MULTIPLE INSTANCE-BASED TUMOR DETECTION FROM MAGNETIC RESONANCE SPECTROSCOPY DATA

Diyuan Lu
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
elu@fias.uni-frankfurt.de

Gerhard Kurz
Independent Researcher
Karlsruhe, Germany
kurz.gerhard@gmail.com

Nenad Polomac, Iskra Gacheva & Elke Hattingen
Institute for Neuroradiology at Frankfurt University Hospital
60528 Frankfurt am Main, Germany
{Nenad.Polomac, elke.hattingen}@kgu.de, s9819495@stud.uni-frankfurt.de

Jochen Triesch
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
triesch@fias.uni-frankfurt.de

ABSTRACT

The application of Deep Learning (DL) for medical diagnosis is often hampered by problems such as scarcity of training data, high inter-individual variability, and various sources of noise. Here, we study the problem of brain tumor detection using magnetic resonance spectroscopy (MRS), which poses all these challenges. Our solution uses a multiple-instance-based (MI-based) classification framework. In particular, we propose to generate bags of multiple spectra for each patient and design the model to be invariant to permutations of spectra within a bag. The bag-level classification is robust to label noise and permits effective data augmentation (DA). We demonstrate that our approach improves the patient-wise diagnosis accuracy from 66% to 75% when compared to single-instance-based classification. We also show that our proposed approach reaches performance levels of human neuroradiologists.

1 INTRODUCTION

We study the problem of brain tumor detection from MRS data. In clinical practice, MRS is a common tool to identify a brain tumor because it can be easily acquired alongside commonplace MR imaging (MRI) procedures and it uniquely reflects the biochemical composition of the brain tissue *in situ*. There has been increasing interest in MRS for clinical use because of the semiautomatic data acquisition, processing, and quantification (Ranjith et al., 2015; Hatami et al., 2018; González-Navarro & Belanche-Muñoz, 2009; Olliverre et al., 2018; Cruz-Barbosa & Vellido, 2011). However, the interpretation of MRS spectra is traditionally performed by human radiologists based on the size and location of certain peaks. In contrast, we train a model to learn informative features from the spectra as a whole. MRS data is that often corrupted by noise from head movements during the procedure or baseline distortions of the spectrum. Additionally, labels are only provided per patient and not per voxel, which could introduce labeling noise as spectra from the tumor-affected hemisphere can be falsely labeled as “tumor” even though they contain healthy brain tissue.

Given the ubiquity and importance of coping with noisy/weak labeling, many works on this topic have been published (Li et al., 2017; Lee et al., 2018; Han et al., 2018; Smyth et al., 2019). Some of them start with a small set of clean expert-labeled data (Li et al., 2017; Albarqouni et al., 2016), but this can be costly to obtain. Thus, models that are robust to noisy labels are highly desirable.

Multiple instance learning (MIL) is a framework to combat the problem, where detailed annotation for each single instance is noisy, laborious to obtain, or simply not available. It tries to make a decision based on a set of single instances instead of a decision for each single instance. MIL has been widely used in medical applications such as breast cancer detection (Sudharshan et al., 2019; Conjeti et al., 2017; Sadafi et al., 2020) and other forms of computer assisted diagnosis (Fung et al., 2007; Liu et al., 2018).

Our contributions are summarized as follows. (1) We present a novel classifier for MRS-based tumor detection that performs patient-wise classification while considering multiple spectra simultaneously. (2) We carefully evaluate the proposed method on data from previously unseen patients and show that it outperforms state-of-the-art methods. (3) We demonstrate that our model reaches performance levels comparable to those of human neuroradiologists.

2 METHODOLOGY

Data. We use 1H-MR-spectroscopy data collected from 422 patients recorded in the Institute for Neuroradiology of the University Hospital in Frankfurt between 01/2009 to 3/2019. The patients were suffering from either glial or glioneuronal first diagnosed tumors (the *tumor* group, 266 patients) or other non-neoplastic lesions (the *non-tumor* group, 156 patients). The tumor group included all spectra from the tumor-affected hemisphere. The non-tumor group consisted of spectra from both hemispheres. As a result, 7442 spectra (3388 non-tumor and 4054 tumor) were selected for further analysis. The obtained MRS examples are saved as 1-*d* arrays with 288 data points. The indices correspond to the position of metabolites and the values indicate signal intensities of corresponding metabolites. We normalize each spectrum to zero mean and unit variance. Thus, some spectra of patients with tumors may actually be recorded from healthy tissue outside the tumor but are still labeled as *tumor*. Another challenge we face with the data set is that the number of spectra from each patient is highly heterogeneous, ranging from 2 to 141 (17 ± 15, mean ± standard deviation).

Patient-wise Data Preparation. We propose to perform classification not on a single spectrum, but on a bag of spectra from the same patient. Among two directions to implement MIL, i.e., instance level and embedding level, we use the instance level approach, where we generate bags of spectra consisting of a fixed number $N \in \mathbb{N}$ of spectra from each patient, and each bag is in the shape $N \times 288$. Since, this is a combinatorial process, we can generate many such samples. This process can be viewed as a form of data augmentation (DA). However, the more bags we generate from one patient, the less diversity we introduce through the DA and the network may not generalize well. Empirically, we set the number of generated bags of one patient to three times his/her total single spectrum count. Further exploration of this choice might be beneficial in the future. Each training bag is assigned a class label $y^p \in \{tumor, non\text{-}tumor\}$ based on the diagnosis of the patient. More formally, our goal is to learn a function *f*, which takes a bag of spectra $\mathbf{x}^p = \{x_i^p, \ldots, x_N^p\}$ from patient p, and outputs the classification decision $\hat{y}^p$. The training objective is the classic cross entropy loss

$$\min_{\theta} \mathbb{E}_{P(\mathbf{x},\hat{y})}[-\log P_\theta(y = \hat{y}|\mathbf{x})] \, , \tag{1}$$

where θ are the parameters in the function *f*.

During training, we draw N spectra randomly with repetition from all the samples of a patient. This allows us to deal with patients that have fewer than N spectra. During testing, we switch off the data augmentation strategy and only allow the minimal repetition of the spectra to make sure that the number of bags to generate for patient p with N_p spectra is 1 if $N_p \leq N$ and $\lceil N_p/N \rceil$, if $N_p > N$.

The ability of the classifier to generalize to new previously unseen patients is of great clinical importance. Therefore, we apply a 5-fold leave-subjects-out cross validation scheme. To be specific, we divide the patient list into 5 sub-lists each with around 80 patients. During training and validation, we adopt a 4:1 split ratio of all generated bags and over-sample the minority class to keep the balance. We also completely withhold one test set (844 spectra) on which we obtain the prediction from neuroradiologists for comparison. The results are reported in Table 2.

Network Structure. When working with bags of MRS samples, we note that the order of the stacked spectra was randomly chosen and should not affect the result of the network. Thus, we design our network structure in such a way that the output of the network is independent of the order of the

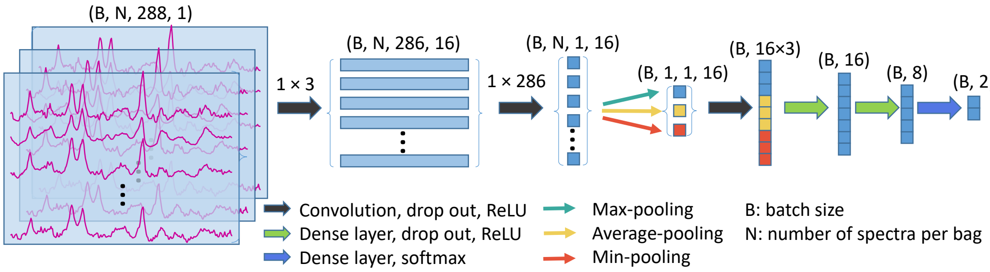

Figure 1: Proposed network structure.

Table 1: Performance on all cross validation sets. The results are shown as **mean ± standard deviation**. MCC: Matthews correlation coefficient, AUC: area under ROC curve. SI: single instance. MI: multiple instance. MLP: multi-layer perceptron. DA: data augmentation

Method	**Bag Accuracy**	**Patient Accuracy**	**AUC**	**F1-score**	**MCC**
SI-MLP	0.67 ± 0.05	0.66 ± 0.06	0.68 ± 0.07	0.65 ± 0.08	0.26 ± 0.01
Ours on SI	0.66 ± 0.04	0.66 ± 0.04	0.70 ± 0.04	0.67 ± 0.04	0.28 ± 0.07
Ours on MI w/o DA	0.62 ± 0.05	0.58 ± 0.08	0.68 ± 0.03	0.66 ± 0.06	0.24 ± 0.08
Ours on MI w/ DA	**0.72 ± 0.06**	**0.75 ± 0.04**	**0.79 ± 0.05**	**0.75 ± 0.06**	**0.43 ± 0.13**

input spectra. Specifically, we first apply 1-*d* convolution on each spectrum in the bag to increase the number of feature maps and then we apply a "large" convolution kernel to the feature map of each spectrum, which mimics a fully-connected network structure while dealing with well-aligned data such as MRS spectra. This way, we compress the feature map in each channel into a scalar value. We then aggregate the compressed minimum, maximum, and mean feature map of all convolutional channels and pass them through two dense layers before the softmax operation. Since the minimum, maximum, and mean operations are invariant to reordering the input spectra, the whole network inherits this invariance. The proposed network structure is shown in Fig. 1. It is trained with the Adam optimizer with default parameters and a batch size of 64.

3 Results

Overall Performance. To evaluate performance, we use the receiver operating characteristic (ROC) curve, which is a gold standard to evaluate the discriminative ability of a classifier. We report classification accuracy, area under the ROC curve (AUC), F1-score, and Matthews correlation coefficient (MCC). The MCC is generally considered as a balanced measure, which takes true negatives, true positives, false negatives, and false positives into account. Empirically, we found that using 16 spectra per bag yields the best results, on which we report our performance, shown in Table 1. For patients with less than 16 spectra, we randomly repeat the available spectra to form one bag.

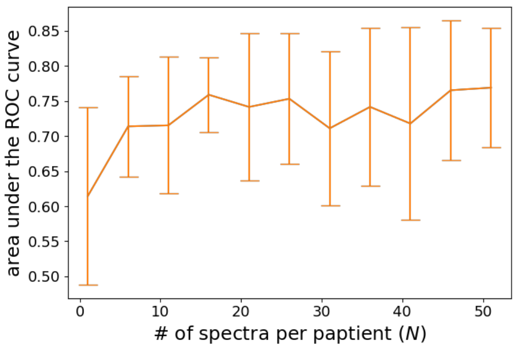

Figure 2: Averaged ROC-AUC across five leave-patients-out cross validation sets w.r.t the number of instances per bag.

Figure 3A shows the PCA visualization of the original data. We can see that the two classes are largely overlapping with each other, which imposes immense difficulties in tumor detection. In Fig. 3B, we show the PCA visualization of the network activity before the softmax output layer. We can see that, the two classes are more separated.

Table 2: Comparison with **neuroradiologists**

Method	**Bag Accuracy**	**Patient Accuracy**	**AUC**	**F1-score**	**MCC**
Neuroradiologists	–	0.69	–	0.56	**0.58**
Ours on SI	0.65 ± 0.02	0.64 ± 0.05	0.70 ± 0.02	0.57 ± 0.02	0.29 ± 0.04
Ours on MI w/o DA	0.57 ± 0.08	0.56 ± 0.06	0.65 ± 0.08	0.53 ± 0.08	0.15 ± 0.15
Ours on MI w/ DA	**0.71 ± 0.03**	**0.70 ± 0.03**	**0.81 ± 0.04**	**0.67 ± 0.02**	0.42 ± 0.05

Human vs. Machine. To assess how well our proposed method works in a more realistic clinical setting, we compared it to human neuroradiologists on one randomly selected test set. The results are shown in Tab. 2. The test set is divided into eight subsets and assigned to eight neuroradiologists. The performance of the model is computed in each corresponding subset for each individual neuroradiologist and averaged across all subsets. The proposed model on single instance classification achieved an AUC of 0.70, a MCC of 0.29, and an F1-score of 0.57. The model on multiple instance learning achieved an AUC of 0.81, an MCC of 0.42, and an F1-score of 0.67. It shows that the performance of our proposed method is at least as good as that of the human neuroradiologists except for the MCC. The reason is that the neuroradiologists achieved a specificity of 0.88 but at the cost of a low sensitivity of 0.54.

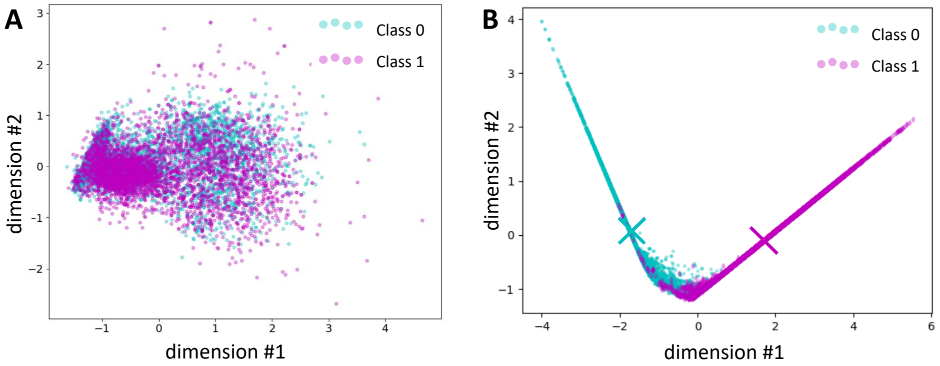

Figure 3: **A.** Principle component analysis (PCA) of original data colored by their labels. **B.** PCA of the network activities before the softmax layer.

Varying the Bag Size. To investigate the effect of the number of samples per bag, we vary the value from one (corresponding to single instance classification) to 51. The AUC as a function of N is shown in Fig 2. We found that there is an increasing trend between one and 16, after which performance plateaus.

4 Conclusion

This paper presents an MI-based tumor detection approach with noisily-labeled MRS data. We generate bags of instances from each patient, which expands the total training data set. The network structure is designed to be permutation-invariant within each bag. We show that our MI-based approach significantly improves the performance compared to SI-based classification and that applying data augmentation for generating more training data is crucial to obtain good results. Our final method achieves patient-wise diagnosis accuracy of 0.75, AUC-ROC of 0.79 and MCC 0.43. We also demonstrate that its performance is at least as good as that of human radiologists on all tested evaluation metrics except the MCC.

In future work, more exploration of other data augmentation strategies such as mixup, adding noise, and generative adversarial networks (GANs) might be interesting. Moreover, applying explainable machine learning methods could promote the transition to clinical practice.

REFERENCES

Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

Sailesh Conjeti, Magdalini Paschali, Amin Katouzian, and Nassir Navab. Deep multiple instance hashing for scalable medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 550–558. Springer, 2017.

Raúl Cruz-Barbosa and Alfredo Vellido. Semi-supervised analysis of human brain tumours from partially labeled mrs information, using manifold learning models. *International journal of neural systems*, 21(01):17–29, 2011.

Glenn Fung, Murat Dundar, Balaji Krishnapuram, and R Bharat Rao. Multiple instance learning for computer aided diagnosis. *Advances in neural information processing systems*, 19:425, 2007.

Félix F González-Navarro and Lluís A Belanche-Muñoz. Using machine learning techniques to explore 1h-mrs data of brain tumors. In *2009 Eighth Mexican International Conference on Artificial Intelligence*, pp. 134–139. IEEE, 2009.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Nima Hatami, Michaël Sdika, and Hélène Ratiney. Magnetic resonance spectroscopy quantification using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 467–475. Springer, 2018.

Kuang Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. ISBN 9781538664209.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li Jia Li. Learning from Noisy Labels with Distillation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1928–1936, 2017.

Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

Nathan Olliverre, Guang Yang, Gregory Slabaugh, Constantino Carlos Reyes-Aldasoro, and Eduardo Alonso. Generating magnetic resonance spectroscopy imaging data of brain tumours from linear, non-linear and deep learning models. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 130–138. Springer, 2018.

G Ranjith, R Parvathy, V Vikas, Kesavadas Chandrasekharan, and Suresh Nair. Machine learning methods for the classification of gliomas: Initial results using features extracted from mr spectroscopy. *The neuroradiology journal*, 28(2):106–111, 2015.

Ario Sadafi, Asya Makhro, Anna Bogdanova, Nassir Navab, Tingying Peng, Shadi Albarqouni, and Carsten Marr. Attention based multiple instance learning for classification of blood cell disorders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 246–256. Springer, 2020.

Luka Smyth, Dmitry Kangin, and Nicolas Pugeault. Training-valuenet: Data driven label noise cleaning on weakly-supervised web images. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 307–312. IEEE, 2019.

PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.