Training a Deep Neural Networks for Small and Highly Heterogeneous MRI Datasets for Cancer Grading

Marek Wodzinski[1], Tommaso Banzato[2], Manfredo Atzori[3],
Vincent Andrearczyk[3], Yashin Dicente Cid[4], Henning Müller[3,5]

***Abstract*— Using medical images recorded in clinical practice has the potential to be a game-changer in the application of machine learning for medical decision support. Thousands of medical images are produced in daily clinical activity. The diagnosis of medical doctors on these images represents a source of knowledge to train machine learning algorithms for scientific research or computer-aided diagnosis. However, the requirement of manual data annotations and the heterogeneity of images and annotations make it difficult to develop algorithms that are effective on images from different centers or sources (scanner manufacturers, protocols, etc.). The objective of this article is to explore the opportunities and the limits of highly heterogeneous biomedical data, since many medical data sets are small and entail a challenge for machine learning techniques. Particularly, we focus on a small data set targeting meningioma grading. Meningioma grading is crucial for patient treatment and prognosis. It is normally performed by histological examination but recent articles showed that it is possible to do it also on magnetic resonance images (MRI), so non-invasive. Our data set consists of 174 T1-weighted MRI images of patients with meningioma, divided into 126 benign and 48 atypical/anaplastic cases, acquired using 26 different MRI scanners and 125 acquisition protocols, which shows the enormous variability in the data set. The performed preprocessing steps include tumor segmentation, spatial image normalization and data augmentation based on color and affine transformations. The preprocessed cases are passed to a carefully trained 2-D convolutional neural network. Accuracy above 74% was obtained, with the high-grade tumor recall above 74%. The results are encouraging considering the limited size and high heterogeneity of the data set. The proposed methodology can be useful for other problems involving classification of small and highly heterogeneous data sets.**

***Index Terms*— deep learning, classification, grading, small data set, meningioma**

I. Introduction

Training a deep neural network using small and highly heterogeneous data set is a challenging task. However, it is a common issue in many medical problems. Very often there is not enough annotated or available data, or otherwise data can easily come from different medical centers, using different acquisition protocols. This problem is not trivial since using small and diverse data sets usually leads to overfitting or lack of knowledge generalization [1]. In this work, we present an approach to correctly train a deep neural network with a small and heterogeneous set of data represented by an exemplary data set dedicated to meningioma grading.

Meningiomas are common cancers and account for 33.8% of all primary intracranical neoplasms in the USA [2], [3]. Meningiomas are histopathologically graded into 3 classes: benign, atypical and anaplastic, abbreviated as Grade I, Grade II and Grade III respectively [2]. Based on the grading, a different treatment needs to be delivered to the patient. Currently, the grading is done by histological examination (a procedure that is in most cases invasive and time-consuming). Grading based on magnetic resonance imaging (MRI) could lead to faster and more efficient treatment planning.

Currently, there are no widely accepted methods to predict histological grading based on MRI, but recent work showed that deep learning classification can be a promising approach to address this problem [1]. Automatic, learned extraction of features that distinguish between differently graded tumors would be a great tool supporting treatment planning. However, predicting histological grading based on MRI is a challenging problem, also due to the fact that data sets are often relatively small (compared to the computer vision domain) and that data can be characterized by high heterogeneity (the images being acquired in many medical centers, using different equipment and protocols).

A. Related work

Several papers targeted the use of small data sets for deep learning in medical imaging domain. In [4], the authors proposed to use deep polynomial networks on small, but homogeneous, breast and prostate ultrasound data sets achieving better performance than traditional architectures. An interesting work about skin tumor diagnosis was presented in [5], where the authors trained a network on a relatively small data set obtaining better classification results than dermatologists. In [6], the authors combined transfer learning with domain adaptation, to perform emotion recognition using a small and heterogeneous data set. Influencing comparison between deep, shallow and deep-regularized networks was discussed in [7]. The authors showed that using shallow networks provides better results on small data sets than a deep, not regularized, model. In [8] a comprehensive study on deep image classification with small data sets was discussed. An interesting concept about supervised layer-wise training for

M. Wodzinski has been partly supported by the EU Project POWR.03.03.00-IP.08-00-P13/18 - PROM NAWA.

[1] AGH University of Science and Technology, Department of Measurement and Electronics, Al. Mickiewicza 30, 30-059 Cracow, Poland

[2] Department Of Animal Medicine, Productions Health, Viale dell'Università 16, 35020, Legnaro (PD), Italy

[3] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

[4] Warwick Manufacturing Group, University of Warwick, Coventry, United Kingdom

[5] University of Geneva, Geneva, Switzerland

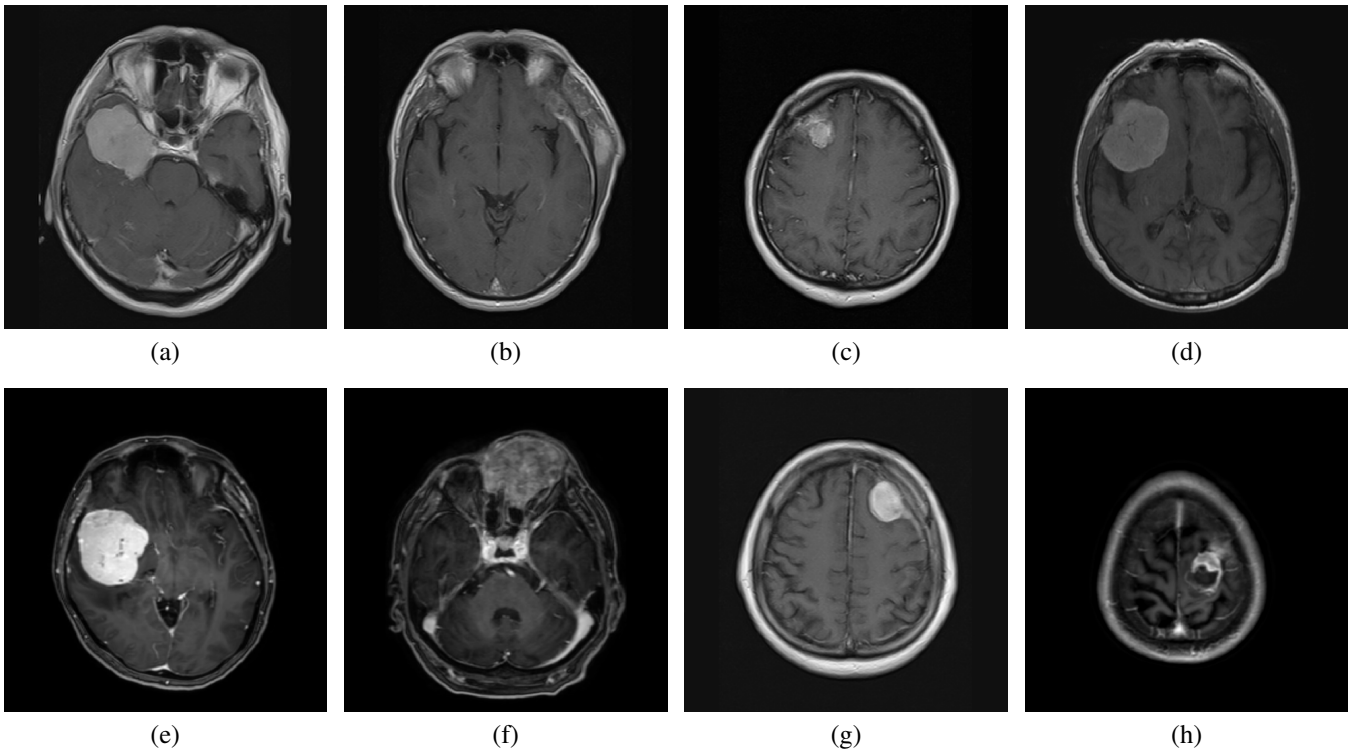

Fig. 1. Visualization of the data set heterogeneity using the slices with the largest tumor area. Images (a)-(e) present low-grade meningiomas while images (f)-(h) show high-grade meningiomas. One can see the data heterogeneity by examining both the intensity distribution and appearance differences between the cases.

deep networks was discussed in [9], showing results comparable to careful fine-tuning but with a significantly lower computational cost. An interesting concept about domain adaptation with adversarial training was discussed in [10], where the authors performed training on data coming from a distribution that was different from the data used for evaluation.

Regarding the meningioma grading, several studies showed that using a machine learning approach can be useful for the MRI-based improvement of the meningioma treatment. In [11], the authors proposed a method to firstly detect and then segment meningiomas in T1-weighted contrast-enhanced MRI using a relatively small data set. In [12], [13], the authors used the pretrained deep networks to perform the grading, independently using the contrasted T1 and ADC maps. The main drawback of the research was connected to manual intensity corrections. In [14], the authors used a traditional machine learning approach (based on hand-crafted features and logistic regression) to analyze a small data set that was acquired in a single medical center. An end-to-end deep network for meningioma grading, with a strong statistical validation, was proposed in [15], and achieved results better than approaches based on hand-crafted features. However, the data set was relatively homogeneous, since it was acquired in two hospitals. In this paper, we show that our approach is able to obtain comparable results using a highly heterogeneous data set, without the need of manual feature extraction.

B. Contribution

In this work, we present a deep learning approach to grade meningiomas using the T1-weighted, post-contrast MRI with an extremely small data set acquired in highly heterogeneous settings (i.e. in different medical centers, using several MRI scanners and acquisition protocols). We use a 2-D network to classify 3-D medical volumes. We show that a proper, careful training approach and a proper preprocessing may lead to a knowledge generalization even for such difficult and small data sets. The preprocessing consists of offline resampling, tumor cropping and strong data augmentation, while the training takes benefits from freezing and scheduling learning rates, oversampling less numerous classes and weighting them accordingly. The proposed methodology can be generalized to other deep learning problems involving small and highly heterogeneous data sets.

II. Methods

A. Method Description

We present a deep learning-based approach for cancer classification using a small meningioma data set acquired in heterogeneous clinical settings. The small data set size and its heterogeneity lead to instant overfitting for both training any deep network from scratch, as well as for a naive fine-tuning after transferring the weights from a pretrained network. As a result of the different acquisition protocols, the naively trained networks tend to learn features connected to a given scanner, not the features distinguishing between differently graded cancers. An exemplary visualization of the data set, showing its diversity, is presented in Figure 1.

The basic idea is to maximize the amount of useful information provided to the network while minimizing the number of parameters being optimized. We start by an initial, offline, preprocessing of the data set. At this point, all cases are resampled to the same physical spacing being the median of the spacing defined during the data acquisition (0.54mm x 0.54mm x 4mm). Eventually intensity normalization is performed. After the offline stage, all the processing is done online in the custom data loader, separately for each epoch. First, during the training set loading, a custom sampler is used to oversample the less numerous class. The meningiomas graded as atypical or anaplastic are occurring less frequently than the benign. In general, in real clinical settings, it is natural that one class is occurring more frequently than others. Therefore, the oversampler is used to ensure an equal number of weight updates per class. Afterwards, the volumes are strongly augmented by affine transformations, random cropping and flipping, and randomly changing the hue, saturation, and contrast. The augmentation is being done at the slice-level because, in further processing, the geometrical information between slices is not directly utilized during the convolutions. Then, only the slices with a present tumor are chosen and the area is cropped to contain only the tumor with a small, predefined, absolute offset equal to about 6mm. In this way, the amount of irrelevant information provided to the network is minimized. We do not utilize the 3-D information directly because: (i) 3-D convolutions often require much more training data to avoid overfitting, (ii) the data is highly anisotropic, (iii) the data sets used for pretraining 3-D networks are much smaller and harder to get than the 2-D ones (e.g. ImageNet [16] which was used for pretraining the used feature extractor). Then, the slices are properly reshaped, stacked in the batch dimension and passed to the network. We experimentally chose the pretrained ResNet-18 [17] as deep neural network, without noticing any substantial changes (apart from a lower training time and slower overfitting) compared to DenseNet-121 [18] or InceptionNet-v3 [19]. For each patient, the classification outcome is combined after the model forward pass, before calculating the loss, by averaging the calculated probabilities for all slices. In this way, it is possible to avoid direct slice-level training and labeling.

Important adjustments were done directly in the training procedure. We used Adam as the optimizer, after an experimental verification that it provides a slightly more stable learning process than SGD or RMSprop [20]. We froze all the parameter updates except the last fully connected layer and the last convolutional layer, together with its batch normalization. We used different learning rates for the fully connected layer and the convolutional layer, 10^{-4} and 10^{-6} respectively. We experimentally verified that using all the convolutional layers or providing too high learning rate for the last convolutional layer resulted in instant overfitting. Moreover, we applied a learning rate scheduler to decrease the learning rates by a constant factor after each epoch. The weighted cross-entropy was used as the cost function. We set a higher weight to the atypical and anaplastic cases than to the benign ones because in practice it is more important to not miss any malignant tumor than incorrectly grade benign cancer as malignant. Finally, the optimization starting point resulted crucial to obtain reasonable results. Since it was not time-consuming for training a network with a limited number of optimized parameters and a small data set, we randomly reinitialized the training procedure a random number of times.

B. Experimental Setup

The data set consists of 174 meningioma cases acquired using 26 different MRI scanners and 125 different acquisition protocols (with different magnetic field strength, acquisition plane, repetition time, echo time, flip angle, pixel size, slice thickness). There is no single acquisition protocol that can be considered as majority. The most frequently occurring protocol comes from a 1 Tesla scanner, using 5.5mm slice thickness, 0.937mm pixel size and echo time, flip angle and repetition time equal to 12ms, 90 degrees and 531ms respectively. The data set is divided into 126 benign and 48 atypical/anaplastic cases which are grouped together into a single class. There are 48, 40, 77 and 9 cases acquired with 1T, 1.5T, 3T and other magnetic field respectively. The slice thickness varies from 0.49mm to 5.5mm and the pixel size is between 0.21mm to 1mm. The data set was randomly divided into 5 folds to do the cross-validation, each with a 4:1 split ratio between the training and the validation set. The experimental procedures involving human subjects described in this paper were approved by the Ethical Committee of the University of Padua.

III. Results

A. Classification Results

All the training and validation procedures were done 5 times, for each fold separately to average the results and make them more realistic. In this way, it is possible to somehow approximate the results as if using a separate test set, which was impossible due to the very limited size of the data set. We show the averaged (average $\pm$ standard deviation) confusion matrix in Table I and the averaged grading results in Table II. We compared the results with two naive approaches: (i) training the network from scratch, (ii) naive fine-tuning the network without using the scheduling, different learning rates and freezing the parameters. The accumulated ROC curves, for the proposed and compared approaches, together with the AUC are shown in Figure 2. The achieved accuracy and recall are at the level of 74% while the accuracy obtained for the naive fine-tuning and training from scratch are 57% and 55% respectively.

TABLE I

THE AVERAGED CONFUSION MATRIX FOR THE VALIDATION SET.

		Predicted	
		Grade I	Grade II/III
Actual	Grade I	19.2 ± 2.0	6.6 ± 2.1
	Grade II/III	2.4 ± 1.2	6.8 ± 1.2

TABLE II

CLASSIFICATION SUMMARY FOR THE VALIDATION SET BY AVERAGING RESULTS FROM DIFFERENT FOLDS.

	Precision	Recall	Support
Grade I	0.89 ± 0.04	0.74 ± 0.08	~ 25
Grade II/III	0.52 ± 0.07	0.74 ± 0.12	~ 10
Accuracy:		**0.74 ± 0.04**	

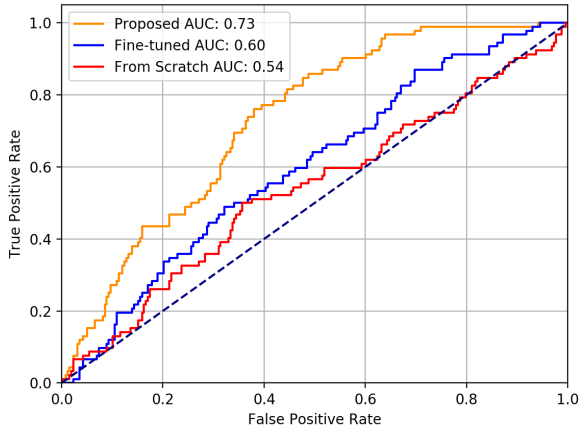

Fig. 2. The accumulated ROC curves of the techniques are shown. The positive class is represented by the high-grade meningiomas.

IV. DISCUSSION AND CONCLUSION

The results show that with a proper, careful training procedure, the classification using deep learning on a highly heterogeneous and relatively small data set can be successful and provide better results than training from scratch or naive fine-tuning. The outcomes are comparable to the results presented in the study [14] where a traditional machine learning approach was used, with time-consuming manual feature extraction and logistic regression. Noteworthy, in the mentioned study, the data set was acquired in a controlled setting in a single medical center, using the same scanner and acquisition protocol, which makes it easier to learn the model.

In this study, we addressed the meningioma grading. However, the same problems occur in practice for many other medical tasks. The problem of very limited, heterogeneous data sets with a strong class imbalance is common. Therefore, the proposed approach to use deep learning for such problems may find its relevance also in other domains.

In future work, we plan to combine several MRI modalities to further improve the results. In general, the T1-weighted post-contrast seems not to be the best modality to do the meningioma grading. Recent studies have shown that the ADC maps are better at differentiating the tumors [21]. However, the meningioma segmentation in ADC is difficult. Therefore, a proper combination of these modalities, the post-contrast T1 where the tumors are easily segmented, with the ADC maps where the grading can be done more reliably, together with registering them together, can lead to even better grading outcomes.

In conclusion, we presented a method based on deep learning to perform meningioma grading, using small and strongly heterogeneous dataset. We showed that by using proper transfer learning, together with a strong data augmentation, carefully choosing the region of interest and proper training scheme, it is possible to achieve reasonable results. The outcomes are comparable to the traditional machine learning approach, without the necessity to use expert knowledge to manually extract the differentiating features.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] R. Goldbrunner *et al.*, "Eano guidelines for the diagnosis and treatment of meningiomas," *The Lancet Oncology*, vol. 17, no. 9, pp. 383–391, 2016.

[3] J. Wiemels, M. Wrensch, and E. Claus, "Epidemiology and etiology of meningioma," *Journal of Neuro-Oncology*, vol. 99, no. 3, pp. 307–314, 2010.

[4] J. Shi *et al.*, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, 2016.

[5] Y. Fujisawa *et al.*, "Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis," *British Journal of Dermatology*, vol. 180, no. 2, pp. 373–381, 2019.

[6] H.-W. Ng, V. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," *Proceedings of the ACM ICMI 2015*, pp. 443–449, 2015.

[7] K. Pasupa and W. Sunhem, "A comparison between shallow and deep architecture classifiers on small dataset," *Proceedings of ICITEE 2016*, 2017.

[8] G. Chandrarathne, K. Thanikasalam, and A. Pinidiyaarachchi, "A comprehensive study on deep image classification with small datasets," *Proceedings of ICCEE 2019*, vol. 619, pp. 93–106, 2020.

[9] D. Rueda-Plata, R. Ramos-Pollán, and F. González, "Supervised greedy layer-wise training for deep convolutional networks with small datasets," *Proceedings of ICCCI 2015*, vol. 9329, pp. 275–284, 2015.

[10] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, 2016.

[11] K. Laukamp *et al.*, "Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric mri," *European Radiology*, vol. 29, no. 1, pp. 124–132, 2019.

[12] T. Banzato *et al.*, "Accuracy of deep learning to differentiate the histopathological grading of meningiomas on mr images: A preliminary study," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 4, pp. 1152–1159, 2019.

[13] T. Banzato, G. Cherubini, M. Atzori, and A. Zotti, "Development of a deep convolutional neural network to predict grading of canine meningiomas from magnetic resonance images," *Veterinary Journal*, vol. 235, pp. 90–92, 2018.

[14] X. Li *et al.*, "Meningioma grading using conventional mri histogram analysis based on 3d tumor measurement," *European Journal of Radiology*, vol. 110, pp. 45–53, 2019.

[15] Y. Zhu *et al.*, "A deep learning radiomics model for preoperative grading in meningioma," *European Journal of Radiology*, vol. 116, pp. 128–134, 2019.

[16] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE CVPR*, vol. 2016, pp. 770–778, 2016.

[18] G. Huang, Z. Liu, and K. Weinberger, "Densely Connected Convolutional Networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[19] C. Szegedy *et al.*, "Going deeper with convolutions," *Proceedings of the IEEE CVPR*, vol. 07-12-June-2015, pp. 1–9, 2015.

[20] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A Survey of Optimization Methods from a Machine Learning Perspective," *CoRR*, vol. abs/1906.06821, 2019. [Online]. Available: http://arxiv.org/abs/1906.06821

[21] R. Huang *et al.*, "Imaging and diagnostic advances for intracranial meningiomas," *Neuro-Oncology*, vol. 21, pp. 44–61, 2019.