








Reliability of Quantification Estimates in MR Spectroscopy: CNNs vs Traditional Model Fitting

Rudy Rizzo^{1,2} , Martyna Dziadosz^{1,2}, Sreenath P. Kyathanahally³ ,
Mauricio Reyes^{4,5} , and Roland Kreis^{1,2}  

¹ MR Methodology, Diagnostic and Interventional Neuroradiology, University of Bern, Bern, Switzerland

`roland.kreis@insel.ch`

² Translational Imaging Center, sitem-insel, Bern, Switzerland

³ Data Science

for Environmental Research Group, Department of System Analysis, Integrated Assessment and Modelling; EAWAG, Dübendorf, Switzerland

⁴ Insel Data Science Center, Inselspital, Bern University Hospital, Bern, Switzerland

⁵ ARTOG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

Abstract. Magnetic Resonance Spectroscopy (MRS) and Spectroscopic Imaging (MRSI) are non-invasive techniques to map tissue contents of many metabolites in situ in humans. Quantification is traditionally done via model fitting (MF), and Cramer Rao Lower Bounds (CRLBs) are used as a measure of fitting uncertainties. Signal-to-noise is limited due to clinical time constraints and MF can be very time-consuming in MRSI with thousands of spectra. Deep Learning (DL) has introduced the possibility to speed up quantitation while reportedly preserving accuracy and precision. However, questions arise about how to access quantification uncertainties in the case of DL. In this work, an optimal-performance DL architecture that uses spectrograms as input and maps absolute concentrations of metabolites referenced to water content as output was taken to investigate this in detail. Distributions of predictions and Monte-Carlo dropout were used to investigate data and model-related uncertainties, exploiting ground truth knowledge in a synthetic setup mimicking realistic brain spectra with metabolic composition that uniformly varies from healthy to pathological cases. Bias and CRLBs from MF are then compared to DL-related uncertainties. It is confirmed that DL is a dataset-biased technique where accuracy and precision of predictions scale with metabolite SNR but hint towards bias and increased uncertainty at the edges of the explored parameter space (i.e., for very high and very low concentrations), even at infinite SNR (noiseless training and testing). Moreover, training with uniform datasets or if augmented with critical cases showed to be insufficient to prevent biases. This is dangerous in a clinical context that requires the algorithm to be unbiased also for concentrations far from the norm, which may well be the focus of the investigation since these correspond to pathology, the target of the diagnostic investigation.

Keywords: Magnetic Resonance Spectroscopy · Convolutional neural networks · Model fitting · Quantification · Reliability · Uncertainties

1 Introduction

Magnetic Resonance Spectroscopy (MRS) and Spectroscopic Imaging (MRSI) are non-invasive methods for determining in-situ metabolic profile maps in humans or animals. A chemical-composition-specific response is evoked from localized tissue regions using an MRI scanner and allows the acquisition of a Voigt-damped time-domain signal, which results from a superposition of multiple metabolite signals. The resonance line patterns are metabolite-specific, reflecting the spin-systems, while their concentrations are proportional to the signal amplitude [1] (Fig. 1).

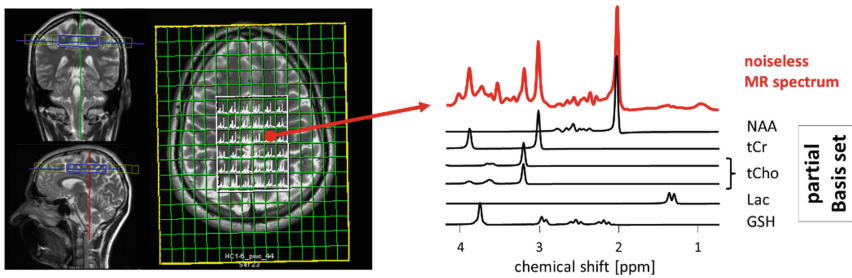


Fig. 1. MRSI acquisition with zoom-in to a sketch of a noiseless MR spectrum (real part) and relative spectral basis set outlined for five metabolites: N-Acetylaspartate (NAA), total-Creatine (tCr), total-Choline (tCho), Lactate (Lac), and Glutathione (GSH).

Quantification is traditionally based on parameter estimation with Model Fitting (MF), minimizing the difference between the data and a parameterized model function. Despite many fitting approaches [2–5], robust and accurate measurement of metabolite concentrations remains challenging [6, 7], mainly due to: (i) severely overlapping metabolite patterns, (ii) poor signal-to-noise ratio (SNR), and (iii) unknown background signals and peak lineshape (incomplete prior knowledge). As a result, the problem is ill-posed, and current techniques still hamper translation to clinical routine.

Supervised Deep Learning (DL) exploits neural networks to find key properties contained in large data sets and to generate complicated nonlinear mappings between inputs and outputs [8]. It thus requires no prior knowledge or formal assumptions. However, it is shown to be frequently biased towards the conditions prevalent in the datasets used in training [9]. DL in MRS quantification is increasingly explored [10–12] and has shown to speed up quantitation while reportedly preserving accurate estimates if compared to MF. Still, questions regarding the reliability of DL quantification have arisen.

Uncertainty measures provide information about how reliably or accurately a given algorithm performs a given task. This information in turn can be used to leverage the decision-making process for a user (e.g., how much to trust the estimated concentration of a metabolite) or to enable optimization of the acquisition technique or the algorithm employed to estimate results (e.g., focusing on areas of high uncertainty [13]). Given MRS restrictions to comply with clinical time frames, the repetition of multiple MRS measurements to determine repeatability is forbidding. Thus, estimates of uncertainty obtained from MRS model fitting of a single measurement are often taken as proxy:

the Cramer Rao Lower Bounds (CRLBs) [14] estimate uncertainties as function of the model (presumed to be true) and SNR; they represent the uncertainty limit for unbiased estimators. It is fundamental to access a CRLB-comparable uncertainty measure for MRS metabolite quantification by DL [15]. Neural network uncertainties originate from noise inherent in the data (aleatoric uncertainty) and uncertainty in the model parameters (epistemic uncertainty) [16, 17]. In the current work, an optimal-performance Convolutional Neural Network (CNN) architecture is designed to quantify metabolites, and metrics based on bias and spread of predicted distribution of concentrations are used to explain aleatoric uncertainties. Epistemic uncertainties are explored via Monte-Carlo dropout [18]. The reliability of MRS quantification is then compared between the two approaches. In-silico simulations guarantee knowledge of Ground Truth (GT).

2 Methods

2.1 Simulations

Spectral patterns were simulated for 16 metabolites recorded at 3T with a semi-LASER protocol [19, 20; TE = 35 ms, 4 kHz spectral width, 4096 points. To mimic pathological conditions, metabolite concentrations are varied independently and uniformly between 0 and twice a normal reference concentration for healthy human brain [1, 21–23]. A constant downsampled water reference (64.5 mM) is added at 0.5 ppm to ease quantitation. Macromolecular background (MMBG) signals and Gaussian broadening mimic in vivo conditions and were independently and uniformly varied (shim 2–5 Hz, MMBG amplitude \pm 33%). Two datasets with 20'000 entries randomly split in training (80%), validation (10%), and testing sets (10%) are generated: one with independent, realistic white Gaussian noise realizations (time-domain water-referenced SNR 5–40), the other noiseless (Fig. 2).

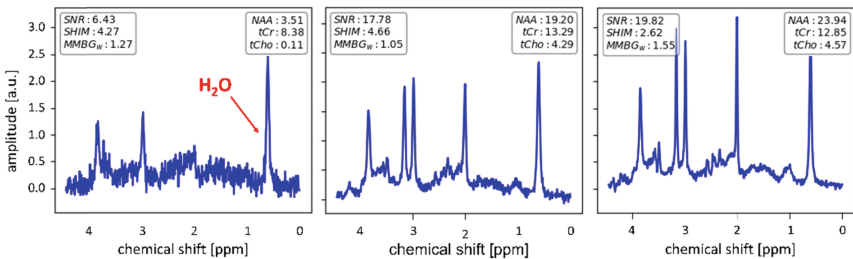


Fig. 2. Samples of realistic simulated spectra. SNR, shim, and MMBG intensity are indicated. Concentrations reported for 3 major metabolites in mM. Downsampled reference peak indicated.

High-frequency-resolved spectrograms [24] were used as input for the CNNs. A spectrogram is a complex, 2D time-frequency domain representation of a spectrum where each row reflects the frequency content of a specific time segment of the MRS signal. It is calculated using the Short-Temporal Fourier Transform (STFT), which allows for varying degrees of frequency and time resolution depending on the size of the Fourier

analysis window. A large window size established through zero-filling is paired with a tiny overlap interval to maximize frequency resolution while compromising temporal resolution (Fig. 3).

2.2 Quantification via Deep Learning

Two Bayesian hyper-parameterized [25] shallow CNNs [26] were trained and tested with the two datasets. They had emerged as optimal DL quantification methods from 24 tested scenarios with different architectures, input forms, and active learning data augmentation. Relative concentrations are provided as output but referencing to the water signal yields absolute concentrations (Fig. 3). Training and validation sets were randomly assigned to train the CNN on a maximum number of 100 epochs and with batch normalization of 50. The learning rate was modulated via the adaptive moment (ADAM) estimation algorithm [27]. Mean-squared error (MSE) served as loss function. Visualization of training and validation loss over epochs combined with the implementation of early-stopping criterion monitoring minimization of validation loss with patience ten was used as a reference for tuning the network parameter space.

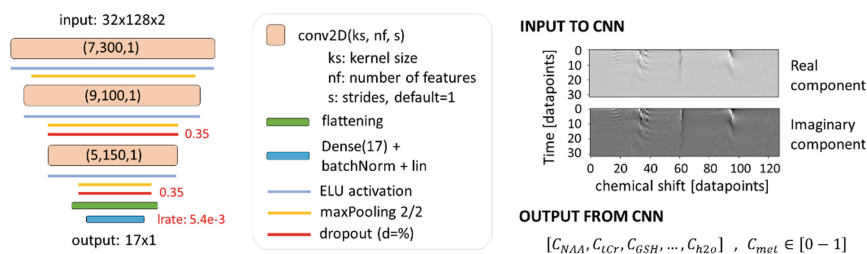


Fig. 3. Shallow CNN architecture, sample of input spectrogram, and output vector description.

Aleatoric uncertainties were evaluated via bias of the DL predictions from GT and spread of these predictions, both as estimated for 20 different bins that cover the whole GT concentration range of each metabolite (called Bin-Spread-Function in Fig. 4). Monte-Carlo dropout consisted of testing the trained model 100 times with activated dropout layers. Thus, the network structure slightly changed for each prediction (i.e., a different set of neurons was switched off) although preserving its weights. The 100 predictions yielded a distribution (called Point-Spread-Function in Fig. 4) for any sample in the test set. The bias and spread of these distributions were then calculated for every test sample, averaged for every GT value, and used as epistemic uncertainties. They highlight the susceptibility of predictions to model variation.

2.3 Quantification via Model Fitting

The test set spectra were fitted using fitAID [5]. The model consisted of a weighted sum of Voigt lines with fixed Lorentzian (GT value) and estimated Gaussian widths. Areas of the metabolites were restricted in $[-0.5 \cdot \mu, +2.5 \cdot \mu]$, where μ is the average

concentration in the testing and training set distribution for each metabolite, aiming to bound the fitting condition to known prior knowledge, mimicking the implicit restrictions of DL algorithms. Bias from GT and CRLB are used as uncertainty measures.

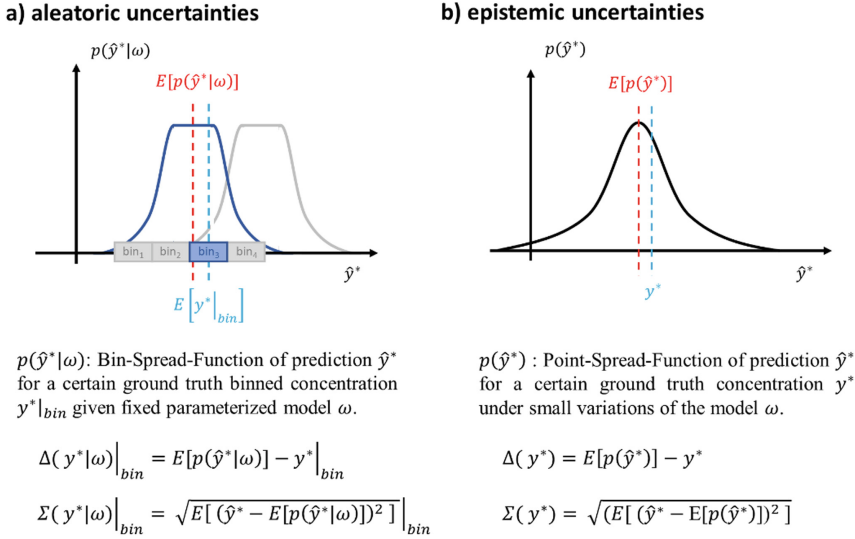


Fig. 4. Distribution of uncertainties: bias (Δ) represents a deviation from ground truth, and spread (Σ) represents variability of predictions around their center: the expected value ($E[\cdot]$).

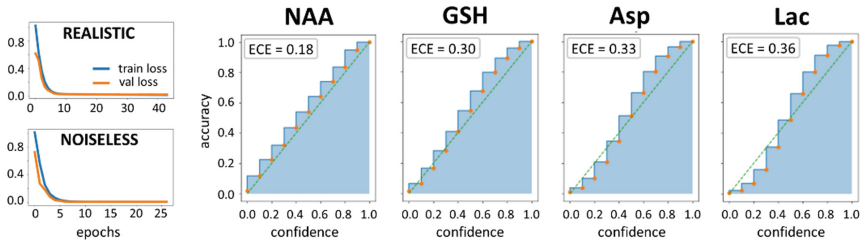


Fig. 5. Training and validation curves. Reliability diagrams and expected calibration error.

3 Results

Results are reported for four metabolites with progressively lower relative SNR: NAA, GSH, Asp, and Lac (c.f. Fig. 1). Figure 5 shows training and validation curves for both networks. Network calibration is investigated for regression, where the design is assumed to predict the Cumulative Distribution Function of relative metabolite concentrations [28, 29]. Reliability diagrams are reported for realistic simulations. Quantification of Lac and Asp is mildly overconfident for low concentrations and underconfident for high

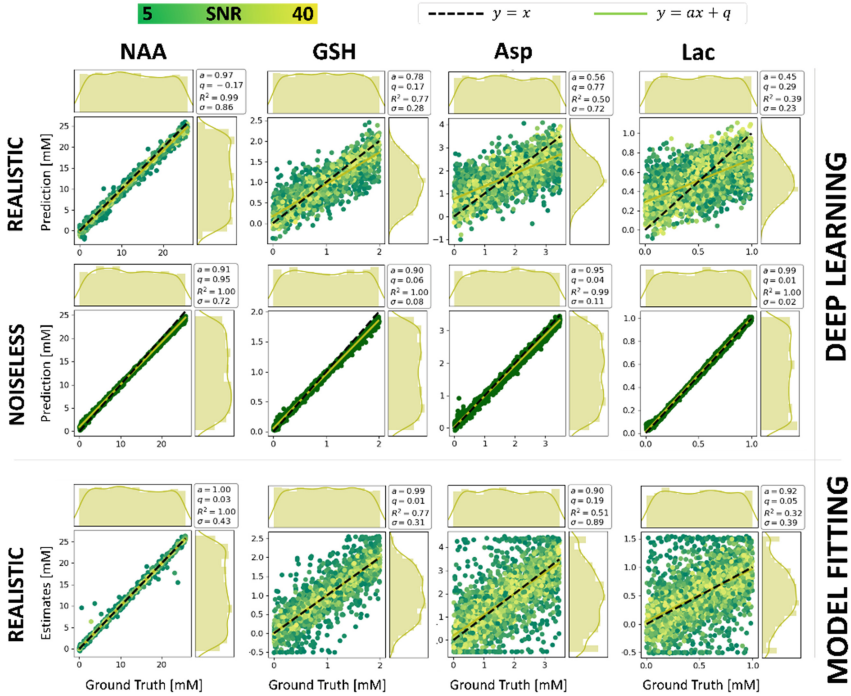


Fig. 6. Maps and marginal distribution of CNN predictions vs. GT for four metabolites and two datasets (realistic, noiseless) and model estimates vs. GT for the realistic case.

concentrations. However, the network can be considered well-calibrated for every target metabolite [13].

Figure 6 reports CNN predictions and MF estimates vs. GT values. A linear regression model is fit to estimate the quality of the prediction. Marginal distributions of GT and predicted values are displayed. Ideal predictions would display as a diagonal line ($y = ax + q$) with minimal spread (RMSE, σ). In line, distributions of predictions would mirror the uniform GT distributions. Considering the realistic case in DL and going from left to right, predicted distributions become less uniform and get more biased towards a mean expected value of the training range. This phenomenon is reflected in lower a and R^2 values and is emphasized for metabolites with low relative SNR (e.g., GSH, Asp, and Lac). Noiseless simulations show a significantly reduced bias. MF estimates show a better spread over the concentration range with $a \rightarrow 1$ and $q \rightarrow 0$ even for metabolites with low relative SNR. RMSE (σ) is lower in MF than DL for NAA but higher in the case of GSH, Asp, and Lac.

Figure 7 maps aleatoric and epistemic uncertainties as function of GT values for DL. Epistemic uncertainties indicate higher variability of predictions at the boundaries of the concentration ranges, which is paired with higher biases for aleatoric uncertainties. In the noiseless scenario, it is evident how the point-spread function is affected by a larger spread at the edges (i.e., U-shape). Training and testing with noise show the same trend if relative SNR is high enough. Model fitting appears unbiased (average bias as orange

line) except for a small effect at the parameter boundaries and biased larger outliers (blue line) for metabolites with low SNR. CRLB are concentration-independent (linear fit parameter). Moreover, average CRLBs are confirmed to represent a lower bound to standard deviation (σ) of the fit error.

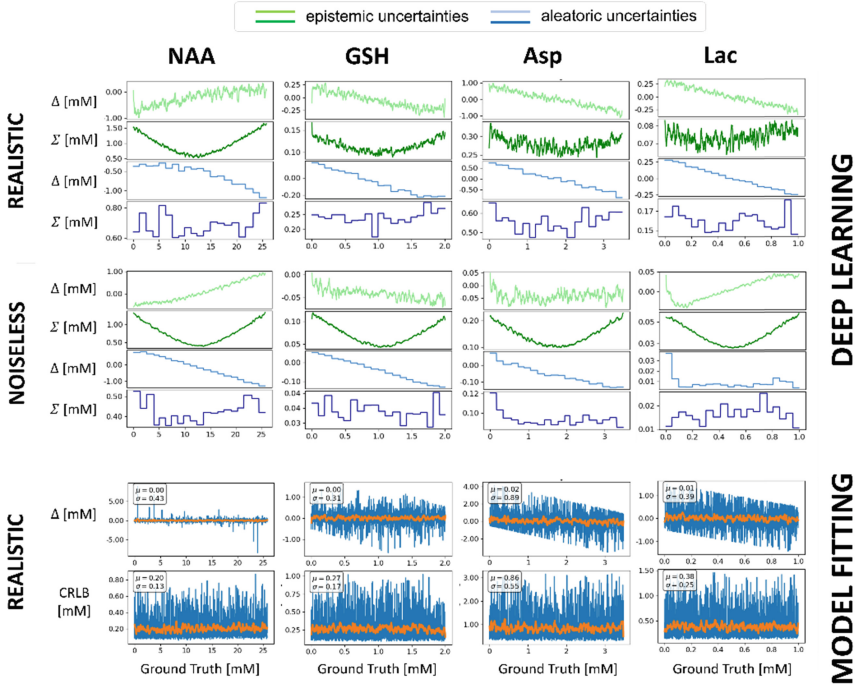


Fig. 7. CNNs' bias and spread of epistemic (green) and aleatoric (blue) uncertainties vs. GT values. MF bias and CRLBs vs. GT: single values (blue) and interpolated (orange). Estimated standard deviation (σ) of the bias can be compared to the estimated average (μ) CRLB. (Color figure online)

4 Discussion

Predicted concentrations should be unbiased, thus returning uniform distributions for uniform training and test data. However, our CNN predictions for real-world simulations tend towards the mean of the test data. Predictions at the boundaries of the testing range are folded back towards the mean value in case of strong uncertainty (i.e., low metabolite SNR), given the lack of knowledge outside the boundaries. Indeed, it is found that the prediction bias is influenced by the limited concentration ranges used in training: epistemic uncertainties indicate higher variability of predictions at the boundaries of the concentration ranges, which is paired with higher biases for aleatoric uncertainties. Though not exploring the exact same architecture, it is suspected that previous DL MRS

approaches may show similar deficiencies [10–12]. An ideal noiseless scenario also shows similar findings: uncertainty is lower in absolute values compared to the realistic case, but a similar dependency on the concentration range is found (U-shape), confirming that the DL prediction constitutes a biased estimator with uncertainties that depend on the placement of the test case in the training range. Training and testing with noise show the same trend if relative SNR is high enough. Metabolites in high concentrations suffer from comparable epistemic spread as those in low concentrations.

MF is confirmed to be unbiased on average. Individual estimates are mildly influenced by restrained concentration ranges (i.e., prior knowledge). CRLBs are confirmed to be a measure of variance that is independent of the estimated concentration.

The current study considers a limited synthetic dataset that does not cover the whole range of possible in-vivo sources of variability despite its aim to mimic realistic performances. Furthermore, a single CNN design tuned for metabolite's quantification is investigated, even if optimized via multiple iterations and with the best combination of input/output spectroscopic information (i.e., spectrograms and relative concentrations), it is not possible to draw general conclusion for MRS quantification via DL algorithms. Uncertainty investigation is limited to two uncertainty measures, which must be taken with their benefits, reliability, and limitations compared to other measures [13].

5 Conclusions

Four measures for aleatoric and epistemic uncertainties are provided, partly representing accuracy and precision of predictions. They scale with metabolite SNR but hint towards bias and increased uncertainty at the edges of the explored parameter space for (these) DL methods in many cases, even at infinite SNR.

Deep Learning does not require feature selection by the user, but the potential intrinsic biases at training set boundaries act like soft constraints in traditional modeling, leading estimated values to an apparently precise (low mean deviation) estimate reflecting an expectation value over the normal concentration range used in training. This is dangerous in a clinical context that requires the algorithm to be unbiased to outliers, which may well be the focus of the investigation corresponding to pathological data.

Further investigation to access more stable predictions is needed: (i) training with even larger concentration ranges, such that the region of interest is well inside the training range where uncertainties are limited, (ii) consider ensemble of networks to strengthen network performances for outliers or (iii) implementation of Batch Nuclear-norm Maximization to improve discriminability and diversity of the predictions [30].

6 Data Availability Statement

The simulated datasets and network architecture that support the findings of this study are available at <https://github.com/bellarude>.

Acknowledgments. This work is supported by the Marie-Sklodowska-Curie Grant ITN-39 237 (Inspire-Med) and the Swiss National Science Foundation (#320030–175984).

References

1. de Graaf, R.A.: In Vivo NMR Spectroscopy: principles and techniques, 3rd ed., WILEY (2018)
2. Ratiney, H., et al.: Time-domain semi-parametric estimation based on a metabolite basis. *NMR Biomed.* **18**, 1–13 (2005)
3. Provencher, S.: Estimation of metabolite concentrations from localized in vivo. *Magn. Reson. Med.* **30**(6), 672–679 (1993)
4. Wilson, M., et al.: A constrained least-squares approach to the automated quantitation of in vivo 1h magnetic resonance spectroscopy data. *Magn. Reson. Med.* **65**(1), 1–12 (2011)
5. Chong, D.G.Q., et al.: Two-dimensional linear-combination model fitting of magnetic resonance spectra to define the macromolecule baseline using FiTAID, a Fitting Tool for Arrays of Interrelated Datasets. *Magn. Reson. Mater. Physics, Biol. Med.* **24**, 147–164 (2011)
6. Bhogal, A.A., et al.: 1H-MRS processing parameters affect metabolite quantification: the urgent need for uniform and transparent standardization. *NMR in Biomed.* **30**, e3804 (2017)
7. Marjanska, M., et al.: Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM. *Magn. Reson. Med.* **87**(1), 11–32 (2022)
8. Wick, C.: Deep Learn. *Informatik-Spektrum* **40**(1), 103–107 (2016). <https://doi.org/10.1007/s00287-016-1013-2>
9. Gyori, N.G., et al.: Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn. Reson. Med.* **87**(2), 932–947 (2022)
10. Lee, H.H., et al.: Deep learning-based target metabolite isolation and big data-driven measurement uncertainty estimation in proton magnetic resonance spectroscopy of the brain. *Magn. Reson. Med.* **84**(4), 1689–1706 (2020)
11. Gurbani, S.S., et al.: Incorporation of a spectral model in a convolutional neural network for accelerated spectral fitting. *Magn. Reson. Med.* **81**, 3346–3357 (2018)
12. Hatami, N., Sdika, M., Ratiney, H.: Magnetic resonance spectroscopy quantification using deep learning. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Lecture Notes in Computer Science, vol. 11070, pp. 467–475. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_53
13. Jungo, A., et al.: Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation, [arXiv:1907.03338v2](https://arxiv.org/abs/1907.03338v2). (2019)
14. Bolliger, C.S., et al.: On the use of Cramér-Rao minimum variance bounds for the design of magnetic resonance spectroscopy experiments. *Neuroimage* **83**, 1031–1040 (2013)
15. Landheer, K., et al.: Are Cramer-Rao lower bounds an accurate estimate for standard deviations in in vivo magnetic resonance spectroscopy? *NMR Biomed.* **34**(7), e4521 (2021)
16. Gal, Y.: *Uncertainty in Deep Learning*, University of Cambridge (2016)
17. Kendall, A.: What uncertainties do we need in Bayesian deep learning for computer vision? [arXiv:1703.04977v2](https://arxiv.org/abs/1703.04977v2). (2017)
18. Gal, Y. et al.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning, [arXiv:1506.02142v6](https://arxiv.org/abs/1506.02142v6). (2016)
19. Soher, B.J., et al.: VeSPA: integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Proc. Int. Soc. Magn. Reson. Med.* **19**(19), 1410 (2011)
20. Oz, G., et al.: Short-echo, single-shot, full-intensity proton magnetic resonance spectroscopy for neurochemical profiling at 4 T: validation in the cerebellum and brainstem. *Magn. Reson. Med.* **65**(4), 901–910 (2011)
21. Marjańska, M., et al.: Region-specific aging of the human brain as evidenced by neurochemical profiles measured noninvasively in the posterior cingulate cortex and the occipital lobe using 1H magnetic resonance spectroscopy at 7 T. *Neuroscience* **354**, 168–177 (2017)

22. Hoefemann, M., et al.: Parameterization of metabolite and macromolecule contributions in interrelated MR spectra of human brain using multidimensional modeling. *NMR Biomed.* **33**(9), e4328 (2020)
23. Oz, G., et al.: Clinical proton MR spectroscopy in central nervous system disorders. *Radiology* **270**(3), 658–679 (2014)
24. Kyathanahally, S.P., et al.: Deep Learning approaches for detection and removal of ghosting artifacts in MR Spectroscopy. *Magn. Reson. Med.* **80**, 851–863 (2018)
25. Snoek, J.: Practical Bayesian optimization of Machine Learning Algorithms. In: 25th International Conference on Neural Information Processing System, vol. 2, pp. 2951–2959 (2012)
26. Espi, M., et al.: Exploiting spectro-temporal locality in deep learning based acoustic event detection. *J. Audio Speech Music Proc.* **26** (2015)
27. Kingma, D.P., et al.: Adam: A method for stochastic optimization. [Arxiv:1412.6980](https://arxiv.org/abs/1412.6980). (2014)
28. Niculescu-Mizil, A., et al.: Predicting good probabilities with supervised learning. In: 22nd ICML, pp.7–11 (2005)
29. Kuleshov, V., et al.: Accurate uncertainties for deep learning using calibrated regression. In: 35th ICML (2018)
30. Cui, S., et al.: Towards discriminability and diversity: batch Nuclear-norm Maximization under label insufficient situations. [Arxiv:2003.12237v1](https://arxiv.org/abs/2003.12237v1). (2020)