample was encountered in the test set.

3 Experimental Details

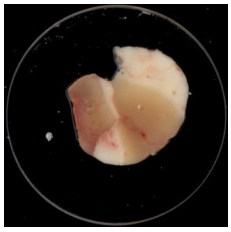
3.1 Samples

The specimen were snap frozen in liquid nitrogen immediately after excision and stored at -80 °C until preparation (approved by the human ethics committee of the Dresden University of Technology). The data comprises nine control samples, *i. e.* samples obtained from patients without any history of brain tumours. Three of these are among the oldest samples of the collection and were seven years old at the time of preparation. All other samples had been stored for less than two years. We included only cerebral tumour samples of patients where no oligodendroglial tumour compound was documented.

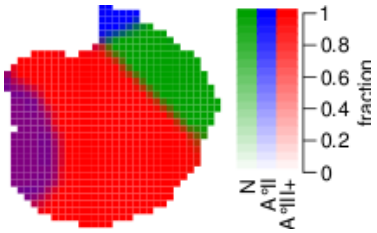
For reference diagnosis, 7 µm cryo sections were cut, fixed (50 vol% EtOH, 4 % formalin) and stained with methylene blue. Detailed diagnoses (fig. 2a) were obtained for each section. The remaining bulk of the specimen was kept frozen until Raman measurement. It was placed on a CaF_2 window with the face adjoining the cryo section (fig. 2b). The CaF_2 serves as lid of a moist chamber that prevents drying of the sample. With the sample hanging from the CaF_2



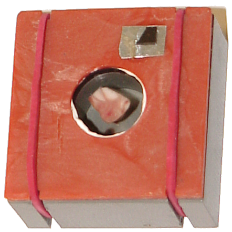
(a) Methylene blue stained section with detailed reference diagnosis



(b) Bulk sample prior to Raman measurement



(c) False colour map indicating reference labels for the sample. The area at the transition between °II and III (left) is labelled 50 % with each grade and thus appears violet. Each pixel corresponds to one spectrum.



(d) Moist chamber with sample. A corner of the Si piece on the Al foil serves as stage coordinate origin.

Fig. 2: Sample preparation. Sub-figures (a) to (c) show the same GBM sample. The heterogeneity is typical for the astrocytomas. The CaF_2 window has a diameter of 12 mm.

window (fig. 2d) the probe's focus can be kept constant during spectra collection. In addition, sample deformation is kept minimal which is crucial for transferring the reference diagnosis onto the spectra. Immersion measurements were not feasible as particularly GBM samples often were all but cell suspensions.

3.2 Spectra Acquisition

Raman spectra were acquired with a f/1.8 spectrograph (Kaiser Optical Systems, Ann Arbor, USA) using a filtered fibre-optic probe (Raman Probe, working distance 5 mm, NA 0.4, 50 and 100 µm excitation and collection fibres; Inphotonics, USA). The focus diameter is ca. 60 µm. The excitation laser (785 nm, multimode; Toptica Photonics Inc., USA) delivers approximately 70 mW below the CaF_2 window. The measurements in this initial study were restricted to 6 h per sample to prevent degradation. The step size of the measurement grid was chosen accordingly and varied between 200 and 333 µm. The excitation time of 20 s per spectrum includes a factor of two for the "Cosmic Ray Filter". Seven measurements were taken without cosmic ray filter (*i. e.*, 10 s exposure). The excitation fibre diameter of the probe was not matched to the laser. A more recently acquired probe (matched, 100 µm diameter) yields the same signal to noise ratio already in about 5 s.

For the collection of the Raman spectra, the sample outline was recorded with the motorized stage (PRIOR Proscan II; Prior, USA) under the spectrometer's microscope (Leica, Germany). A square grid was set up and the points within the outline polygon (plus a "safety margin" of 2 or 3 spectra) were calculated. The stage was then moved to position

Table 1: Overview of the data set.

class	crisp reference patients	crisp reference spectra	crisp + soft reference patients	crisp + soft reference spectra
Normal	16	7 456	35	15 747
thereof controls	9	4 902	9	4 902
Astrocytoma °II	17	4 171	47	19 128
Astrocytoma °III+	27	8 279	53	21 617
total	53	19 906	80	37 015

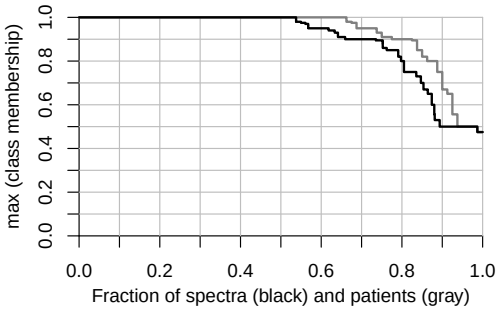


Fig. 3: Only 55 % the spectra could be labelled crisply, and more than a third of the patients do not have any crisp spectrum.

the moist chamber under the fibre-optic for the automated measurement of the defined points. The bottom right corner of a little Si piece served as a coordinate origin (fig. 2d) that is easily found manually in the visible microscope image and automatically with a 2d divide and conquer search using the 520 cm^{-1} band of the Si corner under the probe.

3.3 Spectra Pre-processing and Data Analysis

3.3.1 Reference labels.

The detailed diagnoses given for the parallel sections (fig. 2a) were transferred onto the measurement grid (fig. 2c). The transfer of the diagnosis onto the measurement grid was carried out without any display of the spectra using exclusively the visible images. Tumour tissue diagnosed as borderline case (violet area in fig. 2c A °II/III) was labelled belonging half and half to the respective classes (violet = half red, half blue). The diagnosis "individual tumour cells in normal tissue" and tissue where the histologist was not sure whether it is tumour were labelled as 5 % tumour and 95 % normal.

Some samples were completely round so that the rotation between measurement grid and diagnosed parallel section could not be determined with certainty. Also, deformation of the sample during thawing sometimes cause the exact location of the diagnosis on the measured surface to be ambiguous. Areas affected by such "deformation uncertainty" were labelled with the fractions of the areas occupied by the different classes on the reference section.

Tab. 1 and fig. 3 give an overview of the labelled data.

3.3.2 Data analysis.

All further data processing used the statistical environment R [47], including ggplot2 [48] for graphical display.

Pre-processing. The raw spectra were imported into the spectra handling package hyperSpec [49] using R.matlab [50]. The spectral ranges below 755 cm^{-1} and between 1850 and 2500 cm^{-1} were discarded. Co-additions were multiplicatively signal corrected (package pls [51]) and averaged. The spectra were corrected for the camera's dark current and intensity calibrated. Baseline correction (755 to 1850 cm^{-1} quadratic, 2625 – 3100 cm^{-1} linear; automatically fit) and cutting to 755 – 1800 and 2800 – 3025 cm^{-1} followed.

The spectrometer measures 2680 data points between 125 and 3556 cm^{-1} Raman shift with 4 cm^{-1} resolution and data point spacing between 0.87 cm^{-1} and 1.85 cm^{-1}. A smoothing interpolation (*spc.loess* [49]) created an evenly spaced Raman shift axis with data points every 5 cm^{-1}. This lowers the number of variates in order to stabilise the models. In addition, lower spectral resolution allows wider spectrometer slit and thus faster measurement in the clinical application.

Too intense spectra, spectra from outside the sample, and a few spectra that were contaminated due to accidentally switched on flourescent lamps were removed.

The spectra were normalised on the mean intensity between 2900 and 3025 cm^{-1} to approximately correct overall intensity changes that are not due to the spectral properties of the measured tissue (*e. g.* slightly changing focus). All spectra have good signal to noise ratio, well defined baseline, and also a similar shape in this spectral region: the mean signal to noise ratio is 29, whereas between 755 and 1800 cm^{-1} the average is 6 only. In addition, the spectral region below 1800 cm^{-1} contains residual baseline (most obvious below 900 cm^{-1} and above 1720 cm^{-1}) which would largely affect the normalization. The ratio of the area under the spectrum between this spectral region and the region in between (900 – 1720 cm^{-1}) ranges from 5 to 90 %. Normalization cancels the information of one variate. Consequently, the last data point of each spectrum was excluded.

Finally, we centred the data by subtracting the average spectrum of normal grey matter as a well defined reference that is independent of the fractions of the different tissues in the data set. Fig. 4 shows the pre-processed spectra before centring. The spectral signature of proteins is largely cancelled by subtraction of the average grey matter spectrum (after normalization): the amide I and phenylalanin bands at 1660 and 1005 cm^{-1} are clearly visible before centring (fig. 4), but can hardly be identified afterwards (fig. S.3) though the high grade tumour spectra show a residual feature at 1005 cm^{-1}. Thus, the normalization seems to normalize roughly to the protein content.

The models presented here summarize the spectra by linear combinations. Such models can implicitly do certain baseline corrections and normalization. E. g., a coefficient pattern of $-\frac{1}{2}I_1 + I_2 - \frac{1}{2}I_3$ probes a signal at I_2 corrected by a linear baseline through I_1 and I_3. Predictive quality depends less on the preprocessing than the spectroscopic interpretation of descriptive models. Yet they will profit of external spectroscopic knowledge that enters the data via preprocessing. All preprocessing was decided exclusively using spectroscopic knowledge, and no data-driven preprocessing was performed. Classifier optimization relies on model comparison – which is difficult for predictive classifiers [52]. Solutions are not yet known for "hierarchical" data structures like the present containing multiple patients and large and varying numbers of spectra per patient.

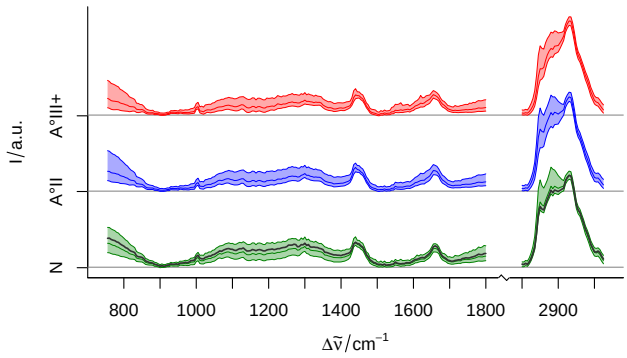


Fig. 4: Weighted median, 16th, and 84th percentile spectra. Thick line: the mean grey tissue spectrum used for centring.

Classification models. The samples are distinguished into normal brain tissues (white or grey matter, and leptomeninges; class "N"), A °II (low-grade morphology; class "A °II"), and high grade astrocytomas (A °III, GBM, and necroses; class "A °III+"). A small amount (0.5 %) of gliotic tissue is comprised in class N. While gliosis is not strictly normal, we retain the name for convenience. The classes reproduce the surgically relevant groups of tissue, and may be thought of as "must be preserved", "remove if possible", and "must be removed". These classes are heterogeneous in terms of tissues, and also in terms of the corresponding spectra. Normal white and grey matter have very different spectra (see *e. g.* [14, 53]), as particularly the different lipid contents and compositions change the Raman spectra. Also for A °III+ tissues differences have been described (*e. g.* [12, 13]).

In order to determine the proper class set up, two effects need to be traded off. On the one hand, heterogeneous classes are often difficult to distinguish as the within-class-variation is high: there should not be too few classes. Otherwise, the model will misclassify certain groups of samples. On the other hand, more classes also come with a penalty. Additional classes multiply the degrees of freedom of the model, requiring also a multiple of training samples in order to keep the models stable. A model with too many classes spends information on determining surgically irrelevant class boundaries. This causes a loss in classification performance also on the relevant class boundaries.

As our data set consists of about an order of magnitude less patients than recommended, the main concern is stability. Heterogeneous classes do not imply problems in the classification. The classification methods used in this paper first (linearly) project the spectra into a classifier space where the classes are as separate (and also as homogeneous) as possible. As long as the classes can be projected into a good ratio of between-class-variance : within-class-variance, LDA will work nicely. LR requires only clear linear class boundaries and is less concerned about the homogeneity within the classes.

Four different models were built:

LR-soft: LR using all spectra, with crisp and soft reference
LR-crisp: LR using spectra with hard labels only

Table 2: Difference between diagnosis of a patient's tumour and detailed histological diagnosis. The patient's tumour grade according to the usual diagnostic procedure (rows) compared to the tissue dominating our reference section (columns). Morphologically varying differentiation within astrocytoma tissues is described in analogy to the grading of the WHO (see text).

Diagnosis for	**Main tissue morphology of the section** ($\geq 50\,\%$ of section's area)									
Patient	Normal	not sure*	Border†	°II	°II-III	°III	°III-IV	°IV	Necrosis	total
Control (normal)	9									9
Astro. °II	3	1	1	2						7
Astro. °III	2	2	3	4	4	5				20
Glioblastoma	4	3	9	5	4	12	6	6	8	60‡
total	18	6	13	11	8	17	6	6	8	96

* neuropathologist was not sure whether the tissue contains tumour cells † tumor cells infiltrate normal tissue ‡ 3 samples too heterogeneous: no dominating tissue

LDA: LDA using the same spectra as LR-crisp
LR-highwn: LR using all spectra, but only $\Delta\tilde{\nu} \geq 2800\,\mathrm{cm}^{-1}$

As a well established technique, LDA serves as a standard. Together with LR-crisp and LR-soft, the difference between LDA and LR on the same samples can be separated from the influence of the borderline samples with soft labels on the (LR) models. LR-highwn models test whether unfiltered probes are a promising direction for Raman guidance in astrocytoma surgery.

10 000 spectra were randomly drawn with replacement from all training patients (see set-up of the cross validation below). To obtain a more balanced training set, the probability to select A °II or N spectra was increased as follows. The odds of crisp spectra were 1.5 for class N, 3 for class A °II, and 1 for A °III+. The odds for soft labelled samples were the weighted average of these values. The resulting training sets consist of ca. 33 % N, 37 % A °II, and 30 % A °III+ (unweighted: 35, 24, and 41 %). LR-soft and LR-highwn models used exactly the same training spectra, and so did the LDA and LR-crisp models (where the training spectra were drawn of the crisp spectra of the same training patients).

LR and LDA models were calculated using the R packages nnet [35] and MASS [35], respectively.

Validation. A 125× iterated 8-fold cross validation scheme was used, randomly splitting patient-wise to ensure statistical independence.

Iterations of the cross validation scheme reduce the random uncertainty on the performance measures. Moreover, they allow to measure model stability (with respect to changing training sets). Stability was calculated as the standard deviation of the predictions for each spectrum across the iterations.

Sensitivity and specificity computations used ROCR [54].

4 Results and Discussion

4.1 Reference Diagnosis

Gliomas are very polymorphous tumours, and thus the samples are heterogeneous: even GBM frequently have regions that are morphologically similar to low grade astrocytomas. We refer to this grading of morphological similarity as "detailed diagnosis".

Table 2 shows the main tissue found in the reference sections compared to the tumour grade reported for the patient. The patient's tumour grade was determined indepently from our samples using using samples taken explicitly for histologic diagnosis and grading of the patient's tumour (yet during the same surgery). This process includes a histological review of ambiguous cases (either due to uncertainty of the neuropathologist or due to differences between histology and clinical and radiology findings) by the brain tumour reference centre in Bonn/Germany.

About 10 % of the tumour samples were actually dominated by normal tissue. For another 5 % of the tumour samples pathologist was not certain whether the predominant tissue contained tumour cells at all. Only $\frac{1}{4}$ of the sections consisted mainly of a tissue that would define the patient's diagnosis, *i. e.*, tissue of a grade that would establish the overall diagnosis for this patient. None of the 96 sections gave evidence of a possibly higher tumour grade than documented for the patient.

Grading a patient's tumour versus grading a tissue for surgical guidance. It is important to realise how much the concept of grading in these detailed diagnoses differs from the usual process of grading a patient's tumour. The diagnosis for the patient is given by the most de-differentiated (highest grade) tissue in the tumour – however small this region may be. In contrast, an intra-operative tool must distinguish different tissues within one tumour. The detailed diagnoses discussed here refer to this second grading concept. The heterogeneity of the samples enlarges the differences of the results of the two diagnostic concepts, resulting in the triangular shape of table 2.

Chemometric implications. From a chemometric point of view, diagnosing the patient's tumour is a rather problematic (ill-posed) procedure. Firstly, the heterogeneity of the tumour tissue leads to a high sampling uncertainty, the more as tumour tissue is usually sampled at the tumour border rather than at the core. Secondly, finding the (possibly tiny region with the) highest grade tissue is an awkward operation, as the maximum-operator captures large amounts of uncertainty. The detailed diagnosis does not share these two problems, as it does not extrapolate a statement for the whole tumour but aims solely at the present tissue.

A third step that introduces additional uncertainty is enforcing a crisp diagnosis. Mathematically, dichotomization

corresponds to an information loss [46]. Thus, hardening *e. g.* ambiguous diagnoses arising from continuous dedifferentiation into a few available groups (*e. g.* WHO grades) causes an information loss, *i. e.* increased variance. *E. g.* Kendall et al [55] find highest disagreement for intermediate grades: the fraction of diagnoses where all pathologists agreed with respect to the number of cases where at least 2 of the 3 pathologists agreed was 91 – 55 – 0 – 51 – 84 % of the spectra for normal tissue – intestinal metaplasia – low grade dysplasia – high grade dysplasia, and adenocarcinoma. Partial memberships avoid the additional variance caused by the hardening as well as information loss due to exclusion of samples.

Astrocytomas are well known to grow infiltratively and de-differentiate, leading to heterogeneous samples. Finding lower grade tissue than the patient's diagnosis states was therefore expected. The extent, however, is consistently over all (patient's) tumour grades much larger than anticipated.

Classification models can be successfully trained even if a certain amount of the training data is mislabelled. However, the fraction of mislabelled data must not be too high, and usually more samples are required to outweigh those "bad" examples. However, even if the high grade astrocytomas are not further distinguished into A °III and GBM, less than half of the high grade tumour samples consist mainly of high grade tissue. And less than a third of the A °II samples is dominated by low grade tumour tissue. In this situation, successful classifier training is impossible. Building a classification model for intra-operative grading of tumour tissue thus needs detailed reference diagnoses across the actually measured tissue.

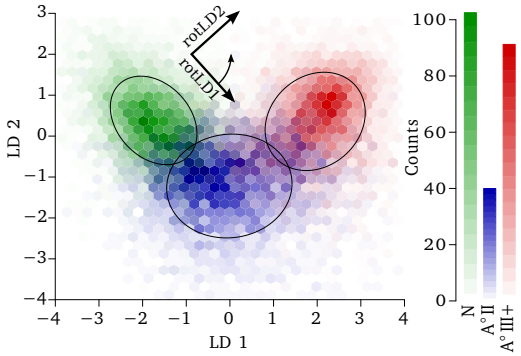


Fig. 5: 2d histogram of the crisp data in LD space. The ellipses mark the area that contains 50 % of the samples of a bi-variate normal distribution with mean and covariance matrix as observed for each class. After rotating the original LD space 48 °counterclockwise (small arrow), the rotated LDs (inset directions) are the horizontal and vertical axes.

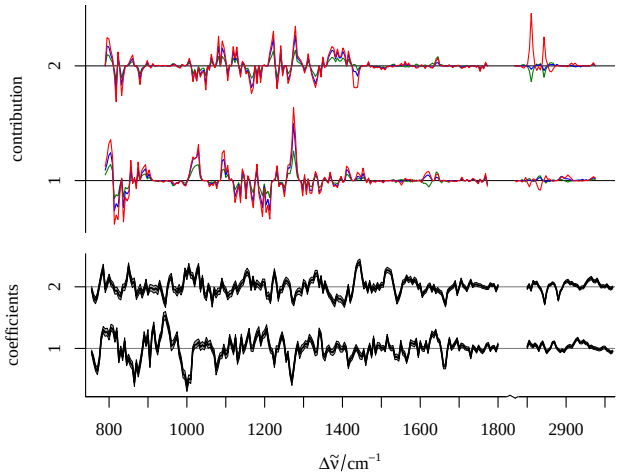


Fig. 6: Coefficient spectra of the rotated LDA (lower two rows, median and quartiles) and median contribution of the spectral regions to the LD score (upper two rows, N: green, A °II: blue, A °III+: red). The quartiles of the contributions are available as fig. S.3. The 1st rotLD has positive scores for tumour tissue and the 2nd positive scores for high grade tumours.

4.2 Spectroscopic Classification of the Astrocytoma

4.2.1 Descriptive LDA and Spectroscopic Interpretation

This study focuses towards a time critical predictive application. In order to give a proof of concept for feasible measurement times, and to gain on overview of the predictive performance for noisy spectra, signal-to-noise ratio and spectral resolution were chosen lower than appropriate for decriptive models aimed at spectral interpretation and discovery of new biochemical features in the samples. Yet the spectroscopic meaning of the models can of course be studied.

LDA models derive the final class membership values (*i. e.* posterior probabilities) from the Euclidean distance between spectrum and class means in LD space [10, 56]. Figure 5 shows a 2d histogram [57] of all crisp spectra in LD space. LD 1 coincides with the direction of increasing malignancy. This has been observed with a four-class model based on infrared data as well [58].

For the interpretation of the probed biochemical properties a rotated version of the LD model is more suitable: the three classes form an approximately orthogonal triangle (this is true for all 8× 125 models). Rotating the projection (*i. e.* fig. 5) about 48 °counterclockwise results in a LD space where the two directions distinguish normal from tumour tissues and (normal and) low grade tissues from high grade tissue (directions inset in fig. 5). LDA is invariant to rotation, mirroring, and translation of the LD space, but not to scaling. All models were rotated so that the mean of the A °II class lies exactly right of the mean of class N in the new rotLD space. Mirroring was not needed, neither were the models shifted. We obtained a set of rotated LDA models that are as similar as possible to each other while retaining prediction and coordinate origin.

These rotated models are now open to spectroscopic interpretation in two ways. On the one hand, the coefficients (lower part of fig. 6) probe the corresponding wavenumbers of the spectra. In order to judge the influence of such a coefficient on the prediction, not only magnitude and sign of the coefficient but also the magnitudes and signs of the spectra must be taken into account. Element-wise multiplication of spectra and coefficients yields "contribution spectra" that give direction and magnitude of the contribution of the respective wavenumber to the rotLD scores. $7.5 \cdot 10^5$ contribution spectra were calculated for each class and direction by

randomly picking a spectrum according to its membership to the class and one of the models. The upper part of fig. 6 shows the median contribution for both rotLD directions and all three classes, fig. S.3 also gives the observed quartiles.

For spectroscopic interpretation, one further issue arises: the intensities are not pure (difference) Raman spectra. The raw spectra showed a high and wavenumber dependent background, particularly in the low wavenumber region below $\Delta\tilde{\nu} = 1800\,\mathrm{cm}^{-1}$. Baseline correction was applied to correct for most of its influence, but the corrected spectra clearly comprise residual background (figs. 4 and S.3). For the spectroscopic interpretation of the models it must be kept in mind that coefficients may "deliberately" probe the background for two opposite reasons. Firstly, the projection can include a background correction. Secondly, the background of the raw spectra increases with malignancy, and the residual background correlates in shape and magnitude with malignancy, too. Such a background signal may be used for classification. Several physical and chemical effects may contribute to the background, *e. g.* fluorescence emission of the sample. In fact, autofluorescence at lower wavelengths has been studied in terms of its brain tumour diagnostic potential [40].

In the spectral region below $\Delta\tilde{\nu} = 900\,\mathrm{cm}^{-1}$, the residual background signal overwhelms any possible Raman contribution. As the coefficients (fig. 6, lower part) are rather large (but changing sign), the models do actually use the residual background. On the other hand, some coefficients in this region correspond to bands of substances that are known to change in the direction probed by the models, *e. g.* the DNA band at $785\,\mathrm{cm}^{-1}$ and the glycogen signal at $850\,\mathrm{cm}^{-1}$ are both expected to be more intense in the tumours (in our models these wavenumbers indicate high grade tissues). In contrast, the signal at $865\,\mathrm{cm}^{-1}$ is probed for a decrease in tumour tissues. This Raman shift is considered typical for phosphatidylethanolamine [59], for which a decrease in malignant astrocytomas was spectroscopically found by Beljebbar et al [60]. Also the series of alternating coefficients between 1100 and $1200\,\mathrm{cm}^{-1}$ seems to probe the background of the signal in addition to probing specific bands.

In the high wavenumber region between 2800 and $3025\,\mathrm{cm}^{-1}$, a decrease in the νCH stretching bands with increasing malignancy is obvious from the spectra. The bands at 2850 ($\nu_s CH_2$) and $2885\,\mathrm{cm}^{-1}$ ($\nu_s CH_3$, νCH_2 in Fermi resonance) are typical for many brain related lipids [53, 59, 61], while the $\nu = CH$ band at 3010 – $3015\,\mathrm{cm}^{-1}$ indicates unsaturated compounds. The respective antisymmetric bands overlap with the CH stretchings of other compounds such as proteins, DNA, RNA, and glycogen (see *e. g.* [13, 62]) and cannot easily be separatedq. Interestingly, the first rotLD (normal vs. tumour) does hardly use this spectral region. Instead, the 2nd rotLD, *i. e.* the recognition of high grade tissues, on median collects about half of the total score by detecting the lack of the $\nu_s CH_2$ and $\nu_s CH_3$ bands. This corresponds to the observation that many low grade tumour spectra show higher intensities here than the gray tissue, while lower quartile of the observed intensities (lower trace of the A °II spectra in fig. S.3) is more similar to the distribution of the high grade tissue spectra. The same holds for νC−C at $1065\,\mathrm{cm}^{-1}$, which is also used by the 2nd rotLD. Biochemically, this translates to the well known general decrease in lipids for the malignant astrocytomas. These contributions are counteracted by the contribution at $1440\,\mathrm{cm}^{-1}$ (CH_2 deformation). This band does not show a consistent intensity for the high grade tumours in the centered spectra. Another interesting property of the models is that they do not use the CH_2 twisting band at $1295\,\mathrm{cm}^{-1}$ though the band is present in the centered spectra.

The most prominent contribution to the distinction between normal and tumour tissues is at $1270\,\mathrm{cm}^{-1}$ (δ=CH of unsaturated fatty acids), where the high grade tumours gain about $\frac{3}{4}$ of their final score, and the low grade tumours about half. In addition, both rotLDs use the band at $1665\,\mathrm{cm}^{-1}$. The νC=C of unsaturated fatty acids ($1660\,\mathrm{cm}^{-1}$), however, hardly contributes to the recognition of tumours, but it serves for the detection of high grade tissues. Colesterol esters give a signal here ($1670\,\mathrm{cm}^{-1}$), too. The band position for cholesterol itself is at $1675\,\mathrm{cm}^{-1}$ [59], where the coefficients of both rotLDs are essentially zero. The centered spectra suggest a lower content of unsaturated lipids (and cholesterol esters) with increasing malignancy, and in white matter. This is in agreement with the results presented by Köhler et al [53] and Beljebbar et al [60]. The second paper [60] reports higher contents of oleic acid in an animal model of GBM than in the surrounding normal grey tissue while cholesteryl oleate decreased, the sum of both being lower in the tumour than in the grey tissue. Our models probe this sum rather than the individual contributions, as they do not try to distinguish between unsaturated fatty acids and their cholesterol esters (which would be indicated by a change in the sign of the coefficients at $1665\,\mathrm{cm}^{-1}$).

As both the δ=CH and the νC=C bands lie in the region of the amide III and amide I bands of proteins (at 1225 – 1300, and 1645 – $1675\,\mathrm{cm}^{-1}$, respectively), the interpretation as lower amounts of unsaturated lipids must be judged against changes in the protein content. The interpretation in favour of the lipids is strengthened by two arguments. Firstly, the centered spectra clearly show that the probed bands are narrow. Secondly, we observed that the normalization and centering effectively deletes the spectral signature of the proteins. Moreover, the tumours have lower lipid : protein ratios than the normal grey tissue [16, 53, 60]. The normalization may be influenced by the lipid content of the tissue as the $\nu_{as} CH_2$ and $\nu_{as} CH_3$ stretching vibrations lie in the spectral range considered for normalization. Thus, the calculated intensity may be higher than it should be for normalizing on the protein content. In this case, the remaining spectral signature of proteins in the centered spectra should be positive. Yet the observed differences at 1270 and $1665\,\mathrm{cm}^{-1}$ are negative.

On the other hand, a close inspection of the phenylalanine band at $1005\,\mathrm{cm}^{-1}$ does shows a weak positive residual signal for the A °III+ tumours which gives an important contribution to the distinction between normal and tumour tissue. We attribute the unusual width to the downsampling of the spectral resolution affecting this sharp and intense band rather heavily (the smoothing interpolation preserves signal area rather than height).

Koljenović et al [12] reported high glycogen contents of (vital) GBM tissue. Our 2thnd rotLD takes bands at 850, 1090, and $1340\,\mathrm{cm}^{-1}$ (for a reference spectrum see *e. g.* [62])

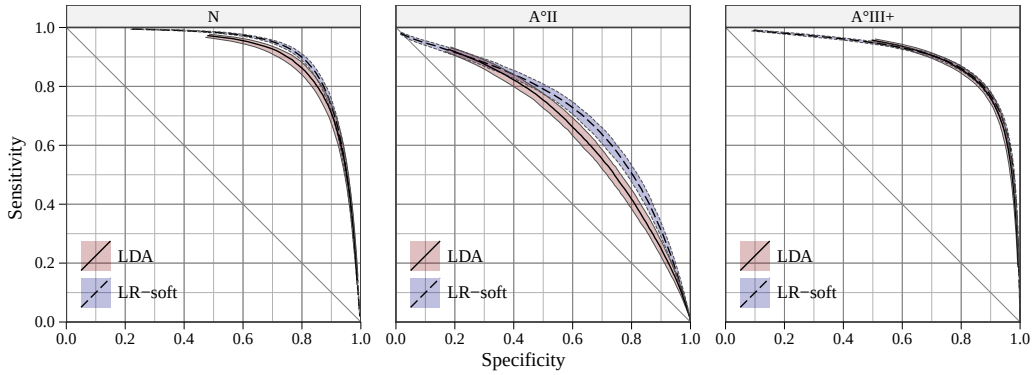


Fig. 7: Performance of the LDA (orange, continuous) and LR-soft (violet, dashed) models: median and quartiles observed over the 125 iterations of the cross validation, hardening thresholds 0.01 – 0.99. The LR-soft models recognize normal tissue more sensitively, and also perform better for the low grade tumours. The performance for the high grade tumours is virtually the same.

as evidence for high grade tumours. However, due to the negative residual baseline of the high grade tissues in these regions the median of the final contribution is negative (towards low grade or normal tissue).

A rather weak but broad contribution to the distinction of normal vs. tumour tissues is at 1635 cm^{-1}, corresponding to the center of the water deformation band. Gliomas have been reported to contain more water than grey and white matter [53]. The water content of tissues is defined, and changes do have diagnostic importance. For intra-operative use two points should be kept in mind, though. Firstly, the normal tissue close to the tumour is often edematous, and additional swelling of the tissue may be caused by the operation. Future models that should use changes in the water content must be trained with samples of swollen normal tissue. Secondly, while tissues have a defined water content, during surgery physiological solution is used to flush and moisten the exposed tissues, and superficial water may show up in the Raman spectra.

The upper quartiles of the Raman spectra (particularly of the high grade tumours, see also fig. S.3) clearly show the typical pattern of hemoglobin which is resonance enhanced with the 785 nm excitation. In general, malignant samples were more bloody but one of the control samples also contains large amounts of blood. In spite of the strong characteristic spectrum and the correlation with the malignancy, hemoglobin hardly contributes to the classification. Again, while growth of new blood vessels indicates malignant tumours, the occurence of blood is not of much diagnostic value during surgery. Diagnostic models must not be confused by the presence of blood (nor by hemoglobin oxidation state, see *e. g.* [62]).

4.2.2 General predictive performance of the models

The overall performance for the crisp spectra is rather similar for the LR-soft and LDA models, with good recognition of normal tissues and almost as good recognition of the high grade morphologies (fig. 7). Normal tissue is recognised slightly more sensitive than specific, while the recognition of high grade tissues is more specific than sensitive.

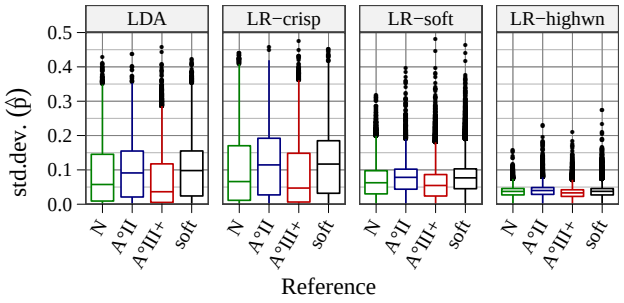


Fig. 8: Model stability: the standard deviation of the predictions observed over the iterations of the cross validation (details see text). A version detailing also the classes in the prediction is available as fig. S.1. The boxes mark median and quartiles, the whiskers extend to the last value inside $1\frac{1}{2}\times$ the inter quartile range (IQR) from the box, all further values are marked by points [47].

The decriptive analysis, particularly fig. 5, showed the low grade tissues rather encompassed between the normal and high grade tumour tissue, while normal and high grade tissues are substantially better separated. Consequently, the low grade morphologies are most difficult to recognize. Both boundaries, against the normal as well as against the high grade tissues, contribute about the same amount to the total misclassifications. This is revealed also by inspection of the MAE and RMSE (fig. S.2).

Model stability. Models built with large numbers of parameters but comparably few training samples may suffer from instability. For crisp samples, unstable predictions imply systematically bad performance. Therefore, the prediction stability is checked by calculating the standard deviation (sd) over all iterations of the cross validation for each spectrum. Fig. 8 summarizes the results for the different reference groups, while fig. S.1 details also the differences between the predicted memberships. Many predictions of models trained on crisp samples only (LDA and LR-crisp) are much less stable than the models trained on the soft spectra as well (LR-soft and LR-highwn): the upper quar-

tile are 0.15 and 0.18 vs. 0.10 and 0.05, respectively. On the other hand, the median stability of LDA and LR-soft is practically the same (0.75 vs. 0.75). The LR-highwn estimate much fewer parameters ($2 \times (46+1) = 94$; vs. > 500 for the other models). Accordingly, their predictions are much more stable. In general, the predictions are more stable for the two "easy" groups of spectra, normal and high grade tissues, than for low grade tissue and soft samples.

4.2.3 Comparison of LDA and LR: models using crisp training samples only.

The LDA and LR-crisp models perform equally for the low grade tumours (fig. S.4), though the LDA recognizes normal and high grade tumour tissues slightly more sensitively. MAE (fig. S.2) is 1.5 % elevated (between 1 % lower for normal and low grade tissues and 3 % higher for high grade tumour tissues), the RMSE 3 % (1 –4.5 %). The higher increase in the RMSE indicates large deviations in few cases, *i. e.* LR-crisp predicts a few more samples grossly wrong.

In conclusion, the unequal covariance structure between the classes (fig. 5) does not disturb the LDA. The gain in power due to the LDA's parametric model outweighs the LR's advantage with heterogeneous classes. Also, the predictions of the LR-crisp models are less stable than the LDA's predictions (median stability 0.09 vs. 0.075, see figs. 8 and S.1). We may safely conclude that a major part of the overall misclassifications is correlated to the models' instability.

4.2.4 Do soft labelled spectra improve the models?

However, the LR models improve when the soft labelled spectra are included into the model training (fig. 7). Detailed specificity-sensitivity-diagrams comparing LDA vs. LR-crisp, and LR-crisp vs. LR-soft are in fig. S.4 in the supplementary material. For the remaining discussion here, we compare the LR-soft to the LDA models as the LDA is superior to LR-crisp (see above).

The most important improvement over the LDA is in the recognition of A °II and the borderline cases (soft spectra), while the recognition of normal tissue shows only a minute increase in sensitivity. The LR-soft models have an slightly increased overall MAE (+1.4 %), while the RMSE improves strongly ($-9\frac{1}{2}$ %), see fig. S.2 for details. More precisely, the MAE is highly increased for the crisp high grade tumour tissues (+25 %). Also the crisply labelled low grade tissue have an elevated MAE ($+5\frac{1}{2}$ %), whereas the normal *tissues* (*i. e.* group of samples as opposed to class membership) are hardly affected (+1 %). This increase for crisply labelled samples is counterweighted by a marked decrease in the MAE for soft labelled samples (−8 %). The RMSE of the crisp A °III+ samples is not affected ($-1\frac{1}{2}$ ‰), while the crisp low grade tissues are recognized much better (−7 %), and the RMSE for normal tissues and soft samples drop strongly about $\frac{1}{7}$ (−14 %) and $\frac{1}{8}$ (−13 %), respectively.

The corresponding changes in the predicted memberships of the three *classes* in the MAE are +3 % (N), −2 ‰(A °II), and +10 % (A °III+). The RMSEs improve by −8 %, $-13\frac{1}{2}$ %, and −6 %, respectively.

The much improved RMSEs in contrast to the slightly worse MAEs mean that the predictions of the LR-soft models deviate slightly from the reference for many samples, whereas the LDA models have larger deviations in fewer cases (fig. S.2).

The stability of predictions (figs. 8 and S.1) for the low grade tumours and the soft samples improves perceivably. The distribution is is much narrower for the LR-soft: the inter quartile range (IQR) of the LR-soft's stability is only half compared to the LDA's, *i. e.* 0.06 instead of 0.13. Compared to the LDA, the upper quartile is lower (0.10 vs. 0.15), but the median is basically unaffected (LR-soft:0.70, LDA: 0.75). Thus, some samples are predicted considerably less stable by LDA and LR-crisp.

This as well as the behaviour of MAE and RMSE are in accordance with the LR-soft models having smoother transitions between the classes than both LDA and LR-crisp models: *i. e.* the posterior probability function (compare fig 1) is steeper for the models built without borderline cases.

An additional benefit for the LR-soft is the 50 % increase in patient numbers. The increased sample base is extremely important for the modeling of the A °II class: LR-soft uses four times as many patients to model the low grade tissue and $4\frac{1}{2}\times$ the number of spectra (tab. 1). The low grade tumour spectra are encompassed between the two other classes, and crisp training samples are rare (tab. 2). Such classes are apt to being ignored between the encompassing classes. While we counteract this risk by drawing more A °II spectra for training, additional patients help much more effectively.

Whether and to what extent soft labelled samples actually help in model training depends also on the uncertainty in the reference labels. If the soft labelled spectra include a disproportionally large amount of wrong labels, the model's performance may even deteriorate. Particularly, inaccurately labelled samples close to the class boundaries are critical. For our data, however, this is outweighted by the advantages: the predictive performance increased.

For our application, the improvement in predicting the borderline cases is the most important advance as this group of samples is the target of the technique. In conclusion, LR successfully uses the additional information supplied by the soft labelled spectra of borderline cases.

4.2.5 Models using the C-H stretching region only.

A diagnostic model using the high wavenumber region above *e. g.* 2600 cm^{-1} Raman shift would allow to use cheaper and smaller unfiltered probes. Unfortunately, the models trained exclusively on the C-H stretching region perform much worse than all other models (fig. 9). Already the low sensitivity for normal tissue immediately disqualifies the model. In fact, the models predict intermediate memberships for nearly all spectra (not shown).

This lack in performance insofar astonishing, as the high wavenumber region not only offers the best signal to noise ratio but has distinct differences between the classes (fig. 4). However, while the class median spectra are different, the distributions of the spectra of the different classes overlap heavily. Different normalisation may help, but the choice is

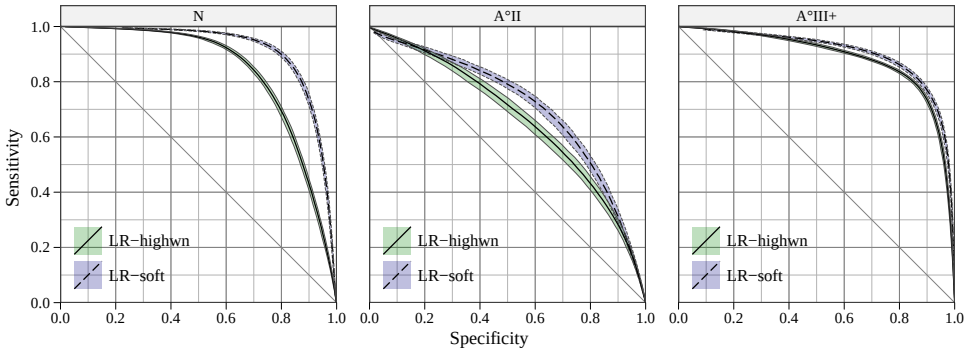


Fig. 9: Performance of the LR using the high wavenumber region only (LR-highwn; green) and LR-soft (using also the spectral range 755 – 1800 cm^{-1}; violet, dashed) models: median and quartiles of the 125 iterations of the cross validation, thresholds 0.01 – 0.99. The fingerprint region below 1800 cm^{-1} is needed for sucessful astrocytoma grading.

restricted (normalising from 2800 to 3025 cm^{-1} did not visibly improve the situation). Note that the restiction of the spectral range obviously includes the preprocessing.

The descriptive interpretation of the LDA models revealed that particularly the distinction between normal and tumour tissues accumulates differences spread out over the fingerprint region below 1800 cm^{-1}, whereas the νC−H bands mainly separated the high grade tumour tissues. Visual examination of the spectra in fig. 4 and spectroscopic knowledge suggest the ν=CH signal at 3010 cm^{-1} as substitute for the rotLDA's δ=CH band at 1270 cm^{-1}. Yet the models are not able to use it efficiently. Also, the relatively few bands in the high wavenumber region seem to limit the LR-highwn models.

The shortcomings of the LR-highwn models cannot be due to instability which limits the hope for improvement with additional patients.

5 Conclusions

We presented *predictive* Raman spectroscopic grading of astrocytomas using fibre-optic probes and moist bulk samples in an experimental setup that is oriented towards the needs of surgeons for prospective intra-operative in-vivo diagnostics.

The detailed histological assessment of the samples revealed that only $\frac{1}{4}$ of the reference sections were dominated by tissues defining the patient's tumour grade. Therefore, detailed reference diagnosis is crucial for the classifier training. The large amount of normal tissue observed in the tumour samples emphasises the need for better and non-invasive intra-operative diagnosis.

Including borderline cases into the LR training (soft models) increased sensitivities and specificities to 85 %, 67 %, and 84 % for normal tissue, low grade and high grade tumour tissues, respectively (median over all iterations, sensitivity equal to specificity). Both fingerprint region and C-H-stretching bands of the Raman spectrum are used.

Soft classification offers several advantages over traditional hard classification. Partial memberships model the spatial transitions (infiltration) and de-differentiation of morphologically different tissues. Thus, borderline cases are used as examples of the sought decision border. This is particularly important as our task is delineating borders rather than recognition of typical examples.

Borderline cases describe class overlap, and are needed to realistically estimate class separation. Excluding borderline cases implies a risk of overestimating the class separation. On the other hand, high amounts of mislabelled (crisp or soft) samples cause overestimation of the class overlap.

Allowing soft training samples also increases the available number of training samples. In small sample size situations every sample is needed. Yet, borderline samples will always enlarge the available sample base.

Partial membership in reference data avoids the information loss associated with forcing the histologist to decide for one of the hard classes for the borderline cases .

LR training including the borderline cases increased the predictive performance: the advantages of having more samples and particularly samples of borderline cases prevail for the astrocytoma grading. Including difficult (borderline) cases into the data analysis model is an important step towards real-word application. Difficult cases not only cannot be avoided in this application, they are the actual target.

6 Acknowledgements

C. Beleites acknowledges financial support by a scholarship of Deutsche Telekom Stiftung, the IRCCS Burlo Garofolo, and the Associazione per i Bambini Chirurgici del Burlo.

References

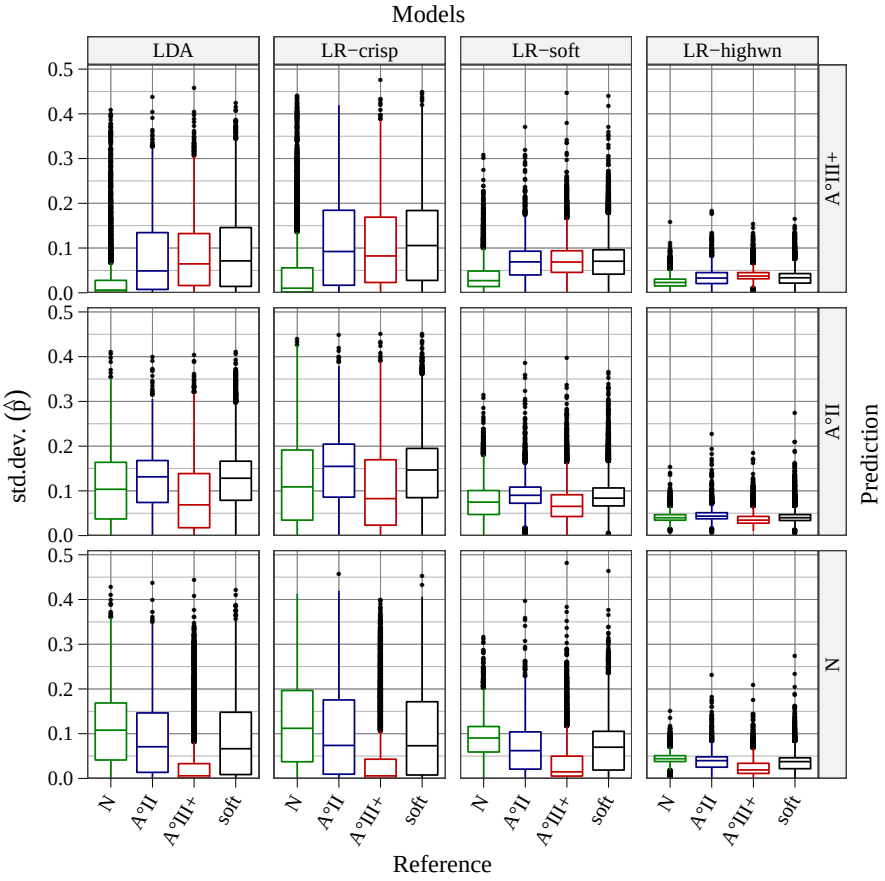
1. VandenBerg SR (1992) Current diagnostic concepts of astrocytic tumors. J Neuropathol Exp Neurol 51(6):644–657
2. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P (2007) The 2007 who classification of tumours of the central nervous system. Acta Neuropathol 114(2):97–109
3. Kros JM, Gorlia T, Kouwenhoven MC, Zheng PP, Collins VP, Figarella-Branger D, Giangaspero F, Giannini C, Mokhtari K, Mørk SJ, Paetau A, Reifenberger G, van den Bent MJ (2007) Panel review of anaplastic oligodendroglioma from european organization for research and treatment of cancer trial 26951: assessment of consensus in diagnosis, influence of 1p/19q loss, and correlations with outcome. J Neuropathol Exp Neurol 66(6):545–551
4. Duran I, Raizer JJ (2007) Low-grade gliomas: management issues. Expert Rev Anticancer Ther 7(12 Suppl):S15–S21

5. Stupp R, Reni M, Gatta G, Mazza E, Vecht C (2007) Anaplastic astrocytoma in adults. Crit Rev Oncol Hematol 63(1):72–80
6. Diener HC, Putzki N, Berlit P, Hacke W, Hufnagel A, Hufschmidt A, Mattle H, Meier U, Oertel W, Reichmann H, Rieckmann P, Schmutzhard E, Wallesch CW, Weller M (2008) Leitlinien für Diagnostik und Therapie in der Neurologie, 4th edn. Thieme, Stuttgart, URL `http://www.dgn.org/-leitlinien-online.html` Accessed 2011-01-08
7. Stummer W, Pichlmeier U, Meinel T, Wiestler OD, Zanella F, Reulen HJ, Group ALAGS (2006) Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase iii trial. Lancet Oncol 7(5):392–401
8. Sobottka SB, Geiger KD, Salzer R, Schackert G, Krafft C (2009) Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery. Anal Bioanal Chem 393(1):187–195
9. Krafft C, Sergo V (2006) Biomedical applications of raman and infrared spectroscopy to diagnose tissues. Spectroscopy 20:195 – 218
10. Krafft C, Steiner G, Beleites C, Salzer R (2009) Disease recognition by infrared and raman spectroscopy. Journal of Biophotonics 2(1-2):13–28
11. Kendall C, Isabelle M, Bazant-Hegemark F, Hutchings J, Orr L, Babrah J, Baker R, Stone N (2009) Vibrational spectroscopy: a clinical tool for cancer diagnostics. Analyst 134(6):1029–1045
12. Koljenović S, Choo-Smith LP, Schut TCB, Kros JM, van den Berge HJ, Puppels GJ (2002) Discriminating vital tumor from necrotic tissue in human glioblastoma tissue samples by raman spectroscopy. Lab Invest 82(10):1265–1277
13. Koljenović S, Schut TCB, Wolthuis R, de Jong B, Santos L, Caspers PJ, Kros JM, Puppels GJ (2005) Tissue characterization using high wave number raman spectroscopy. J Biomed Opt 10(3):031,116–1 – 031,116–11
14. Krafft C, Sobottka SB, Schackert G, Salzer R (2005) Near infrared raman spectroscopic mapping of native brain tissue and intracranial tumors. Analyst 130(7):1070–1077
15. Amharref N, Beljebbar A, Dukic S, Venteo L, Schneider L, Pluot M, Manfait M (2007) Discriminating healthy from tumor and necrosis tissue in rat brain tissue samples by raman spectral imaging. Biochim Biophys Acta 1768(10):2605–2615
16. Krafft C, Kirsch M, Beleites C, Schackert G, Salzer R (2007) Methodology for fiber-optic raman mapping and ftir imaging of metastases in mouse brains. Anal Bioanal Chem 389(4):1133–1142
17. Beljebbar A, Dukic S, Amharref N, Manfait M (2010) Ex vivo and in vivo diagnosis of c6 glioblastoma development by raman spectroscopy coupled to a microprobe. Anal Bioanal Chem
18. Kirsch M, Schackert G, Salzer R, Krafft C (2010) Raman spectroscopic imaging for in vivo detection of cerebral brain metastases. Anal Bioanal Chem 398(4):1707–1713
19. Krafft C, Sobottka SB, Schackert G, Salzer R (2004) Analysis of human brain tissue, brain tumors and tumor cells by infrared spectroscopic mapping. Analyst 129(10):921–925
20. Beleites C, Salzer R (2008) Assessing and improving the stability of chemometric models in small sample size situations. Anal Bioanal Chem 390(5):1261–1271
21. Utzinger U, Richards-Kortum RR (2003) Fiber optic probes for biomedical optical spectroscopy. J Biomed Opt 8(1):121 – 147
22. Santos LF, Wolthuis R, Koljenović S, Almeida RM, Puppels GJ (2005) Fiber-optic probes for in vivo raman spectroscopy in the high-wavenumber region. Anal Chem 77(20):6747–6752
23. Prayson RA, Agamanolis DP, Cohen ML, Estes ML, Kleinschmidt-DeMasters BK, Abdul-Karim F, McClure SP, Sebek BA, Vinay R (2000) Interobserver reproducibility among neuropathologists and surgical pathologists in fibrillary astrocytoma grading. J Neurol Sci 175(1):33–39
24. Louis DN, Cavenee WK, Ohgaki H, Wiestler OD (eds) (2007) WHO Classification of tumors of the central nervous system, Word Health Organization, chap Astrocytic tumors
25. Wrensch M, Rice T, Miike R, McMillan A, Lamborn KR, Aldape K, Prados MD (2006) Diagnostic, treatment, and demographic factors influencing survival in a population-based study of adult glioma patients in the san francisco bay area. Neuro Oncol 8(1):12–26
26. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning; Data mining, Inference and Prediction. Springer Verlag, New York
27. Beleites C, Baumgartner R, Bowman C, Somorjai R, Steiner G, Salzer R, Sowa MG (2005) Variance reduction in estimating classification error using sparse datasets. Chemom Intell Lab Syst 79:91 – 100
28. Kowalski B, Wold S (1982) Pattern recognition in chemistry. In: Krishnajah PR, Kanal LN (eds) Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics, vol II, North-Holland, Amsterdam, pp 673 – 697
29. Huberty CJ (1994) Applied Discriminant Analysis. John Wiley & Sons, Inc., New York
30. Vittinghoff E, McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and cox regression. Am J Epidemiol 165(6):710–718
31. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49(12):1373–1379
32. Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sensing of Environment 80(1):185–201
33. Sattlecker M, Bessant C, Smith J, Stone N (2010) Investigation of support vector machines and raman spectroscopy for lymph node diagnostics. Analyst 135(5):895–901
34. Hedegaard M, Krafft C, Ditzel HJ, Johansen LE, Hassing S, Popp J (2010) Discriminating isogenic cancer cells and identifying altered unsaturated fatty acid content as associated with metastasis status, using k-means clustering and partial least squares-discriminant analysis of raman maps. Anal Chem 82(7):2797–2802
35. Venables WN, Ripley BD (2002) Modern Applied Statistics with S, 4th edn. Springer, New York, URL `http://www.stats.ox.ac.uk/pub/MASS4` Accessed 2010-02-21
36. Teh S, Zheng W, Ho K, Teh M, Yeoh K, Huang Z (2009) Near-infrared raman spectroscopy for gastric precancer diagnosis. J Raman Spectrosc 40(8):908–914, cited By (since 1996) 4
37. da Silva Martinho H, de Oliveira Monteiro da Silva CM, Yassoyama MCBM, de Oliveira Andrade P, Bitar RA, do Espírito Santo AM, Arisawa EAL, Martin AA (2008) Role of cervicitis in the raman-based optical diagnosis of cervical intraepithelial neoplasia. J Biomed Opt 13(5):054,029
38. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z (2010) Near-infrared raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach. Br J Surg 97(4):550–557
39. Nijssen A, Schut TCB, Heule F, Caspers PJ, Hayes DP, Neumann MHA, Puppels GJ (2002) Discriminating basal cell carcinoma from its surrounding tissue by raman spectroscopy. J Invest Dermatol 119(1):64–69
40. Majumder SK, Gebhart S, Johnson MD, Thompson R, Lin WC, Mahadevan-Jansen A (2007) A probability-based spectroscopic diagnostic algorithm for simultaneous discrimination of brain tumor and tumor margins from normal brain tissue. Appl Spectrosc 61(5):548–557
41. Majumder SK, Keller MD, Boulos FI, Kelley MC, Mahadevan-Jansen A (2008) Comparison of autofluorescence, diffuse reflectance, and raman spectroscopy for breast tissue discrimination. J Biomed Opt 13(5):054,009
42. Kanter EM, Majumder S, Vargis E, Robichaux-Viehoever A, Kanter GJ, Shappell H, III HWJ, Mahadevan-Jansen A (2009) Multiclass discrimination of cervical precancers using raman spectroscopy. J Raman Spectrosc 40:204–211
43. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish CS (ed) Artificial Intelligence Proceedings 14^{th} International Joint Conference, 20 – 25. August 1995, Montréal, Québec, Canada, Morgan Kaufmann, USA, pp 1137 – 1145
44. Kohavi R (1995) Wrappers for performance enhancement and oblivious decision graphs. Phd thesis, Department of Computer Science, Stanford University
45. Ellison S, Fearn T (2005) Characterising the performance of qualitative analytical methods: Statistics and terminology. TrAC 24(6):468–476
46. Fedorov V, Mannino F, Zhang R (2009) Consequences of dichotomization. Pharm Stat 8:50–61

47. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0
48. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New York, URL http://had.co.nz/ggplot2/book Accessed 2010-07-08
49. Beleites C, Sergo V (????) hyperspec: a package to handle hyperspectral data sets in r, in preparation
50. Bengtsson H, Riedy J (2008) R.matlab: Read and write of MAT files together with R-to-Matlab connectivity. URL http://www.braju.com/R/ Accessed 2009-07-10, r package version 1.2.4
51. Mevik BH, Wehrens R (2007) The pls package: Principal component and partial least squares regression in r. Journal of Statistical Software 18(2):1 – 24
52. Salzberg S (1997) On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1(3):317–328
53. Köhler M, Machill S, Salzer R, Krafft C (2009) Characterization of lipid extracts from brain tissue and tumors using raman spectroscopy and mass spectrometry. Anal Bioanal Chem 393(5):1513–1520
54. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940–3941, http://bioinformatics.oxfordjournals.org/cgi/reprint/21/20/3940.pdf
55. Kendall C, Stone N, Shepherd N, Geboes K, Warren B, Bennett R, Barr H (2003) Raman spectroscopy, a potential tool for the objective identification and classification of neoplasia in barrett's oesophagus. J Pathol 200(5):602–609
56. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17(3):166–173
57. Carr D, Lewin-Koh N, Maechler M (2010) hexbin: Hexagonal Binning Routines. URL http://CRAN.R-project.org/package=hexbin Accessed 2010-07-08, r package version 1.22.0, ported by Nicholas Lewin-Koh and Martin Maechler
58. Beleites C (2003) Chemometrische Auswertung von IR-Images und -Maps. Master's thesis, Technische Universität Dresden
59. Krafft C, Neudert L, Simat T, Salzer R (2005) Near infrared raman spectra of human brain lipids. Spectrochim Acta A Mol Biomol Spectrosc 61(7):1529–1535
60. Beljebbar A, Amharref N, Lévèques A, Dukic S, Venteo L, Schneider L, Pluot M, Manfait M (2008) Modeling and quantifying biochemical changes in c6 tumor gliomas by fourier transform infrared imaging. Anal Chem 80(22):8406–8415
61. Socrates G (2001) Infrared and Raman Characteristic Group Frequencies, 3rd edn. Wiley
62. Diem M, Griffiths PR, Chalmers JM (eds) (2008) Vibrational Spectroscopy for Medical Diagnosis. Wiley

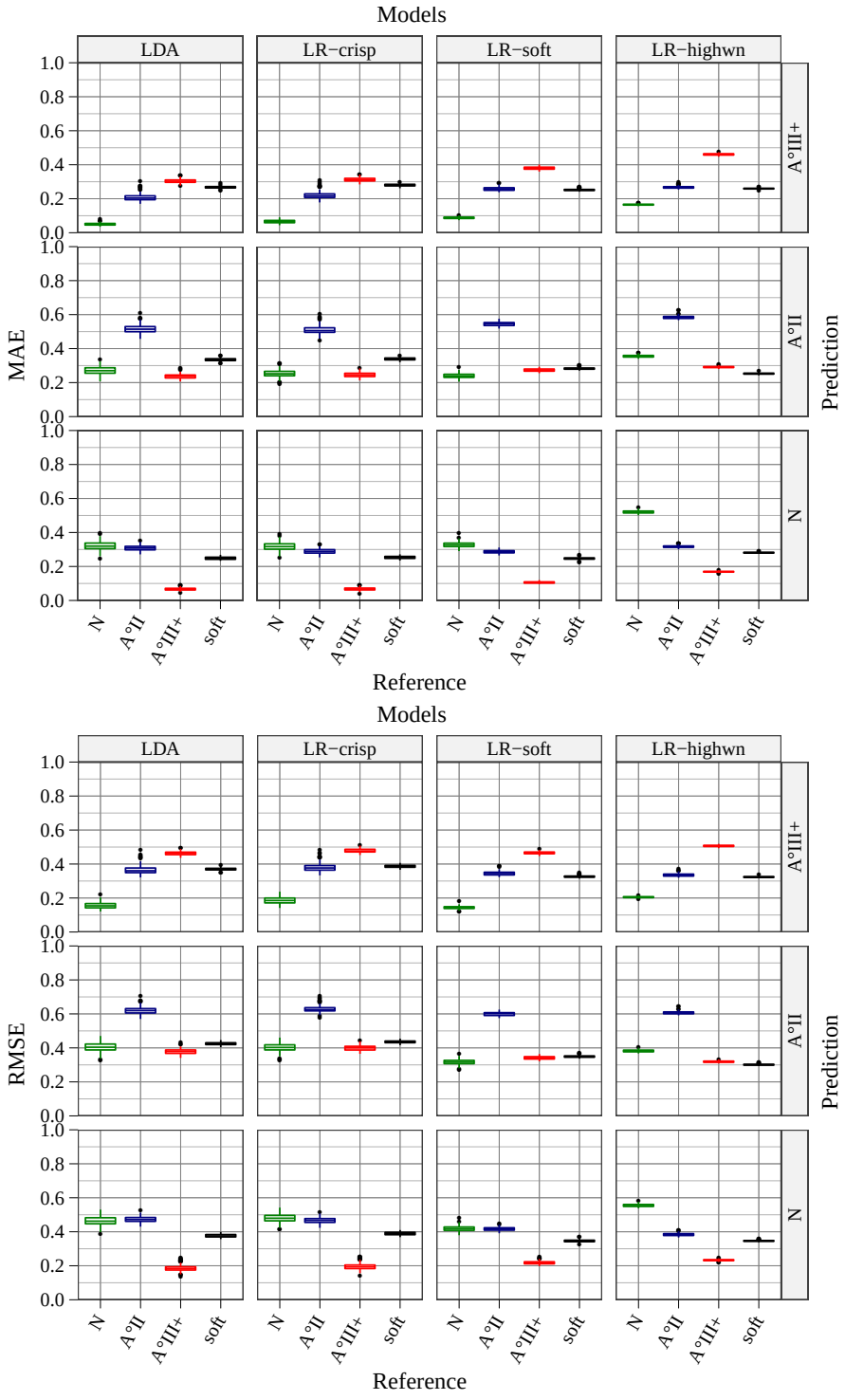
S Supplementary Material

S.1 Model Stability



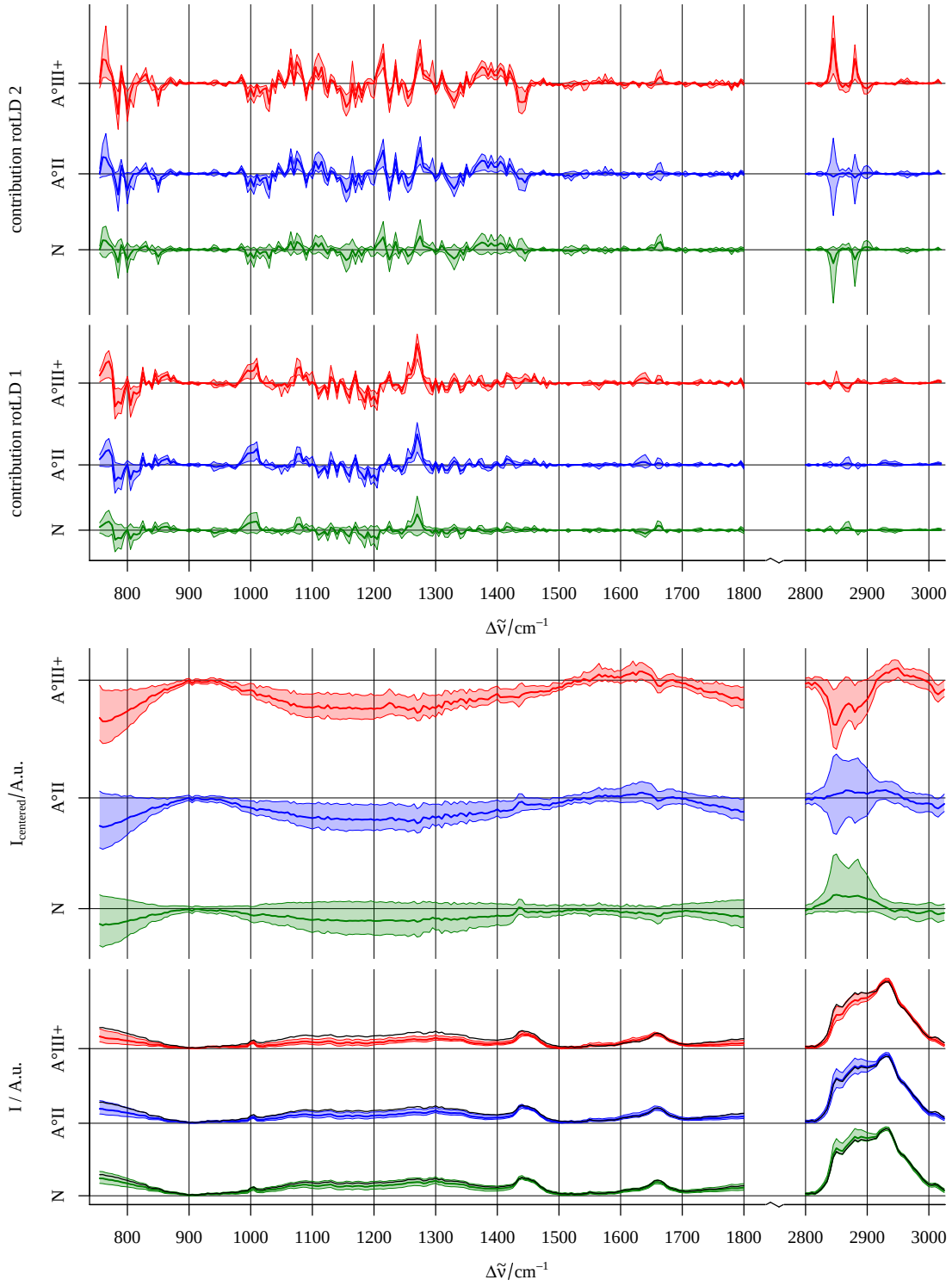
Supplementary Figure S.1: More detailed version of the diagram in fig. 8: the stability of the predictions of the different models (columns) for each predicted class separately (rows). The boxes mark median and quartiles, the whiskers extend to the last value inside $1\frac{1}{2}$ IQR from the box, all further values are marked by points [47].

S.2 *MAE* and *RMSE*



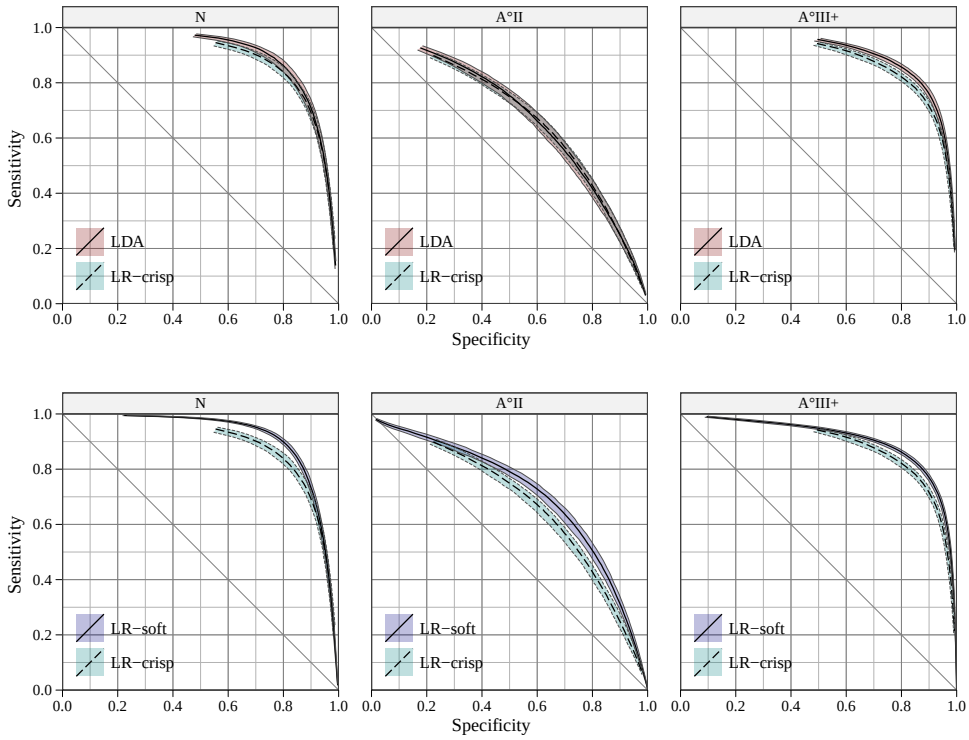
Supplementary Figure S.2: Both *MAE* and *RMSE* have the same pattern across the models, and are very similar for all but the LR-highwn models. Low grade tissues are the most difficult class as they lie in between the normal and high grade tumour tissue: normal tissue is is confused with low grade tumours,but not with high grade tissue, and vice versa. The soft LR models have slightly increased *MAE* for the predicted memberships of crisply labelled spectra, which is counterweighted by a decreased *MAE* of the soft spectra. Their *RMSE* does not show this increase in these cases, but rather an improvement. Also the improvement for soft samples is more pronounced. Together, these findings allow the conclusion that the LR-soft models have more small deviations from the reference, while the LR-crisp and LDA models have larger deviations for fewer spectra. This is in accordance with the the LR-soft modeling smoother class transitions.
The boxes mark median and quartiles, the whiskers extend to the last value inside $1\frac{1}{2}$ IQR from the box, all further values are marked by points [47].

S.3 Spectra and rotated LDA Contributions in Detail



Supplementary Figure S.3: Upper part: detailed Versions of the contributions to the rotated LDA model: median and quartiles, lower part: spectra before (bottom; black: mean normal grey matter spectrum subtracted for "centering") and after "centering" (2nd lowest row).

S.4 Specificity-Sensitivity Diagrams LDA, crisp, and soft LR



Supplementary Figure S.4: Performance of the LDA (continuous) and crisp LR (dashed) models (top row), and crisp and soft LR (lower row), respectively : median, 5^{th} and 95^{th} percentiles observed over the 125 iterations of the cross validation, thresholds 0.01 – 0.99. Both LDA and crisp LR models have virtually the same performance for the low grade tumours, and LDA reaches slightly higher sensitivities for normal tissue. The advantage of the LDA models is most pronounced for the high grade tumours. Soft LR outperforms crisp LR for all classes, the improvement is most pronounced for normal tissue, but most important for A °II.