Received: 20 February 2023 | Revised: 15 June 2023 | Accepted: 10 July 2023

DOI: 10.1002/nbm.5012

RESEARCH ARTICLE

NMR IN BIOMEDICINE WILEY

Deep learning pipeline for quality filtering of MRSI spectra

Mladen Rakić[1,2] | Federico Turco[3] | Guodong Weng[3] | Frederik Maes[2] | Diana M. Sima[1] | Johannes Slotboom[3]

[1]Research and Development, Icometrix, Leuven, Belgium

[2]Department of Electrical Engineering (ESAT), Processing Speech and Images (PSI) and Medical Imaging Research Center, KU Leuven, Leuven, Belgium

[3]Institute for Diagnostic and Interventional Radiology, Support Center for Advanced Neuroimaging (SCAN), University of Bern, Bern, Switzerland

Correspondence
Mladen Rakić, Icometrix, Kolonel Begaultlaan 1b/12, 3012 Leuven, Belgium.
Email: mladen.rakic@icometrix.com

Funding information
H2020 Marie Skłodowska-Curie Actions; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung; Swiss National Science Foundation (SNSF), Grant/Award Number: 182569

Abstract

With the rise of novel 3D magnetic resonance spectroscopy imaging (MRSI) acquisition protocols in clinical practice, which are capable of capturing a large number of spectra from a subject's brain, there is a need for an automated preprocessing pipeline that filters out bad-quality spectra and identifies contaminated but salvageable spectra prior to the metabolite quantification step. This work introduces such a pipeline based on an ensemble of deep-learning classifiers. The dataset consists of 36,338 spectra from one healthy subject and five brain tumor patients, acquired with an EPSI variant, which implemented a novel type of spectral editing named SLOtboom–Weng (SLOW) editing on a 7T MR scanner. The spectra were labeled manually by an expert into four classes of spectral quality as follows: (i) noise, (ii) spectra greatly influenced by lipid-related artifacts (deemed not to contain clinical information), (iii) spectra containing metabolic information slightly contaminated by lipid signals, and (iv) good-quality spectra.

The AI model consists of three pairs of networks, each comprising a convolutional autoencoder and a multilayer perceptron network. In the classification step, the encoding half of the autoencoder is kept as a dimensionality reduction tool, while the fully connected layers are added to its output. Each of the three pairs of networks is trained on different representations of spectra (real, imaginary, or both), aiming at robust decision-making. The final class is assigned via a majority voting scheme.

The F1 scores obtained on the test dataset for the four previously defined classes are 0.96, 0.93, 0.82, and 0.90, respectively. The arguably lower value of 0.82 was reached for the least represented class of spectra mildly influenced by lipids.

Not only does the proposed model minimise the required user interaction, but it also greatly reduces the computation time at the metabolite quantification step (by selecting a subset of spectra worth quantifying) and enforces the display of only clinically relevant information.

Abbreviations: AI, artificial intelligence; CAE, convolutional autoencoder; CSF, cerebrospinal fluid; CPU, central processing unit; EPSI, echo planar spectroscopic imaging; FFT, fast Fourier transform; GPU, graphics processing unit; jMRUI, java magnetic resonance user interface; MIDAS, metabolic imaging data analysis system; MLP, multilayer perceptron; MR, magnetic resonance; MRS, magnetic resonance spectroscopy; MRSI, magnetic resonance spectroscopy imaging; RF, radiofrequency; ROI, region of interest; SELU, scaled exponential linear unit; SLOW, SLOtboom–Weng; SNR, signal-to-noise ratio; TE, echo time; TR, repetition time; UHF, ultrahigh field.

KEYWORDS
convolutional autoencoders, MRSI, quality filtering, spectral quality

1 | INTRODUCTION

As 3D magnetic resonance spectroscopy imaging (MRSI) gains more popularity and presence in the clinical and research practice, with novel acquisition protocols which enable the recording of a large number of spectra from a single subject, there is an increased need for fast, robust, and reliable processing pipelines.[1,2] Such pipelines would include some form of quality assessment of the spectra followed by the quantification of the metabolites, as opposed to the commonly used post-hoc quality-control pipelines.

There is increasing evidence that suggests that MRS image quality has a significant impact on diagnostic outcomes. In a study by Shakir et al,[3] the grading accuracy of low- and high-grade glioma increased by 21% and 15%, respectively, after considering the MRS image-quality assessment. It is known that spectral quality in spectroscopic imaging can be influenced and deteriorated by numerous factors. A review article by Kreis et al[4] provides a systematic overview of various artifacts and quality assessment, as well as detailed guidelines containing quality-check considerations, which are used when evaluating individual datasets/spectra. The artifacts mentioned commonly include (and are not limited to) region of interest (ROI) location and shape,[5,6] movements at scan time,[7–9] chemical shift artifacts,[10] outer volume signal bleed and ghosts,[11] eddy currents,[12] and others. Numerous articles focus on the quality-filtering task in the MRSI domain. Most of these, however, rely on manually crafted and extracted features from the spectra, followed by specific machine-learning approaches for their acceptance/rejection, typical choices being random forest classifiers.[13,14] Alternatively, they depend on particular exclusion criteria based on Cramér–Rao bounds,[15] metabolite peak linewidths,[16] or reliability testing,[17] to name a few. To the best of our knowledge, only one study has addressed the quality-filtering task from the deep-learning perspective, employing convolutional neural networks to detect artifact-contaminated MRSI spectra.[18]

The main requirements of a quality-assessment pipeline, in this case, are its robustness and reliability (i.e., the quality checks should be as accurate as possible) as well as its speed (i.e., the computational time should be as short as possible). In this work, we present a pipeline, based on deep-learning classifiers, for the classification of the spectra according to their spectral content and quality. Namely, the architecture we propose is composed of an ensemble of paired networks, each pair consisting of a convolutional autoencoder (CAE) for feature extraction[19] and a multilayer perceptron (MLP)[20] added to the encoding path of the autoencoder for classification. Autoencoding networks are proven to be useful for unsupervised dimensionality reduction and feature extraction in numerous applications.[19,21–25] The ensemble includes three pairs of the networks mentioned, trained on either the real or imaginary parts of the spectra or both together. A simple voting scheme is then employed to assign the final label to each spectrum. An underlying hypothesis is that the ensemble approaches add value to the model's output regarding its accuracy and robustness.[26–28] The spectra are analysed in the frequency domain. To this end, we define four spectral quality classes, including noise, two classes containing either major or minor lipid-related artifact components, and good-quality spectra that are directly quantifiable without applying additional processing steps. Further, we will explore the binary class consideration instead of the proposed four-class approach by classifying spectra as either quantifiable or not, regardless of additional processing that may be needed.

With the number of spectra that can be acquired always increasing,[29–31] it is becoming more and more relevant to filter out the spectra deemed unquantifiable, such that the quantification step[2] is optimised and only relevant data are presented to the user. Our proposed method speeds up this process significantly by automatically classifying a large number of spectra in a short time. The model learns directly from the expert's considerations used at the labeling step, which intrinsically reflect the guidelines mentioned above. A quality-filtering tool such as the one we propose in this article can also be employed to mitigate misdiagnosis errors to a certain degree.

2 | MATERIALS AND METHODS

2.1 | Spectral quality classes

For the purpose of the automatic classification and filtering of brain MRSI spectra, we propose four different spectral quality classes, defined as follows.

1. Noise: these are primarily spectra stemming from the extracranial region, as well as any intracranial spectra that carry no metabolic information due to either their localisation (e.g., nasal cavity) or inadequate acquisition.
2. Spectra greatly influenced by lipid-related artifacts: these are most commonly spectra coming from the skull and subcutaneous regions, characterised by broad components high in intensity, overpowering and overlapping the metabolic signals and thus rendering the quantification useless.

3. Spectra mildly influenced by lipid components: these are normal-appearing spectra coming from the brain region with a noticeable lipid component, which can be removed using certain postprocessing steps to make them suitable for quantification. Lipid components are typically extracted using lipid reduction algorithms, namely lipid *k*-space interpolation[32] or constrained reconstruction.[1,33–35]
4. Good-quality spectra: these are spectra suitable for quantification without applying additional processing steps when using state-of-the-art metabolite quantification algorithms.

Typical representations of each of the four defined classes of spectra are shown in Figure 1.

While the spectral classification is presented and analysed in this four-class context in most of the experiments presented in this work, it is good to note that we can think of the problem in a more straightforward, binary way. It treats the first two classes as spectra that are not worth quantifying. On the other hand, the last two classes are those spectra worth putting through the (rather time-consuming) quantification step, with or without application of specific restorative preprocessing steps beforehand. This binary perspective will be referred back to when discussing the results of the proposed model.

2.2 | Dataset

The brain MRSI data used in this research were collected with an EPSI variant, which implements a novel type of spectral editing named SLOtboom–Weng (SLOW) editing[29] on a UHF 7T MR scanner (Terra, Siemens Healthineers, Erlangen, Germany), using the CE-labeled clinical mode. Whole-brain spectral editing can be obtained within 10 minutes of acquisition time (TE = 68 ms, TR = 1500 ms). Due to the implicit water suppression of SLOW editing, no additional water removal postprocessing step is necessary. The raw SLOW EPSI data were processed with the metabolic imaging data analysis system (MIDAS),[36] including the following steps: (i) EPSI *k*-space regridding, (ii) echo combination and drift correction, (iii) spatial fast Fourier transform (FFT), (iv) multichannel acquisition combination, and (v) eddy-current correction.

The dataset used in the initial development stages to assess intrareader variability consisted of 26,460 spectra from a single healthy subject, manually labeled twice by an expert rater (JS). The labeling here refers to assigning one of the four class labels to each spectrum in the frequency domain, as defined in the previous section. It was performed by evaluating the spectra using the built-in labeling tool available in the SpectrIm plugin[37] of the java magnetic resonance user interface (jMRUI) software.[38] Double labeling was performed to improve certainty and reduce the intrarater variability. A total of 22,878 spectra were labeled the same in both iterations and kept for training, validating, and testing the model. We conducted a simple intrarater variability experiment, comparing the labels in two labeling iterations. It serves as an interesting potential benchmark for our model.

Further, an extended dataset, which includes an additional 13,460 spectra from five tumor patients, was labeled to account for class imbalance and data variability. It is also used to test the robustness and generalisation of the proposed model in the cross-validation experiments. These

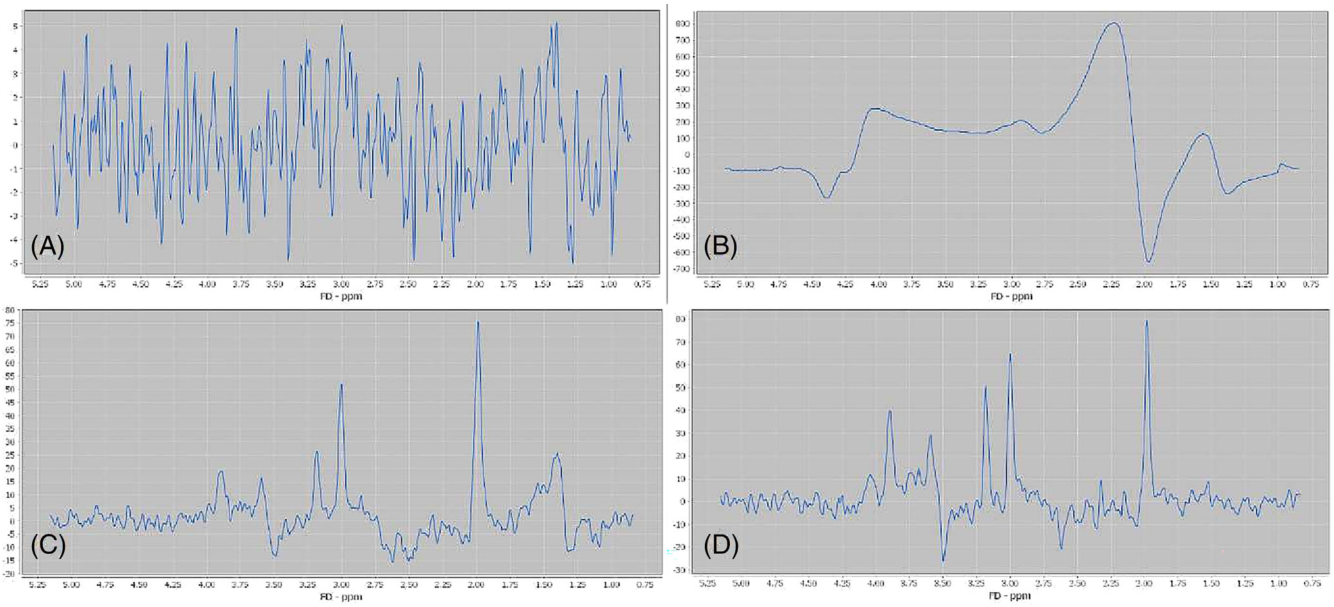


FIGURE 1 Typical examples of the four quality classes defined, showing the real parts of the corresponding spectra: (A) noise, (B) spectra greatly influenced by lipid-related artifacts, (C) spectra mildly influenced by lipid-related artifacts, (D) good-quality spectra. Note that the good-quality spectrum has phasing distortions, but, as long as the metabolite content is clearly distinguishable, the proposed method does not require zero-phased spectra.

spectra originate from one axial slice per patient, which contains the largest region of interest (i.e., a cranial region with or without tumor, as opposed to axial slices with a greater number of extracranial voxels containing noise).

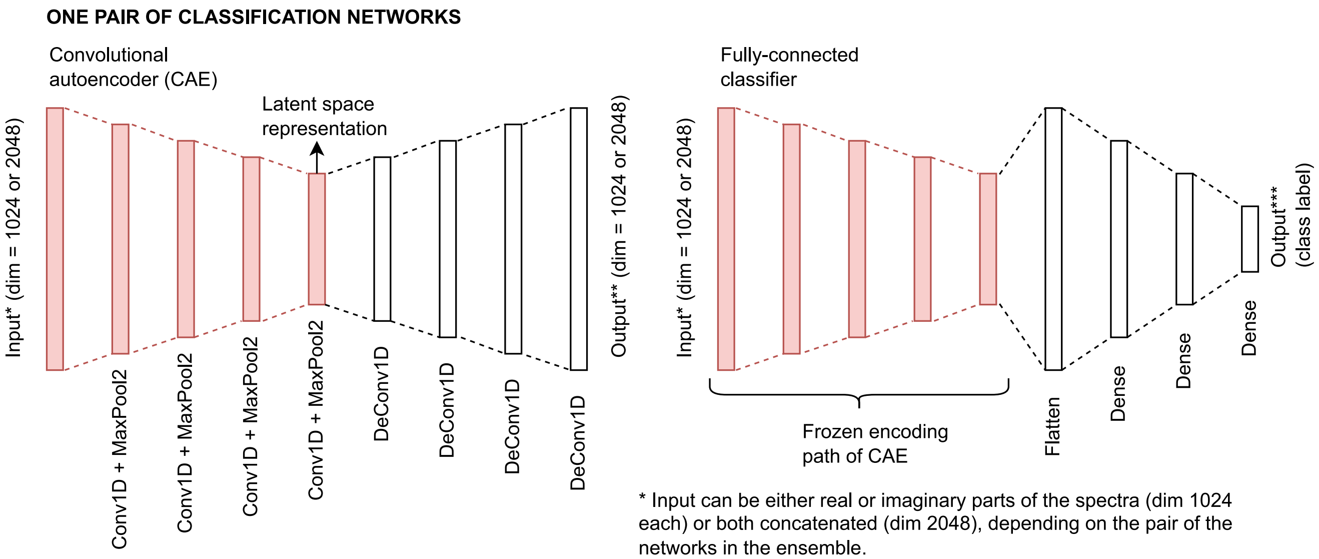
Additionally, we demonstrate the qualitative aspects of the developed model on an unlabeled dataset originating from a second healthy subject. Therefore, this dataset is not used in the training phase but only to assess the model's capabilities when evaluated on unseen data.

2.3 | Model architecture

The spectra are classified using a deep-learning ensemble model of three pairs of networks, illustrated in Figure 2. Each pair consists of two modules: a CAE for feature extraction and an MLP classifier network for labeling the spectra. Each pair of networks is trained on different aspects of the spectra in the frequency domain, namely (i) real, (ii) imaginary, and (iii) both real and imaginary concatenated parts of the spectra. The principal reason behind the introduction of the three pairs of networks is to establish a voting scheme that would help improve the robustness and margin of error in the classification model, while at the same time not increasing the computational resources/time significantly. It should be noted that the spectra were normalised to the unit norm (using the *l*2 normalisation) prior to passing them to the network in order to increase the robustness to outliers and to avoid difficulties in training due to the high values typically present in spectra influenced greatly by lipids.

CAEs are a popular type of neural networks used for compression and feature extraction tasks. An autoencoder, by definition, tries to reconstruct the signal presented at its input as its output by inferring essential features in the latent space. A CAE comprises an encoding path, which compresses the input signal through a series of convolutional and max pooling layers into a lower-dimensional representation of the extracted features of importance, followed by a decoding path, which, again through a chain of transposed convolutional layers,[39] tries to reconstruct the input from its latent space representation.

A classifier network in the form of an MLP is then trained in a supervised manner to learn to differentiate the four defined classes of spectra. It takes the input signals and their corresponding labels and tries to update the layer weights in order to minimise the loss function. Instead of feeding the entire spectra to the classification network directly, our approach presents their latent space versions learned by the appropriate CAE as inputs to the classifier. That way, the features of interest that are distinguishable among the classes are preserved in the input signal, while the noise and redundancy are minimised. In other words, in the classification step, we "freeze" the encoding path of the already trained CAEs, and replace the decoding path with the trainable classifying layers. Note that the CAEs and the MLPs are trained separately. This is because we


ONE PAIR OF CLASSIFICATION NETWORKS
Convolutional autoencoder (CAE)
Latent space representation
Input* (dim = 1024 or 2048)
Conv1D + MaxPool2
Conv1D + MaxPool2
Conv1D + MaxPool2
Conv1D + MaxPool2
DeConv1D
DeConv1D
DeConv1D
DeConv1D
Output** (dim = 1024 or 2048)
Fully-connected classifier
Input* (dim = 1024 or 2048)
Frozen encoding path of CAE
Flatten
Dense
Dense
Dense
Output*** (class label)

* Input can be either real or imaginary parts of the spectra (dim 1024 each) or both concatenated (dim 2048), depending on the pair of the networks in the ensemble.
** Output is always a reconstructed input spectrum of matching dimensionality.
*** Class label can be 1, 2, 3 or 4 and the labels from the three pairs of the networks are subjected to a majority voting scheme to determine the final label.

FIGURE 2 A visual representation of a pair of classification networks. Three pairs of networks process either real, imaginary, or both spectral parts concatenated, respectively. CAEs serve as feature extractors and a dimensionality reduction tool, while the fully connected classifiers label the spectra by training on their latent space representations obtained from autoencoders. (Conv1D, one-dimensional convolution layer; MaxPool2, max pooling layer of size 2; DeConv1D, one-dimensional deconvolution, i.e., transposed convolution layer).

wanted to avoid introducing potential bias toward the ground truth. Specifically, by training the networks simultaneously, the feature extraction learned by the autoencoders would be biased toward the ground-truth labels, as the joint training would not be unsupervised.

At the validation and testing stages, the corresponding spectra are given to the three pairs of feature extraction/classification modules, and a simple majority voting scheme is applied to assign the final labels. In the case of a tie when counting the votes from individual classifiers, the highest assigned label is kept as final, since we deem it more important not to miss a potentially good-quality spectrum among the noisy ones than the other way around.

The network is implemented using the Keras open-source software library, which provides a Python interface for artificial neural networks.[40] The CAEs are trained by monitoring and minimising the mean-squared error loss function. The input vectors to the CAEs are of size either 1024 or 2048, depending on whether the real and imaginary parts of the spectra are processed independently or in a concatenated manner. Each CAE comprises four encoding layers containing 32, 64, 64, and 128 convolutional kernels (Conv1D layers, kernel size 2), each followed by a max pooling layer (MaxPooling1D) of size 2. Note that we tested different kernel sizes (2, 3, and 4) and selected the one that yielded the best performance. A smaller receptive field seems more favorable in our task, where local characteristics can differentiate among different spectral quality classes better than global ones. Scaled exponential linear unit (SELU) activation is also applied in each convolutional layer.[41] It includes self-normalising properties and has no problem with vanishing gradients. The decoding path also includes four analogously defined transposed convolutional layers (Conv1DTranspose), each comprising 128, 64, 64, and 32 kernels, respectively. On the other hand, the classifier networks try to minimise the categorical cross-entropy, a typical loss function present in multiclass classification approaches. Three dense layers with 256, 64, and 4 nodes, respectively, are trained using stochastic gradient descent as the optimiser. SELU activation is also incorporated in the classification modules of the model, except after the last dense layer, which is followed by a softmax activation.

2.4 | Experiment design

This section provides an overview of the experiments conducted, the results of which are discussed later. As a brief reminder, the dataset consists of one healthy subject, labeled twice by the expert, and five tumor subjects, partially labeled based on a single slice.

Firstly, we analyse the intrarater variability by comparing the labels assigned by the expert during two labeling iterations of the healthy subject. The spectra assigned the same label both times are kept in the dataset, and the rest are discarded.

Secondly, we analyse the performance of a model using only the spectra retained from the healthy subject. This was done by randomly splitting the healthy subject's dataset into 64% training, 16% validation, and 20% testing spectra.

Thirdly, we observe a model's performance when using a more diverse dataset. Here we combine the healthy subject's spectra with the five partially labeled tumor subjects. A model is trained similarly to the previous experiment by randomly splitting all the spectra into 64% training, 16% validation, and 20% testing spectra. It is important to note that this is the most complete and the largest of the datasets we used in this article. If an experiment does not explicitly state otherwise, it is understood that the dataset composition used in it is the same as in this one.

Additionally, cross-validation experiments are conducted to evaluate the performance of a model on unseen data, reflecting its robustness and generalisation abilities. We propose fivefold cross-validation by splitting the dataset five times into training, validation, and testing sets. In each fold, we keep the healthy subject and four tumor subjects for training and validation, while the remaining patient is used as a testing dataset. This testing subject is a different one in each of the five folds.

Further, we want to investigate the added value of using the ensemble of networks. To do this, we compare the performance of a model when using only one pair of networks with the performance when using all three pairs.

Moreover, we show the importance of having a well-labeled dataset. We do this by comparing the model's performance on the healthy subject after only one labeling iteration by the expert with the performance of the model trained on retained data after two labeling iterations.

Lastly, we present qualitative results when evaluating the model (trained on the labeled healthy subject and the five partially labeled tumor subjects) on an unseen, unlabeled, healthy subject. The qualitative analysis is done to observe the model's generalisation capabilities and robustness when applied to independent datasets.

2.5 | Performance metrics

The model's performance is evaluated by comparing its output with the expert's labels on the test set (which comprises 20% of the total spectra in the dataset) or on a single held-out patient dataset in the case of the cross-validation experiment. Quantitative results are demonstrated as confusion matrices, showing the number of spectra labeled a certain way by the model against the true labels assigned by the expert. The matrices allow for the computation of F1 scores, indicative of the number of spectra classified correctly and, conversely, the number of misclassified spectra.

The F1 score measures the test's accuracy and is commonly computed by applying the following formula for each of the four classes defined:

$$F1=\frac{2TP}{2TP+FN+FP}. \tag{1}$$

In the formula above, *TP*, *FN*, and *FP* signify counts of spectra labeled as true positives, false negatives, and false positives, respectively, from the perspective of the observed class of spectra. For example, considering the confusion matrix, the *TP* count for the specific class is the value found on the main diagonal in the row/column of that class. In contrast, *FN* and *FP* are all the values found in the row/column for that class, but outside the main diagonal.

3 | RESULTS

Table 1 contains a confusion matrix providing deeper insight into the intrarater variability for two labeling iterations for the spectra from a single healthy subject. It also shows interesting information pertaining to class representation in the dataset. We can observe a relative imbalance, evident by the fact that almost half of the spectra in the dataset are noise. In contrast, about 14.9% of the spectra are of good quality, and only about 7.7% of the dataset contains spectra mildly influenced by lipid components. However, the addition of another 13,460 labeled spectra from five patients with brain tumor changed the class representation.

TABLE 1 Intrarater variability confusion matrix, showing the number of spectra that were labeled a certain way during two labeling iterations. In an ideal scenario with an expert rater with no doubts as to how to label spectra, all values greater than zero would be found on the main diagonal of the matrix. F1 scores for each of the four classes are given, as a way to quantify consistency.

		Iteration 2				
	Classes	**1**	**2**	**3**	**4**	**F1 score**
Iteration 1	1	10,845	264	1	0	**0.92**
	2	1744	6862	419	10	**0.84**
	3	4	231	1753	468	**0.69**
	4	0	1	440	3418	**0.88**

TABLE 2 Confusion matrix and the corresponding class F1 scores coming from the model trained and validated on a single, twice labelled healthy subject. The values are computed on the test set, comprising 20% of the total number of spectra.

		Model output: one subject				
	Classes	**1**	**2**	**3**	**4**	**F1 score**
Expert labels	1	2116	32	0	0	**0.98**
	2	67	1312	24	0	**0.94**
	3	0	40	265	42	**0.81**
	4	0	0	17	661	**0.96**

TABLE 3 Confusion matrix and the corresponding class F1 scores coming from the model trained and validated on all the available labeled data, comprising one healthy and five tumor subjects. The values are computed on the test set, comprising 20% of the total number of spectra.

		Model output: all subjects				
	Classes	**1**	**2**	**3**	**4**	**F1 score**
Expert labels	1	2705	153	3	0	**0.96**
	2	95	2435	30	0	**0.93**
	3	3	71	701	117	**0.82**
	4	0	0	81	874	**0.90**

The initial results were obtained via training, validating, and testing the model on the dataset coming only from the healthy subject (more specifically on the subset of the spectra labeled the same in both labeling iterations) by splitting the spectra randomly into 64% training, 16% validation, and 20% testing cases. The confusion matrix after training the three classifiers and applying the majority voting to assign the final labels is shown in Table 2.

Additionally, a model was trained and evaluated using the entire dataset comprising all the retained spectra from the healthy subject and five partially labeled tumor patients. Again, the entire dataset was split randomly into 64% training, 16% validation, and 20% testing spectra. The corresponding confusion matrix and the F1 scores obtained are shown in Table 3. Note that the performance stays relatively similar between Tables 2 and 3.

When it comes to cross-validation experiments, to simplify the outcome and avoid showing five individual confusion matrices from each of the folds, a matrix obtained by summing up the five is shown in Table 4, together with the corresponding F1 class scores. A crucial performance remark of the model is that it does not discard spectra coming from tumor regions during the quality control. To this end, we show in Figure 3 qualitative results when evaluating the model in two of the cross-validation folds on the single left-out subject for testing purposes. These examples demonstrate that the model deems tumor spectra quantifiable (class 3 or 4) and show how they compare with the spectra from brain tissue unaffected by a tumor. The testing subjects in the remaining folds have significant differences in MRSI grid and brain orientation, which do not allow for nice visual single-slice representation, though the tumor prediction remains unchanged.

If we consider spectra given to the network as inputs, there is arguably redundancy that should be remarked on. While two of the classifying modules take real and imaginary parts of the spectra, the third one takes both concatenated together. It is hence essential to demonstrate whether there is any improvement when using an ensemble of classifiers. Table 5 compares the F1 scores obtained in the experiments analysed

TABLE 4 Confusion matrix and the corresponding class F1 scores coming from the fivefold cross-validation experiments, obtained by summing up the five individual matrices.

		Model output: cross-val				
	Classes	**1**	**2**	**3**	**4**	**F1 score**
Expert labels	1	3241	531	20	5	**0.90**
	2	195	5451	72	1	**0.92**
	3	3	148	2124	364	**0.81**
	4	1	2	409	893	**0.70**

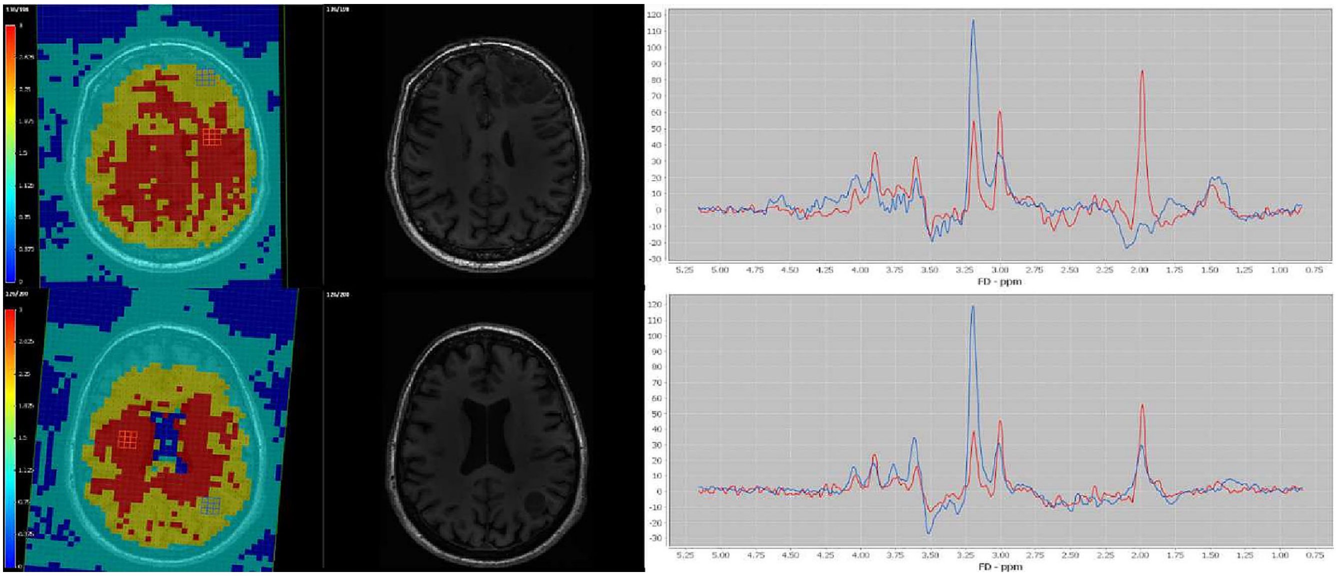


FIGURE 3 Example results from two folds of the cross-validation experiments. Each row corresponds to a single left-out tumor patient used for testing in each of the folds. Images in the left column show a map overlay of the labels produced by the model (class 1: dark blue, class 2: light blue, class 3: yellow, class 4: red). The middle column includes appropriate anatomical images of the same axial slices. Spectra in the right column are obtained by averaging the blue and red marked voxels in the leftmost images. Blue selection relates to the tissue affected by the tumor, while the voxels selected in red are a typical representation of healthy-appearing tissue, classified as good-quality spectra by the classifier network.

TABLE 5 Comparison of the F1 scores across different experiments when using three pairs of classifying modules, or just one that is trained on both real and imaginary parts of the spectra. Values listed in the cross-validation columns are the mean values across five folds of the corresponding experiment.

		One subject One pair	One subject Three pairs	All subjects One pair	All subjects Three pairs	Cross-val One pair	Cross-val Three pairs
F1 scores	Class 1	0.98	0.98	0.96	0.96	0.89	**0.90**
	Class 2	0.94	0.94	0.93	0.93	0.92	0.92
	Class 3	0.81	0.81	0.81	**0.82**	0.80	**0.81**
	Class 4	0.96	0.96	0.89	**0.90**	0.67	**0.70**

TABLE 6 Confusion matrix and the corresponding F1 scores obtained from the model trained on the healthy subject's dataset after the first manual labeling iteration. The values are computed on the test set, comprising 20% of the total number of spectra.

		Model output: one subject, first labeling iteration				
	Classes	1	2	3	4	F1 score
Expert labels	1	2456	63	0	3	**0.96**
	2	146	1307	29	0	**0.89**
	3	0	87	339	93	**0.71**
	4	0	0	69	700	**0.89**

above, which use majority voting of all three networks, with the F1 scores coming from the corresponding experiments, which use only the third pair of the deep-learning modules for classification (the one trained on concatenated real and imaginary spectra). The best-performing cases are highlighted appropriately in the table.

Referring to the intrarater assessment, we want to note the importance of a well-labeled dataset. As mentioned previously, an expert manually labeled one healthy subject in the dataset twice. While Table 2 shows the results of the model trained on a subset of spectra labeled the same in both labeling iterations, it is interesting to showcase the difference when training on all spectra of the healthy subject after only the first labeling iteration. Table 6 demonstrates these values and indicates a decreased agreement compared with the results shown in Table 2.

Further, regarding the class mixing, we observe that, in all of the confusion matrices, most of the values outside the main diagonals tend to be concentrated in the matrix fields corresponding to adjacent classes. This is true for both the model outputs and the intrarater variability analysis. Going back to the binary consideration of the classes in place of the four-class approach introduced earlier, we can indeed think of classes 1 and 2 as "unquantifiable" and classes 3 and 4 as "quantifiable", albeit with or without certain additional processing steps. Repurposing the confusion matrices for this binary perspective (by considering each of the four two-by-two quadrants of a matrix as a single cell), we can obtain some remarkable F1 scores, presented in Table 7 for easy comparison.

Moreover, a qualitative assessment is shown in Figure 4. It shows the same examples as in Figure 3, but, while the color maps on the left visualise the model output, the color maps in the middle column represent the ground truth, as labeled by the expert. The color maps on the right are the absolute voxelwise difference of the previous two. Finally, Figure 5 shows the performance of the model on an unseen, unlabeled, healthy subject. It serves as an additional qualitative assessment of the model performance when introducing new data from a subject that was not part of training/validation datasets.

4 | DISCUSSION

Before the in-depth analysis of the observed results, we want to note an essential aspect of the proposed method. By employing this deep-learning approach, there is no need to distinguish specific artifacts (e.g., noise or lipid contamination) except during the labeling step of the training set. The model learns from an expert and produces quality-control classification mimicking the expert's knowledge.

While it is true that the F1 scores obtained from the model trained and tested on a single healthy subject are the highest among the experiments presented, the interpretation of the scores observed in the other experiments should not be diminished. A model trained and evaluated on the same data type is bound to be biased toward the data seen during the training.[42,43] It is also noteworthy that the F1 score for class 3 in that experiment was significantly lower than those of the other three classes, but the dataset arguably suffers from class imbalance.

TABLE 7 F1 scores comparing different experiments, inferred using the binary approach, which considers classes 1 and 2 as unquantifiable and classes 3 and 4 as quantifiable.

		Intrarater	One subject	All subjects	Cross-validation
Binary	Classes 1+2	0.98	0.99	0.99	0.99
F1 scores	Classes 3+4	0.95	0.97	0.97	0.97

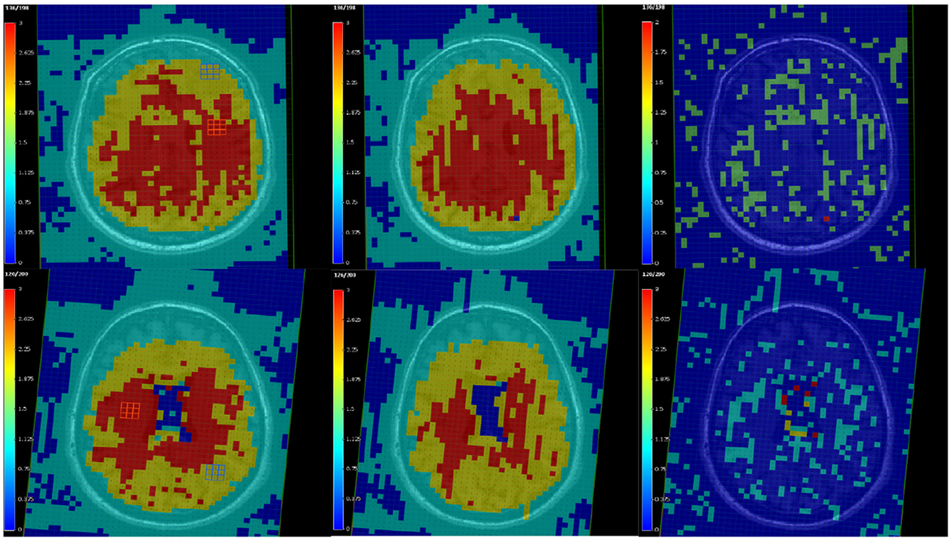


FIGURE 4 A qualitative comparison of model output (left) and ground truth labeled by the expert (middle) shows that most differences tend to cluster at the boundaries of regions with particular predominant classification. Absolute voxelwise difference of the corresponding left and middle color maps is shown on the right.

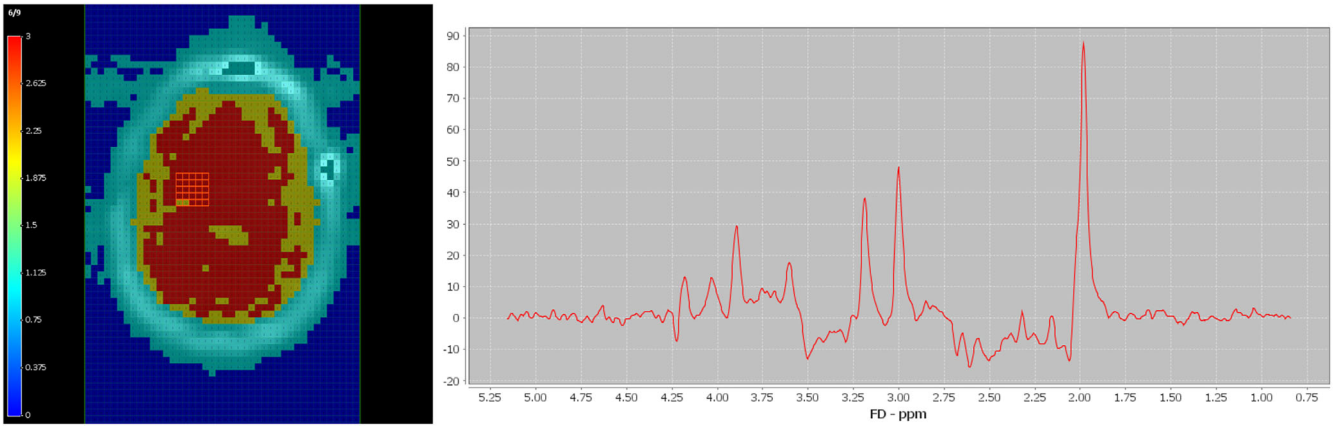


FIGURE 5 The quality filtering model was applied on an unseen dataset coming from a healthy subject, and the spatial representation matches what is expected (left). The average spectrum of the selected voxels is shown on the right.

Though slightly lower, the F1 scores from the model trained and tested on six subjects are still comparable with the ones obtained in the original experiment. They demonstrate the generalisation and robustness capabilities of the model when introducing more diverse data.

If we look further into the confusion matrices presented, some additional conclusions can be inferred. While the values on the main diagonals represent true positives (i.e., the cases where the model output matched the expert's labels), we dub all other values "class mixing". This class mixing is arguably the result of two possible scenarios. The first one relates to a simple misclassification of the model when labeling spectra with labels that do not match the ones from the expert. The second one, alternatively, is associated with potential mislabeling by the expert. We should consider the fact that all supervised models "suffer" from their inherent reliability on what is considered to be the ground truth. Referring back to

Table 1, we can see that the F1 scores calculated from two independent labeling iterations are lower than the ones the models yielded for certain classes.

If we consider the scores obtained in the cross-validation experiments, we can notice the drop in performance in class 4. This drop happens in two of the five folds of the cross-validation. However, it should be taken into consideration that the model classifies the spectra coming from completely unseen data during its training phase. This can mean that some interpatient variability is more difficult for the model to capture. However, class mixing also happens mostly between adjacent classes, and the binary class performance is comparable with other results, as shown in Table 7. This means that even the interpatient variability, which caused the drop in the performance in class 4, does not result in the model discarding the useful spectra as unquantifiable. For the qualitative assessment and the demonstration of the method in practice, we show in Figure 5 the output of the model (trained on the intersection of the twice labeled dataset) on an unlabeled healthy subject that has not been used during the training steps. In the example, dark blue voxels correspond to the noisy class 1 spectra, while the light blue, yellow, and red voxels relate to classes 2, 3, and 4, respectively. The image shows an axial slice from the middle region of the cerebrum at the approximate height of the slice directly above the lateral ventricles. Some of the voxels are highlighted, and the mean spectrum from the highlighted region is shown in the same figure on the right. Yellow voxels in the middle of the axial slice classified as class 3 are a likely result of a partial volume effect from cerebrospinal fluid (CSF) or low signal-to-noise ratio (SNR) volume due to RF inhomogeneity of B_1^+.

We can be confident in the model's performance on unseen data acquired using the same protocol as the datasets we used during the training, as demonstrated in the experiment on data from an independent, healthy subject and the cross-validation experiments. However, regarding the generalisation of our method, it is difficult to say how the model would translate to spectra acquired using different sequences. A solution would be to expand the training and validation datasets by adding more labeled spectra to account for new sources of variability and to retrain the model. Additionally, we would like to address the fact that our current dataset is labeled by a single expert and, as a result, our model is essentially reproducing the expert's opinion. While that might arguably be considered a limitation, we want to highlight the rater's expertise, which is reflected, among other things, in them having analysed clinical routine neuroradiologic MRSI data of more than 4000 patients over the span of several decades.

Comparison of our method with the existing ones is challenged for several reasons. Firstly, the method was developed on EPSI 3D-MRSI data acquired with a novel SLOW spectral editing, which introduces unique features to our dataset. Secondly, the comparison is limited by our definition of classes and the unique four-class framework. Lastly, validating the method on publicly available datasets would also be influenced by the rater bias. Finally, one additional significant benefit of the proposed model is its computational speed. On an average CPU, the model can classify 3D MRSI datasets similar to the ones used in this research (which contain about 25,000 spectra) in under a minute. The performance is naturally considerably faster when running the model on a GPU, where the same results can be obtained within seconds of processing time.

5 | CONCLUSIONS

First and foremost, we highlight the proposed approach's uniqueness. While acknowledging the previous works that have attempted similar goals, we note that this is the first to target SLOW MRSI spectra. Moreover, with the introduction of more labeled data to the training set, this simple deep-learning approach should deal efficiently with vast MRSI datasets.

We attempted to approach the problem from the four-class perspective, which would distinguish the defined classes of spectral quality successfully. As mentioned previously, the aim was to split the usable spectra class into two, thus inferring the extent to which the spectra require simpler or more complex postprocessing steps for metabolite quantification. It is apparent, though, that such a distinction was hard to make. Both the intrareader assessment and the model results demonstrate this.

However, we have proven the model's generalisation capabilities in terms of having different subjects and anatomic variability that can influence the results. It is important to stress that the model does not discard spectra coming from tumor regions and still deems them proper for quantification purposes.

Another thing to consider is how difficult it would be to adapt the method if something changed in the acquisition protocol of brain MRSI (e.g., different echo time or other spectral editing). The model could be retrained to incorporate such variations by adding additional labeled data at the training stages. It allows for the processing of different kinds of spectra as long as the spectral dimensions match the input dimensions of the network, which can be adjusted with sampling techniques should they not match. The most time-consuming step in this scenario would be to label the new data manually.

Lastly, we want to highlight the importance of a pipeline such as the one we present for when MRSI becomes an integral part of the clinical routine. Having a robust quality-control model that can process a large number of spectra in a short time would be very effective in saving time and guiding the clinician toward valuable spectral information for decision-making. The method presented in this article is being integrated into the SpectrIm plugin for the jMRUI software.

ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska–Curie grant agreement No. 813120. The research was additionally supported by the Swiss National Science Foundation (SNSF) grant number 182569. Open access funding provided by Inselspital Universitatsspital Bern.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.

ORCID

Mladen Rakić https://orcid.org/0000-0002-1950-6760

Johannes Slotboom https://orcid.org/0000-0001-5121-9852

REFERENCES

1. Kreis R, Boer V, Choi I-Y, et al. Terminology and concepts for the characterization of in vivo MR spectroscopy methods and MR spectra: Background and experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4347.
2. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4257.
3. Shakir TM, Fengli L, Chenguang G, Chen N, Zhang M, Shaohui M. ^{1}H-MR spectroscopy in grading of cerebral glioma: A new view point, MRS image-quality assessment. *Acta Radiol Open*. 2022;11(2):20584601221077068.
4. Kreis R. Issues of spectral quality in clinical ^{1}H-magnetic resonance spectroscopy and a gallery of artifacts. *NMR Biomed*. 2004;17(6):361-381.
5. Keevil SF. Spatial localization in nuclear magnetic resonance spectroscopy. *Phys Med Biol*. 2006;51(16):R579.
6. Landheer K, Schulte RF, Treacy MS, Swanberg KM, Juchem C. Theoretical description of modern ^{1}H in vivo magnetic resonance spectroscopic pulse sequences. *J Magn Reson Imaging*. 2020;51(4):1008-1029.
7. Felblinger J, Kreis R, Boesch C. Effects of physiologic motion of the human brain upon quantitative ^{1}H-MRS: analysis and correction by retro-gating. *NMR Biomed Int J Devoted Dev Appl Magn Reson Vivo*. 1998;11(3):107-114.
8. Haupt CI, Kiefer AP, Maudsley AA. In-plane motion correction for MR spectroscopic imaging. *Magn Reson Med*. 1998;39(5):749-753.
9. Kim D-H, Adalsteinsson E, Spielman DM. Spiral readout gradients for the reduction of motion artifacts in chemical shift imaging. *Magn Reson Med: Off J Int Soc Magn Reson Med*. 2004;51(3):458-463.
10. Goelman G, Liu S, Fleysher R, Fleysher L, Grossman RI, Gonen O. Chemical-shift artifact reduction in hadamard-encoded MR spectroscopic imaging at high (3T and 7T) magnetic fields. *Magn Reson Med: Off J Int Soc Magn Reson Med*. 2007;58(1):167-173.
11. Posse S, Otazo R, Dager SR, Alger J. MR spectroscopic imaging: principles and recent advances. *J Magn Reson Imaging*. 2013;37(6):1301-1325.
12. Klose U. In vivo proton spectroscopy in presence of eddy currents. *Magn Reson Med*. 1990;14(1):26-30.
13. Menze BH, Kelm BM, Weber M-A, Bachert P, Hamprecht FA. Mimicking the human expert: pattern recognition for an automated assessment of data quality in MR spectroscopic images. *Magn Reson Med: Off J Int Soc Magn Reson Med*. 2008;59(6):1457-1466.
14. Pedrosa de Barros N, McKinley R, Wiest R, Slotboom J. Improving labeling efficiency in automatic quality control of MRSI data. *Magn Reson Med*. 2017;78(6):2399-2405.
15. Jiru F, Skoch A, Klose U, Grodd W, Hajek M. Error images for spectroscopic imaging by LCModel using Cramer–Rao bounds. *MAGMA*. 2006;19(1):1-14.
16. Maudsley AA, Domenig C, Govind V, Darkazanli A, Studholme C, Arheart K, Bloomer C. Mapping of brain metabolite distributions by volumetric proton MR spectroscopic imaging (MRSI). *Magn Reson Med: Off J Int Soc Magn Reson Med*. 2009;61(3):548-559.
17. Slotboom J, Nirkko A, Brekenfeld C, Van Ormondt D. Reliability testing of in vivo magnetic resonance spectroscopy (MRS) signals and signal artifact reduction by order statistic filtering. *Meas Sci Technol*. 2009;20(10):104030.
18. Gurbani SS, Schreibmann E, Maudsley AA, et al. A convolutional neural network to filter artifacts in spectroscopic MRI. *Magn Reson Med*. 2018;80(5):1765-1775.
19. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: International Conference on Artificial Neural Networks; 2011:52-59.
20. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533-536.
21. Pintelas E, Livieris IE, Pintelas PE. A convolutional autoencoder topology for classification in high-dimensional noisy image datasets. *Sensors*. 2021;21(22):7731.
22. Rakić M, Cabezas M, Kushibar K, Oliver A, Lladó X. Improving the detection of autism spectrum disorder by combining structural and functional MRI information. *NeuroImage: Clin*. 2020;25:102181.
23. Meng Q, Catchpoole D, Skillicom D, Kennedy PJ. Relational autoencoder for feature extraction. In: International Joint Conference on Neural Networks (IJCNN); 2017:364-371.
24. Supratak A, Li L, Guo Y. Feature extraction with stacked autoencoders for epileptic seizure detection. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE; 2014:4184-4187.
25. Xiong Y, Lu Y. Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. *IEEE Access*. 2020;8:27821-27830.
26. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. *Int MICCAI Brainlesion Workshop*. 2017;3:450-462.
27. Valverde S, Cabezas M, Roura E, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*. 2017;155:159-168.

28. Rakić M, Vercruyssen S, Van Eyndhoven S, et al. icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. *NeuroImage: Clinical*. 2021;31:102707.
29. Weng G, Radojewski P, Sheriff S, et al. SLOW: A novel spectral editing method for whole-brain MRSI at ultra high magnetic field. *Magn Reson Med*. 2022;88(1):53-70.
30. Lam F, Ma C, Clifford B, Johnson CL, Liang Z-P. High-resolution ^{1}H-MRSI of the brain using SPICE: data acquisition and image reconstruction. *Magn Reson Med*. 2016;76(4):1059-1070.
31. Ebel A, Maudsley AA. Improved spectral quality for 3D MR spectroscopic imaging using a high spatial resolution acquisition strategy. *Magn Reson Imaging*. 2003;21(2):113-120.
32. Haupt CI, Schuff N, Weiner MW, Maudsley AA. Removal of lipid artifacts in 1H spectroscopic imaging by data extrapolation. *Magn Reson Med*. 1996; 35(5):678-687.
33. Bilgic B, Chatnuntawech I, Fan AP, Setsompop K, Cauley SF, Wald LL, Adalsteinsson E. Fast image reconstruction with l2-regularization. *J Magn Reson Imaging*. 2014;40(1):181-191.
34. Ma C, Lam F, Ning Q, Johnson CL, Liang Z-P. High-resolution ^{1}H-MRSI of the brain using short-TE SPICE. *Magn Reson Med*. 2017;77(2):467-479.
35. Tkáč I, Deelchand D, Dreher W, et al. Water and lipid suppression techniques for advanced ^{1}H MRS and MRSI of the human brain: experts' consensus recommendations. *NMR Biomed*. 2021;34(5):e4459.
36. Maudsley AA, Darkazanli A, Alger JR, et al. Comprehensive processing, display and analysis for in vivo MR spectroscopic imaging. *NMR Biomed*. 2006; 19(4):492-503.
37. Pedrosa de Barros N, McKinley R, Knecht U, Wiest R, Slotboom J. Automatic quality control in clinical ^{1}H MRSI of brain cancer. *NMR Biomed*. 2016; 29(5):563-575.
38. Stefan DDCF, Di Cesare F, Andrasescu A, et al. Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. *Meas Sci Technol*. 2009;20(10):104035.
39. Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition IEEE; 2010:2528-2535.
40. Chollet F. Keras: The Python deep learning library. *Astrophys Source Code Libr*. 2018:ascl-1806.
41. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. *Adv Neural Inform Process Syst*. 2017:30.
42. Tommasi T, Patricia N, Caputo B, Tuytelaars T. A deeper look at dataset bias. *Domain Adapt Comput Vision Appl*. 2017:37-55.
43. Torralba A, Efros AA. Unbiased look at dataset bias. *CVPR*. 2011;2011:1521-1528.

How to cite this article: Rakić M, Turco F, Weng G, Maes F, Sima DM, Slotboom J. Deep learning pipeline for quality filtering of MRSI spectra. *NMR in Biomedicine*. 2023;e5012. doi:10.1002/nbm.5012