Bioinformatics, 38(12), 2022, 3238–3244
https://doi.org/10.1093/bioinformatics/btac309
Advance Access Publication Date: 5 May 2022
Original Paper

Systems biology

Targeted metabolomics analyses for brain tumor margin assessment during surgery

Doruk Cakmakci[1,†], Gun Kaynar[2,†], Caroline Bund[3,4,5], Martial Piotto[6], Francois Proust[7], Izzie Jacques Namer[3,4,5] and A. Ercument Cicek [2,8,*]

[1]School of Computer Science, McGill University, Montreal, QC H3A 0E9, Canada, [2]Computer Engineering Department, Bilkent University, Ankara 06800, Turkey, [3]MNMS Platform, University Hospitals of Strasbourg, Strasbourg 67098, France, [4]ICube, University of Strasbourg/CNRS UMR 7357, Strasbourg 67000, France, [5]Department of Nuclear Medicine and Molecular Imaging, ICANS, Strasbourg 67000, France, [6]Bruker Biospin, Wissembourg 67160, France, [7]Department of Neurosurgery, University Hospitals of Strasbourg, Strasbourg 67091, France and [8]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Olga Vitek

Received on December 13, 2021; revised on April 13, 2022; editorial decision on April 27, 2022; accepted on May 2, 2022

Abstract

Motivation: Identification and removal of micro-scale residual tumor tissue during brain tumor surgery are key for survival in glioma patients. For this goal, High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HRMAS NMR) spectroscopy-based assessment of tumor margins during surgery has been an effective method. However, the time required for metabolite quantification and the need for human experts such as a pathologist to be present during surgery are major bottlenecks of this technique. While machine learning techniques that analyze the NMR spectrum in an untargeted manner (i.e. using the full raw signal) have been shown to effectively automate this feedback mechanism, high dimensional and noisy structure of the NMR signal limits the attained performance.

Results: In this study, we show that identifying informative regions in the HRMAS NMR spectrum and using them for tumor margin assessment improves the prediction power. We use the spectra normalized with the ERETIC (electronic reference to access *in vivo* concentrations) method which uses an external reference signal to calibrate the HRMAS NMR spectrum. We train models to predict quantities of metabolites from annotated regions of this spectrum. Using these predictions for tumor margin assessment provides performance improvements up to 4.6% the Area Under the ROC Curve (AUC-ROC) and 2.8% the Area Under the Precision-Recall Curve (AUC-PR). We validate the importance of various tumor biomarkers and identify a novel region between 7.97 ppm and 8.09 ppm as a new candidate for a glioma biomarker.

Availability and implementation: The code is released at https://github.com/ciceklab/targeted_brain_tumor_margin_assessment. The data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.5781769.

Contact: cicek@cs.bilkent.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gliomas are the most prevalent brain tumor type (Yan *et al.*, 2009). They are graded between I and IV according to the histological and clinical criteria established by the World Health Organization (WHO) (Louis *et al.*, 2016). High-grade gliomas (i.e. Grade III and IV) are malignant tumors with poor prognosis and patient survival rate (Sim *et al.*, 2018) as opposed to low-grade gliomas (Lote *et al.*, 1997). Unfortunately, even low-grade gliomas have the potential to progress (Claus *et al.*, 2015). Hence, the surgical management of the tumor is important for the survival of the patient regardless of its grade and pathology.

Even though gross total resection of the tumor attenuates recurrence risk drastically, residual tumor tissue left on the excision cavity constitutes a risk for patient survival. Spectroscopy-based intraoperative feedback mechanisms have been useful in detecting bounds of tumor infiltration which can guide the surgical decision process. Many techniques based on mass spectrometry (Brown *et al.*, 2012; Calligaris *et al.*, 2015a,b; Fatou *et al.*, 2016; Jarmusch *et al.*, 2016; Pirro *et al.*, 2017; Santagata *et al.*, 2014; Schafer *et al.*,

2011) and optical spectrometry (Chan *et al.*, 2018; Colditz and Jeffree, 2012; Hollon *et al.*, 2018; Jermyn *et al.*, 2016, 2017; Li *et al.*, 2014; Lu *et al.*, 2016; Poulon *et al.*, 2017; Stummer *et al.*, 2006; Tsugu *et al.*, 2011; Xue *et al.*, 2018) have been proposed for this goal. A relatively recent technique, High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HRMAS NMR) spectroscopy is a good-fit for use in tumor surgeries because of its ability to analyze small, intact and unprocessed tissue samples in minutes while allowing other analyses to be conducted on the same tissue sample (Gogiashvili *et al.*, 2019). The machine outputs a free induction decay (FID) signal whose frequency domain representation can be analyzed by a technician and a pathologist in ∼20 min to pinpoint the presence of a few biomarker metabolites. This pipeline fits into the time frame of a surgery (Battini *et al.*, 2017).

This feedback mechanism is not bulletproof and can be hindered by problems such as overlapping metabolite peaks in the spectrum which can prevent the expert to decide whether the biomarker is present (Karakaslar *et al.*, 2020). Only peaks for a few metabolites can be checked and quantification of these metabolites is often not possible due to the strict time constraint. Moreover, the system is constrained by the availability and the proficiency of human experts during surgery. A Random Forest (RF)-based approach was proposed to automate this process and pinpoint residual tumor tissue using the raw NMR signal as input (Cakmakci *et al.*, 2020). This is an *untargeted* approach that inputs and analyzes the full raw NMR spectrum. The model was trained on the largest glioma HRMAS NMR dataset available to date which contains close to six hundred samples and achieved over 85% AUC-ROC on distinguishing tumor and healthy samples (i.e. downstream task). Despite the high performance attained by the well-motivated use of the untargeted metabolomics, the study is limited by the inherent noise and high dimensionality of the raw spectrum (over 16k) and the size of the dataset. These limitations prohibit fitting more complex end-to-end models with these dataset sizes (Cakmakci *et al.*, 2020).

In this study, we hypothesized that techniques that can denoise the signal and mimic the targeted approach of an expert technician can boost the performance of the system. In order to test this hypothesis, first, we manually quantify 37 metabolites in the above-mentioned dataset which includes spectra obtained using Carr–Purcell–Meiboom–Gill (CPMG) pulse sequence Cakmakci et al. (2020). HRMAS NMR signatures for these metabolites are obtained from Ruhland *et al.* (2019). Training an RF-based model using manually quantified metabolite concentrations as features has indeed provided an AUC-ROC improvement of 3.8% over the state-of-the-art tumor detection approach which uses the raw spectrum Cakmakci *et al.* (2020). Unfortunately, quantification of this many metabolites is a time-consuming task and does not fit into the surgery time-frame. To overcome this issue, we perform HRMAS NMR metabolite concentration prediction using neural networks by training on actual labeled glioma data (see Fig. 1). Our results show that models have a median of mean absolute percentage errors (MAPEs) ranging from 0.15 to 1.38. Using the predicted metabolite levels for tumor margin assessment provides 4.6% AUC-ROC and 2.8% AUC-PR improvements (median AUC-ROC 88.7%) and can easily fit in the time frame of the surgery as inference only takes seconds.

In addition, we introduce an external, artificial signal into the CPMG spectrum as a scaling agent using ERETIC (electronic reference to access *in vivo* concentrations) method (Akoka *et al.*, 1999). We use and release a new dataset of half a thousand ERETIC-CPMG spectra samples of glioma and control patients (107 of them are new samples) along with quantified metabolite concentrations. Tumor margin assessment with predicted quantities of metabolites we obtain from ERETIC-CPMG spectra has even a better performance (median AUC-ROC 90.2%). We also show that we obtain a similar performance in distinguishing malignant and benign tumors.

Finally, a close investigation of the raw CPMG spectrum via feature importance analysis reveals a very short and novel region between 7.97 ppm and 8.09 ppm which very effectively distinguishes tumor and healthy tissues. While this region has neither been associated with any known metabolite nor used for distinguishing tumor and healthy tissues, using only this region for targeted analyses results in a median AUC-ROC of 88.5%. Our initial analysis using a TOCSY spectrum indicates that the region is associated with the *N–H* group of N-Acetylaspartate but further research is required to establish it as a novel biomarker that can be used for targeted analysis.

2 Materials and methods

2.1 Dataset

Two sample sets were generated from the acquired spectra: (i) CPMG sample set consists of 393 samples which contains 76 control samples, 152 tumor samples (115 aggressive and 37 benign) and 165 samples from the excision cavity (3 aggressive and 1 benign tumors along with 161 control samples); and (ii) ERETIC-CPMG sample set consists of 500 samples which contain 87 control samples, 258 tumor samples (208 aggressive and 50 benign) and 155 samples from the excision cavity (3 aggressive and 1 benign tumors along with 151 control samples). Please see Supplementary Figure S1 for details on sample counts and

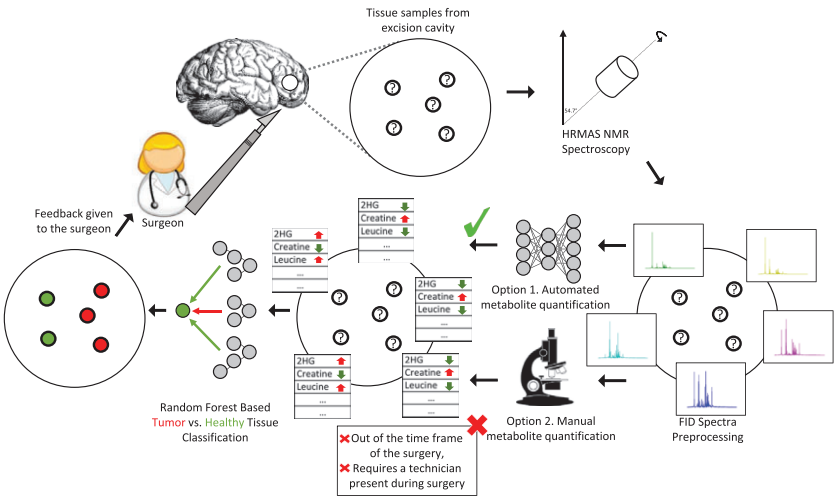

Fig. 1. Proposed pipeline for a feedback mechanism based on predicted metabolite levels during surgery. The surgeon first removes the tumor and then extracts tissue samples from the excision cavity. Samples are analyzed via HRMAS NMR spectroscopy. The resulting spectra are processed by a set of neural networks for metabolite quantification, as opposed to a manual approach which requires a technician to be present and does not fit into the time frame of surgery. Excision cavity samples are labeled as tumor or healthy via a random forest classifier built on the predicted metabolite levels. Guidance on tumor margins is then provided to the surgeon by the predicted labels. The surgeon continues to resect tissue from tumor labeled regions

Supplementary Figure S2 for the age distribution of the samples. Note that the CPMG sample set is a subset of the data released at https://zenodo.org/record/3951448 ($n = 568$) which contains only 393 samples with quantified metabolite concentrations. For more information regarding the dataset including patient's cohort, tissue sample collection, ethics statement, HRMAS NMR spectrum acquisition and normalization and HRMAS NMR Spectrum Preprocessing, please see Sections S1.1, S1.2, S1.3 and S1.7 of the Supplementary Text, respectively. Please see Supplementary Figure S12 in which we took one control and one tumor sample randomly from the dataset and showed the raw ERETIC spectrum on the frequency domain.

2.2 Metabolite quantification

2.2.1 Metabolite quantification and challenges

Manual metabolite profiling is a time-consuming process. The procedure for metabolite profiling requires a separate analysis per spectrum due to changes in the chemical shift of spectrum peaks according to the micro-environment in tissue or cells, particularly its pH (van der Graaf and Heerschap, 1996). A given target metabolite can be profiled from a spectrum in a single execution of the procedure with the assistance of an NMR technician. Moreover, a database of metabolite signatures, to be aligned with the target peaks of the spectrum, must be accessible during the procedure. Frequently, the profile of a single metabolite is not sufficient for a robust analysis. Hence, conducting separate analyses per metabolite is a requirement that increases the complexity of the task drastically. Consequently, application of this procedure during surgery is not feasible and one needs to opt for automated approaches for metabolite quantification.

We consider 37 metabolites for our analyses in this work. A complete list of these metabolites along with the quantification procedure used to measure concentrations of them is provided in Section S1.4 of the Supplementary Text. These are a subset of the metabolite signatures provided by Ruhland *et al.* (2019). The reason for using these 37 metabolites is that for every sample in our dataset all of these metabolites are quantified by an expert and these are deemed important to assess tumor metabolism. Other metabolites can be added to this list if their measurements are available and they are important for the classification task of interest (e.g. for other diseases or cancer types). Metabolite concentrations measured with this procedure were used as ground truth for the automated metabolite quantification methods.

2.2.2 Background on automated techniques for metabolite quantification

A few related methods were recently proposed on automated metabolite quantification problems in the literature. First is the application of RF regression on the raw spectrum which estimates concentrations of N-acetyl-L-aspartic-acid (NAA), Creatine, Choline and myo-Inositol metabolites (Das *et al.*, 2017). Two other works use convolutional neural networks (CNNs) for the same purpose (Hatami *et al.*, 2018; Lee and Kim, 2019). However, their settings and datasets are different from ours. First of all, they all work on brain imaging data (Magnetic Resonance Spectroscopy Imaging) rather than HRMAS NMR technology. Second, all methods heavily depend on simulated data for training where data sizes range from tens of thousands to millions. Only Das *et al.* used 287 samples for training purposes along with the simulated data, and Lee *et al.* use 40 samples obtained from five individuals for testing. As also Lee *et al.* indicate that the simulation procedure has limitations. For instance, in their setting, they leave out several factors such as spectroscopic artifacts like residual water signal, and first-order phase distortion which has the potential to affect the accuracy of the methods. Finally, the number of metabolites quantified is smaller, ranging from 5 to 20.

2.2.3 Problem formulation

We represent each HRMAS NMR spectrum i in the dataset with two variables: a feature vector, $X^{(i)}$, and target metabolite concentrations, $M^{(i)}$. The feature vector is a k-dimensional vector: $X^{(i)} = [X_1^{(i)}, X_2^{(i)}, \ldots, X_k^{(i)}] \in \mathbb{R}^k$ corresponding to either whole or parts of the full spectrum or frequency-binned spectrum. Target metabolite profiles are represented by a d-dimensional vector: $M^{(i)} = [M_1^{(i)}, M_2^{(i)}, \ldots, M_d^{(i)}]$ corresponding to the ground truth concentrations of d metabolites measured from the HRMAS NMR spectrum with the manual metabolite quantification procedure. We formulate the metabolite quantification problem, using the above-mentioned variables, as a set of regression problems such that each regression problem is tackled in an: (i) independent, or (ii) dependent manner. Regardless of the formulation, our goal is to approximate a function f such that $f(X^{(i)}) = M^{(i)}$. For (i), we learn d models for d metabolites where each model approximates a function $f_j(X^{(i)}) = M_j^{(i)}, j \in \{1, 2, \ldots, d\}$. On the other hand, for (ii), we learn a single quantifier model for all metabolites by approximating the function f.

2.2.4 Learning to quantify metabolites from the HRMAS NMR spectrum

Here, we provide the details of the automated metabolite quantification methodology as well as the input modalities used for their development. Note that our approaches are formulated based on the feature vector X with k dimensions and target metabolite concentrations M with d dimensions, as described in Section 2.2.3. In particular, we consider the multivariate multiple elastic net regression trained on the full spectrum as well as fully-connected two-layer perceptrons trained on (i) metabolite peak regions (i.e. target metabolite signatures); and on (ii) frequency-binned spectrum for learning target metabolite concentrations in an independent manner, as formulated in Section 2.2.3. We pick Elastic-net as our baseline as it is a commonly used method that applies regularization to avoid overfitting. Note that in our problem the number of features is larger than the number of training samples.

The multivariate multiple regression is concerned with finding the linear relationship between multiple response variables and predictor variables. Karakaslar *et al.* (2020) used the multivariate multiple regression approach to predict one-dimensional ^{13}C HRMAS NMR signal intensity using also one-dimensional ^{1}H HRMAS NMR spectrum elements as the predictor variables. We have experimented with a multivariate multiple linear regression model without regularization. However, the model was omitted due to poor generalization to unseen samples. Eventually, we used a multivariate multiple elastic-net regression which is a regularized version of the former using a convex combination of L_1 and L_2 priors. Predictor variables for the model (i.e. feature vector X) consist of the full spectrum (i.e. $X^{(i)}$, $k = 16,\ 314$), as defined in Section S1.7 of Supplementary Text, with an extra (1) padded to the beginning (i.e. $X\prime^{(i)}$, $k = 16,\ 315$) for sample i. Response variables of the model are target metabolite concentrations (i.e. $M^{(i)}$, $d = 37$). The Elastic-net objective function we use is obtained from Friedman *et al.* (2010) and is defined as follows:

$$\min_W \frac{1}{2N} \| X\prime^{(1:N)} W - M^{(1:N)} \|_2^2 + \alpha\rho \| W \|_1 + \frac{\alpha(1-\rho)}{2} \| W \|_2^2 \qquad (1)$$

where N is the dataset size, W is the learnable weights of the model, ρ is defined as the constant weighting parameter for the convex combination of L_1 and L_2 priors and α controls the strength of penalization.

Target peak regions, for fully-connected two-layer perceptron models, were selected based on metabolite signatures recorded on this database (Ruhland *et al.*, 2019). We extract these regions from both the frequency-binned spectrum (i.e. $k = 1,\ 401$) and the full spectrum (i.e. $k = 16,\ 314$) by masking relevant ppms. Metabolite signatures used for each metabolite are provided in Supplementary Table S3. We train quantifier models using peak regions from each source separately. That is, d models were trained for each metabolite-specific feature type. Each of the trained models approximates the function f_j, defined in Section 2.2.3, and can be summarized using a dataset of size N as follows:

$$\tilde{X}^{(1:N)} = Mask_j\left(X\prime^{(1:N)}, S_j\right) \quad (2)$$

$$X_{encoding}^{(1:N)} = ReLU\left(FC_j^{(192)}\left(\tilde{X}^{(1:N)}\right)\right) \quad (3)$$

$$\hat{M}_j^{(1:N)} = ReLU\left(FC_j^{(1)}\left(X_{encoding}^{(1:N)}\right)\right) \quad (4)$$

where $FC_j^{(\cdot)}$ represents a fully-connected layer with $(\cdot)$ neurons used only for model j, $ReLU$ stands for the rectified linear unit, S_j represents the signature (peak regions in the HRMAS NMR spectrum) of metabolite j and $Mask_j$ stands for a function to remove elements that do not belong to S_j and concatenate the remaining features. Predicted metabolite concentrations are then a combination of all model outputs:

$$\hat{M}^{(1:N)} = \left[\hat{M}_1^{(1:N)}, \hat{M}_2^{(1:N)}, \ldots, \hat{M}_d^{(1:N)}\right] \quad (5)$$

2.3 Pathological classification

Samples in the dataset are characterized as either healthy (i.e. control group) or tumor (i.e. glioma) tissue. Tumor samples are further characterized as malignant or benign with respect to the labeling of a pathologist. The methods to be proposed in this section are performed for distinguishing: (i) tumor samples from healthy samples; (ii) malignant tumor samples from benign tumor samples. We denote the former as *the main task* due to its importance for brain tumor margin assessment and the latter as the subsidiary task to provide optional and extensive feedback to the surgeon.

2.3.1 Problem formulation

Both tasks are modeled as binary classification problems. Let the class labels for the main and subsidiary tasks, and the feature vector for sample i from the dataset D be $Y_m^{(i)}$, $Y_s^{(i)}$ and $X^{(i)}$ respectively.

The feature vector $X^{(i)}$ is a k-dimensional vector $[X_1^{(i)}, X_2^{(i)}, X_3^{(i)}, \ldots, X_k^{(i)}] \in \mathbb{R}^k$ where $X_j^{(i)}$ correspond to: (i) measured or predicted biomarker metabolite levels (i.e. M_j^i or $\hat{M}_j^i$) (see Section 2.3.4); or (ii) HRMAS NMR spectrum intensity for both tasks (see Sections 2.3.2 and 2.3.3).

The class label for the main task, $Y_m^{(i)}$, is set to 1 if the sample i originated from a glioma tissue and 0 otherwise. Then, given a dataset of size N, the model we learn for the main task approximates a function f such that $f(X^{(1:N)}) = Y_m^{(1:N)}$. Similarly, the class label for the subsidiary task, $Y_s^{(i)}$, is set to 1 if the sample i originated from a malignant glioma and 0 otherwise. Then, the model we learn for the subsidiary task is a function g such that $g(X^{(1:N)}) = Y_s^{(1:N)}$ for a dataset of size N.

2.3.2 Using the raw spectrum as features

In this section, we describe the baseline methods for learning to distinguish pathological labels. Recently, an RF-based method had been shown to detect brain tumor margins more accurately compared to CNNs, fully-connected neural networks, Support Vector Machine (SVM) and Partial Least Squares-Discriminant Analysis (PLSDA) methods (Cakmakci *et al.*, 2020). For this reason and also for fair comparison purposes, we also adopt an RF-based method which is run using different input modalities. We consider both an untargeted approach that works on the raw spectrum and several targeted approaches which use carefully selected regions on the spectrum. First, we consider two untargeted RF-based models for both main and subsidiary tasks: learning from the cropped spectrum (i.e. $k = 8$, 172, only regions with variance) and learning from the full spectrum (i.e. $k = 16$, 314). Note that the former (cropped spectrum) corresponds to the state-of-the-art method proposed by Cakmakci *et al.* (2020). The CPMG spectrum preprocessing routine for the main and subsidiary tasks contains all steps described in Section S1.7 of Supplementary Text with the exception of the spectrum binning step. For the case of ERETIC-CPMG spectrum, a similar preprocessing was applied with the exception of constant factor normalization. Instead of normalizing with a constant factor, each spectrum was normalized with respect to the inherent ERETIC signal located at 10 ppm. For the full spectrum method, we measure feature importance on a left out validation dataset for both main and subsidiary tasks using SHapley Additive exPlanations (SHAP) values (Lundberg *et al.*, 2020). We then construct a targeted RF model on the most important $t \in \{5, 10, 20, 100, 200, 500\}$ spectrum features calculated on a left out validation set. In this case, feature vector $X^{(1:N)}$ $(k = t)$ becomes a discontinuous vector of concatenated t most important features.

2.3.3 Using the uncharacterized region as features

The feature importance analysis (see Section S2 of Supplementary Text) yielded an ignored continuous region from the spectra between 7.97 and 8.09 ppm (i.e. a continuous region of length 142 from the raw spectrum). We train a separate RF model using only this continuous region from the raw spectrum (i.e. $k = 142$).

2.3.4 Using the metabolite quantities as features

We utilize again an RF-based approach on both the main and subsidiary tasks. We train the models using: (i) ground truth metabolite profiles; and (ii) predicted metabolite profiles. The feature vector for the former is a vector of manually quantified biomarker metabolite levels (i.e. $X^{(i)} = M^{(i)}$). The feature vector for the latter is a vector of metabolite levels predicted by the network per metabolite approach (see Section 2.2.4). Additionally, we conduct feature importance analysis on a validation set using either ground truth metabolite levels or predicted metabolite levels. For each feature type, we also construct targeted, RF-based models on the most important $t \in \{1, 3, 5, 10, 20, 30\}$ metabolites to investigate the performance gain.

3 Results

3.1 Experimental setup

In this section, we provide the experimental setup for methods presented in Sections 2.2.4 and 2.3. First, we relabeled tissue samples according to their pathological assessment using the procedure described in Section S1.5 of Supplementary Text. After arranging main and subsidiary task labels, we create two datasets from CPMG and ERETIC-CPMG samples, respectively. Implementation details can be seen in Section S1.6 of Supplementary Text.

3.1.1 Automated metabolite quantification

Regardless of the task and dataset, we adopt a 5-fold cross-validation scheme (repeated three times) for evaluating our methods. Each fold is stratified with respect to the pathological class labels (i.e. control, aggressive and benign). There was not any patient and sample overlap between folds. In each repetition of the framework, we first shuffle the dataset, generate the data folds and perform metabolite quantification analyses. Then, samples contained in each metabolite quantification fold are mapped to main and subsidiary task folds.

During iteration i of the setup, neural network models were tested on fold i, validated on fold $(i + 1) \bmod 5$ and trained on the remaining folds; whereas the multivariate multiple elastic-net regression model was tested on fold i and trained on the remaining folds.

MAPE was calculated on each test fold using measured metabolite concentrations as the ground truth. During cross-validation of automated metabolite quantification methods, we record the predicted concentrations for each test fold. We shuffle the dataset and repeat the cross-validation setup three times for a robust comparison of the methods.

The multivariate multiple regression model was trained using the default values of parameters ($\alpha = 1$, $\rho = 0.5$), which led 2-fold weight for L_1 and 1-fold weight for L_2 losses (after a grid search on $\alpha = \{0.01, 0.1, 1\}$ and $\rho = \{0.1, 0.2, \ldots, 0.9\}$). We experimented with the mean absolute error, mean squared error (MSE), logcosh and huber as loss functions; and on the learning rates between 10^{-2} and 10^{-4} for neural network models. We proceeded with the MSE

loss and learning rate of $10^{-2.1}$ due to the highest overall observed performance. For all models, Adam optimizer (Kingma and Ba, 2014) with a batch size of 64 (selected among 32, 64 and 128) and weight decay of 10^{-5} (selected among 10^{-4}, 10^{-5} and 10^{-6}) were used during training. The maximum epoch limit was 2000 (selected among 1000, 2000 and 3000) and early stopping based on validation loss monitoring was applied on a moving window of 100 (selected among 75, 100 and 125) epochs. We also employ a learning rate reduction on plateau by a factor of 0.2 (selected among 0.1, 0.2 and 0.3) with 50 (selected among 25, 50 and 100) epoch patience period, 25 (selected among 15, 25 and 35) epoch cool-down period with a minimum learning rate of 10^{-4} (selected among 10^{-3}, 10^{-4} and 10^{-5}). We suggest using cross-validation to select these parameters for other applications.

3.1.2 Tumor margin assessment

The 5-fold cross-validation setup used for the metabolite quantification task was also used for main and subsidiary pathology classification tasks. Folds were mapped from the metabolite quantification task and matched with main and subsidiary task labels. During iteration i of the setup, each pathological classification method was tested on fold i, validated on fold $(i+1) \bmod 5$ and trained on the remaining folds. The overall setup was repeated three times.

The training scheme involved a grid search-based hyper-parameter selection on the validation set. Grid search routine was performed on the following hyperparameter space: (i) number of estimators: 50, 150, 300, and 400; (ii) maximum tree depth: 10, 15, 25 and 30; (iii) minimum number of samples required to split an internal node: 5, 10 and 15; (iv) criterion to measure the quality of a split: gini-index and entropy; and (v) the number of samples required to be at a leaf node: 2, 10 and 20. Models were selected based on the AUC-ROC metric calculated on the validation set. As the result of the grid search on our 5-fold cross-validation setup, the following parameters are picked as the best performing hyper-parameter setting for control and tumor classification: Number of estimators = 50, maximum tree depth = 10, minimum number of samples to split a node = 5, minimum number of samples in a leaf node = 20 and the impurity measure = Gini index. For benign and aggressive classification, the following parameters are picked as the best performing hyper-parameter setting: Number of estimators = 300, maximum tree depth = 15, minimum number of samples to split a node = 5, minimum number of samples in a leaf node = 2 and the impurity measure = Gini index. Finally, a model with the decided parameter setting was trained on the full training set and released.

The performance of the models was measured in terms of AUC-ROC and AUC-PR calculated on the test fold. The training and test folds are synced between two tasks. That is, for the samples that are in the test fold for the latter task, we use the metabolite concentration predictions obtained when they are on the test set for the former task. Please see Section S3.1 of Supplementary Text for the experimental setup used for feature importance based models.

3.2 Metabolite quantification

We compare the performances of baseline methods and the proposed neural network architecture (Section 2.2.4) using the MAPE metric. Please see Supplementary Table S3 for a tabular representation of the results. The proposed method (i.e. neural network per metabolite) achieves the lowest median MAPE for 32 and 33 out of 37 metabolites on CPMG and ERETIC-CPMG sample sets, respectively. Please see in Supplementary Figure S11 that the convergence of the training and validation losses for 37 metabolites were shown. We provide the performance comparison of metabolite networks for the proposed approach on the CPMG sample set in Figure 2. The lowest and highest median MAPE achieved were 0.15 and 1.38, respectively. We observe that model could not quantify 2-hydroxyglutarate and acetate as well as other metabolites. On the other hand, creatine, glutamate, glutamine and lactate were the most successfully quantified metabolites in the cohort. Please see Supplementary Figure S3 for the results of the same analysis on ERETIC-CPMG sample set. Overall, we observe that results obtained on each sample set are on par in terms of median MAPE. While our results are not directly comparable with the methods in the literature due to differences in the selection of the performance metrics, the number of metabolites, training setups, and their use of simulated data; we checked the performance of the latest work in Lee and Kim (2019) obtained on simulated spectra and observed that our method performed better in terms of MAPE for alanine and lactate (improvement up to ~0.4 and ~0.24, respectively); worse for some metabolites including glutathione and glutamate (decline up to ~0.26 and ~0.05, respectively). Note that while we would like to perform well on the quantification, it is not the ultimate goal of this study. The ultimate goal is to perform well on the pathology classification tasks using the models trained with these predictions.

3.3 Pathologic classification

We compare the performance of the above-mentioned methods on main and subsidiary pathology classification tasks with respect to AUC-ROC and AUC-PR metrics.

The performance of RF models trained for the main pathology classification task using various feature types obtained from CPMG dataset is given in Figure 3. We observe that the model that uses ground truth metabolite concentrations performed the best on the main task with a median AUC-ROC of 89.4% and AUC-PR of 94.7%. Moreover, the second best model in terms of AUC-ROC is the one built on predicted metabolite concentrations. In this case, the model achieved a median AUC-ROC of 88.7% and AUC-PR of 93.6%. The uncharacterized region also performs similarly. We further observe that the performance gap between the models trained on ground truth and predicted metabolite concentrations are small

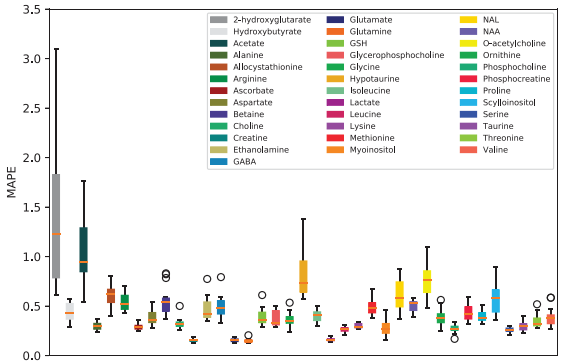

Fig. 2. MAPE of the metabolite concentrations predicted by the proposed method (see Section 2.2.4) on the CPMG sample set. Box plots represent the performance of models obtained on the test folds, in a 5-fold cross-validation setup, which is repeated three times

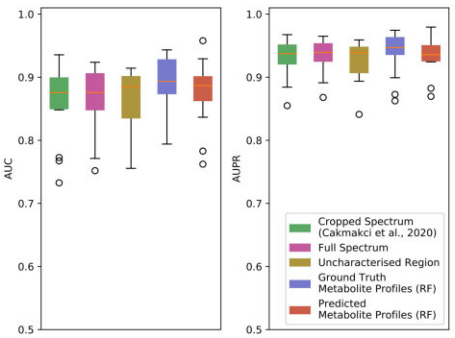

Fig. 3. Performance comparison of models on the task of distinguishing tumor samples from control samples (main task), with respect to AUC-ROC and AUC-PR metrics. Results obtained on CPMG sample set is provided. Please see Supplementary Figure 13 where results obtained on ERETIC-CPMG sample set are provided. Box plots represent the performance obtained on test fold, in a 5-fold cross validation setup, which is repeated 3 times.

(median difference up to 0.7% AUC-ROC and 1.1% AUC-PR). This result shows that predicted metabolite levels provided by the automated metabolite quantification models can be used to distinguish control and tumor samples with high performance.

We observe that distinguishing benign tumor and control samples is more difficult than distinguishing aggressive tumor and control samples. We used the model we trained to classify tumor and healthy samples. Then, we removed all the aggressive tumor samples from the test set. When we classified benign tumors against control samples, we obtained an AUC-ROC of 70% and AUC-PR of 46%, which is lower than the result reported above.

In the case of the raw spectrum-based approaches, the model trained on the cropped spectrum [i.e. the model presented in Cakmakci *et al.* (2020)] had a median AUC-ROC of 87.6% and AUC-PR of 93.8%. Using the full spectrum gives a similar performance which is tied with the model built on the uncharacterized region in terms of AUC-PR but achieved a lower median AUC-ROC. This shows the benefit of our targeted analyses. Note that training with the raw spectrum is also computationally more demanding than the models proposed here—up to 10× running time.

We investigate the benefit of quantifying and using more metabolites in the prediction task and show that adding more metabolites improves the results (Supplementary Figures S8a and S9a). We find that 2-hydroxyglutarate and NAA metabolites are important for tumor and healthy sample distinction (see Supplementary Table S4) which is also supported by the literature (Bulik *et al.*, 2013; Choi *et al.*, 2012). We also show that when using only informative parts of the raw spectrum (discontinuous) also improves the classification task performance (see Supplementary Figure S7a). These results also show that the raw spectrum is noisy and a carefully designed targeted analysis performs better. See Section S3 of Supplementary Text for details of this analysis.

We demonstrate the results of the proposed methods trained on ERETIC-CPMG sample set in Figure S13 of Supplementary Text. The best performing model, which was also trained on ground truth metabolite concentrations, has a median AUC-ROC of 91.2% and AUC-PR of 96.7%. Using predicted metabolites and uncharacterized region resulted in a very similar performance (∼90% AUC-ROC and ∼96% AUC-PR). While overall, using ERETIC-CPMG is better than CPMG, the performance gain is mainly due to the larger dataset size (393 vs 500).

We repeat this analysis for distinguishing benign and aggressive tumors (subsidiary task) and the results are presented in Section S4 of Supplementary Text. As a result, we detected that myo-Inositol metabolite was important for discrimination between aggressive and benign tumors, which is also supported by the literature (Castillo *et al.*, 2000). We also observed that the model constructed on the uncharacterized region performed poorly on the subsidiary task.

3.4 Validation on an independent dataset

We tested the performance of the pipeline on an independent HRMAS NMR dataset obtained from Firdous *et al.* (2021). The raw FID files are preprocessed as we did for our own dataset and obtained the full CPMG spectrum vector of length 16 314. We normalized the signal intensities with respect to the maximum intensity observed in our dataset and we shifted the signal 1515 ppm to the left for calibration and to align with the signal we used during training.

This dataset contains HRMAS NMR plasma samples of 42 individuals. First, we predicted the quantities of 37 metabolites considered for all samples, using the models described in Section 3.2. Out of 42, 26 samples are glioma samples and 16 samples are from healthy controls. Using the best parameter settings for the RF model (see Section 3.1.2), we trained a final model to distinguish tumor and control samples using the full CPMG training set ($n = 393$). We obtained an AUC-ROC of 79% and AUC-PR of 82%.

Out of the 26 glioma samples, 9 are low grade and 17 are high grade. Using the best parameter settings for the RF model (see Section 3.1.2), we trained a final model that uses the full CPMG glioma training set ($n = 156$) to distinguish benign and aggressive tumors. We obtained an AUC-ROC of 72% and AUC-PR of 89%. These results are lower than the performance we obtained on our dataset. Nevertheless, we show that despite the variation in the data due to the cross-site analysis, we were able to obtain arguably high AUC-ROC and AUC-PR values.

4 Discussion

Targeted analysis has the advantage of discarding the noise regions and reducing the dimension of the spectrum over the untargeted analysis. It is also more interpretable as variables of interest can be more easily detected. On the other hand, selecting targeted metabolites and the quantification procedure potentially adds bias to the downstream analyses. Yet, we show that predicting quantities of certain metabolites and using these to distinguish control and tumor samples perform better than the untargeted approach as machine learning techniques can effectively deal with the above-mentioned bias and also fit in the time frame of surgery.

One limitation of the study is the limited size of the training dataset to train end-to-end complex machine learning models. While the dataset we use is the largest of its kind, it prohibits using deep learning models. Hence, we opt for an RF-based model and had to incorporate metabolite quantification task. We foresee that with increasing dataset sizes such problems will attenuate and more complex models will be able to fit directly to the raw signal.

Our metabolite quantification learning attempt is the first for this type of data. Other methods rely mainly on work on image data and simulation to attain higher sample sizes. For these reasons, they are not directly comparable to our approach which works on HRMAS NMR data and is trained on a case control group. There is uncertainty added to the input once the predicted metabolite levels are used for the downstream classification task. However, as our results show, the predicted metabolite levels work on par with using ground truth levels. While the expert decision on ground truth metabolite levels is the best we can obtain, there might be systematic differences that lead to noise in these labels. The added uncertainty in the input that comes with the predictions can help the downstream task to generalize better and avoid over-fitting.

Funding

This work was supported by grants from the BPI France (ExtempoRMN Project), Hôpitaux Universitaires de Strasbourg, Bruker BioSpin, Univ. de Strasbourg and the Centre National de la Recherche Scientifique; also by TUBA GEBIP, Bilim Akademisi BAGEP and TUSEB Research Incentive awards to AEC.

Conflict of Interest: Martial Piotto works for Bruker Biospin.

References

Akoka,S. *et al.* (1999) Concentration measurement by proton NMR using the ERETIC method. *Anal. Chem.*, **71**, 2554–2557.

Battini,S. *et al.* (2017) Metabolomics approaches in pancreatic adenocarcinoma: tumor metabolism profiling predicts clinical outcome of patients. *BMC Med.*, **15**, 56.

Brown,M.V. *et al.* (2012) Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. *Genome Med.*, **4**, 33.

Bulik,M. *et al.* (2013) Potential of MR spectroscopy for assessment of glioma grading. *Clin. Neurol. Neurosurg.*, **115**, 146–153.

Cakmakci,D. *et al.* (2020) Machine learning assisted intraoperative assessment of brain tumor margins using HRMAS NMR spectroscopy. *PLoS Comput Biol.*

Calligaris,D. *et al.* (2015a) MALDI mass spectrometry imaging analysis of pituitary adenomas for near-real-time tumor delineation. *Proc. Natl. Acad. Sci. USA*, **112**, 9978–9983.

Calligaris,D. *et al.* (2015b) Molecular typing of meningiomas by desorption electrospray ionization mass spectrometry imaging for surgical decision--making. *Int. J. Mass Spectrom.*, **377**, 690–698.

Castillo,M. *et al.* (2000) Correlation of myo-inositol levels and grading of cerebral astrocytomas. *AJNR Am. J. Neuroradiol.*, **21**, 1645–1649.

Chan,D.T.M. *et al.* (2018) 5-aminolevulinic acid fluorescence guided resection of malignant glioma: Hong Kong experience. *Asian J. Surg.*, **41**, 467–472.

Choi,C. *et al.* (2012) 2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas. *Nat. Med.*, **18**, 624–629.

Claus,E.B. *et al.* (2015) Survival and low-grade glioma: the emergence of genetic information. *Neurosurg. Focus*, **38**, E6.

Colditz,M.J. and Jeffree,R.L. (2012) Aminolevulinic acid (ALA)-protoporphyrin IX fluorescence guided tumour resection. Part 1: clinical, radiological and pathological studies. *J. Clin. Neurosci.*, **19**, 1471–1474.

Das,D. *et al.* (2017). Quantification of metabolites in magnetic resonance spectroscopic imaging using machine learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 462–470.

Fatou,B. *et al.* (2016) *In vivo* real-time mass spectrometry for guided surgery application. *Sci. Rep.*, **6**, 25919–25914.

Firdous,S. *et al.* (2021) Dysregulated alanine as a potential predictive marker of glioma—an insight from untargeted HRMAS-NMR and machine learning data. *Metabolites*, **11**, 507.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gogiashvili,M. *et al.* (2019) HR-MAS NMR based quantitative metabolomics in breast cancer. *Metabolites*, **9**, 19.

Hatami,N. *et al.* (2018). Magnetic resonance spectroscopy quantification using deep learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 467–475.

Hollon,T.C. *et al.* (2018) Rapid intraoperative diagnosis of pediatric brain tumors using stimulated raman histology. *Cancer Res.*, **78**, 278–289.

Jarmusch,A.K. *et al.* (2016) Lipid and metabolite profiles of human brain tumors by desorption electrospray ionization-MS. *Proc. Natl. Acad. Sci. USA*, **113**, 1486–1491.

Jermyn,M. *et al.* (2016) Raman spectroscopy detects distant invasive brain cancer cells centimeters beyond MRI capability in humans. *Biomed. Opt. Express*, **7**, 5129–5137.

Jermyn,M. *et al.* (2017) Highly accurate detection of cancer *in situ* with intraoperative, label-free, multimodal optical spectroscopy. *Cancer Res.*, **77**, 3942–3950.

Karakaslar,E.O. *et al.* (2020) Predicting carbon spectrum in heteronuclear single quantum coherence spectroscopy for online feedback during surgery. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 719–725.

Kingma,D.P. and Ba,J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee,H.H. and Kim,H. (2019) Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. *Magn. Reson. Med.*, **82**, 33–48.

Li,Y. *et al.* (2014) Intraoperative fluorescence-guided resection of high-grade gliomas: a comparison of the present techniques and evolution of future strategies. *World Neurosurg.*, **82**, 175–185.

Lote,K. *et al.* (1997) Survival, prognostic factors, and therapeutic efficacy in low-grade glioma: a retrospective study in 379 patients. *JCO*, **15**, 3129–3140.

Louis,D.N. *et al.* (2016) The 2016 World Health Organization Classification of tumors of the central nervous system: a summary. *Acta Neuropathol.*, **131**, 803–820.

Lu,F.-K. *et al.* (2016) Label-free neurosurgical pathology with stimulated raman imaging. *Cancer Res.*, **76**, 3451–3462.

Lundberg,S.M. *et al.* (2020) From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–5839.

Pirro,V. *et al.* (2017) Intraoperative assessment of tumor margins during glioma resection by desorption electrospray ionization-mass spectrometry. *Proc. Natl. Acad. Sci. USA*, **114**, 6700–6705.

Poulon,F. *et al.* (2017) Optical properties, spectral, and lifetime measurements of central nervous system tumors in humans. *Sci. Rep.*, **7**, 1–8.

Ruhland,E. *et al.* (2019) A metabolic database for biomedical studies of biopsy specimens by high-resolution magic angle spinning nuclear MR: a qualitative and quantitative tool. *Magn. Reson. Med.*, **82**, 62–83.

Santagata,S. *et al.* (2014) Intraoperative mass spectrometry mapping of an onco-metabolite to guide brain tumor surgery. *Proc. Natl. Acad. Sci. USA*, **111**, 11121–11126.

Schafer,K.-C. *et al.* (2011) Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry. *Anal. Chem.*, **83**, 7729–7735.

Sim,H.-W. *et al.* (2018) Contemporary management of high-grade gliomas. *CNS Oncol.*, **7**, 51–65.

Stummer,W. *et al.* (2006) Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial. *Lancet Oncol.*, **7**, 392–401.

Tsugu,A. *et al.* (2011) Impact of the combination of 5-aminolevulinic acid-induced fluorescence with intraoperative magnetic resonance imaging-guided surgery for glioma. *World Neurosurg.*, **76**, 120–127.

van der Graaf,M and Heerschap,A. (1996). Effect of cation binding on the proton chemical shifts and the spin-spin coupling constant of citrate. *J. Magn. Reson. B*, **112**, 58–62.

Xue,Z. *et al.* (2018) Fluorescein-guided surgery for pediatric brainstem gliomas: preliminary study and technical notes. *J. Neurol. Surg. B*, **79**, S340–S346.

Yan,H. *et al.* (2009) IDH1 and IDH2 mutations in gliomas. *N Engl. J. Med.*, **360**, 765–773.