MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220163

Noninvasive Glioma Grading with Deep Learning: A Pilot Study

Gleb Danilov[a], Vladislav Korolev[b], Michael Shifrin[a], Eugene Ilyushin[b], Narek Maloyan[b], Daniel Saada[b], Timur Ishankulov[a], Ramin Afandiev[a], Alexander Shevchenko[a], Tatyana Konakova[a], Tatyana Tsukanova[a], Svetlana Shugay[a], Igor Pronin[a], Alexander Potapov[a]

[a] *Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation,*
[b] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russian Federation*

Abstract

Gliomas are the most common neuroepithelial brain tumors, different by various biological tissue types and prognosis. They could be graded with four levels according to the 2007 WHO classification. The emergence of non-invasive histological and molecular diagnostics for nervous system neoplasms can revolutionize the efficacy and safety of medical care and radically reduce healthcare costs. Our pilot study aimed to evaluate the diagnostic accuracy of deep learning (DL) in subtyping gliomas by WHO grades (I-IV) based on preoperative magnetic resonance imaging (MRI) from Burdenko Neurosurgery Center's database. A total of 707 MRI studies was included. A "3D classification" approach predicting tumor type for the entire patient's MRI data showed the best result (accuracy = 83%, ROC AUC = 0.95), consistent with that of other authors who used different methodologies. Our preliminary results proved the separability of MR T1 axial images with contrast enhancement by WHO grade using DL.

Keywords:

Magnetic Resonance Imaging, Tumor Grading, Deep Learning.

Introduction

Gliomas are the most common neuroepithelial brain tumors, different by various biological tissue types and prognosis. A widely used classification of central nervous system tumors shaped by the World Health Organization (WHO) in 2007 was based on histogenesis and microscopic similarities of neoplasms [4]. According to the 2007 WHO classification, gliomas are graded with four levels [4]. WHO grades I-II are typically referred by experts to low-grade gliomas (LGG), WHO grades III-IV – to high-grade (HGG). This classification was revisited in 2016: the new system appeared more complex and relied on molecular subtypes associated with treatment outcomes [5]. Despite the rationale behind the new classification, its application is costly, requiring expensive analyzes. That is the reason the old WHO grade classification is still in use, especially in countries with limited resources.

The emergence of non-invasive histological and molecular diagnostics for nervous system neoplasms can revolutionize the efficacy and safety of medical care and radically reduce healthcare costs. The current state of artificial intelligence applications in neuroimaging research opens up a promising perspective in that direction [3]. The scientific results in radiomics-based brain malignancy typing are also inspirative [6]. The introduction of this quantitative neuroimaging analysis derived new diagnostic, prognostic and predictive tools and advocated for using them to enhance a personalized tumor diagnosis and management plan design in neurooncology [2]. Thus, radiomics, involving MRI as the most common imaging for brain tumors, is in the focus of researchers. However, most data-driven studies in neurooncology are limited with the amount of data available, which prevents the confident application of "big data" and deep learning approaches.

N.N. Burdenko Neurosurgery Center (Moscow, Russia) – is a leading neurosurgical facility in Russia and one of the biggest in the world. Over 20-year exploiting electronic health records and PACS, Neurosurgery Center has accumulated a large archive of nervous system images, which are usable in radiomics research. Our pilot study aimed to evaluate the diagnostic accuracy of deep learning (DL) in subtyping gliomas by WHO grades (I-IV) based on preoperative magnetic resonance imaging (MRI) from Burdenko Neurosurgery Center's database.

Methods

The primary dataset for pilot machine learning experiments was obtained from Burdenko Neurosurgery Center's PACS for 1,280 patients with glial tumors and WHO grade verified by morphological descriptions who underwent preoperative MRI scanning in a period between 2009 and 2018. The data were MR (DICOM) images in series of brain sections identified for each patient, collected on several devices in different modes. Exploratory data analysis revealed inconsistency in the number of slices between patients, even in one modality (ranged 18 to 30 images). A greater difference in slice number appeared for different modes. The sets of MRI modalities also differed between patients. We decided to use only T1 axial images with contrast enhancement, since they were found in most subjects (n = 707) and were collected primarily on the same device. However, the remaining difference in scanning devices influenced the selected data to differ in a certain number of sections and image resolution.

We attempted to solve the technical task of classifying the patient's neuroimaging data (a series of MR images from a single study) by four classes (WHO grades): I, II, III, and IV. To accomplish this goal, two basic approaches were proposed. The

first one - a "3D classification" – predicted tumor type for the entire patient's MRI data as a single object. This approach exploited 707 volumetric images (each collected from the series of patient preoperative slices). We advocated for the first approach as no manual slice labeling was done before the experiments. The second - a "2D classification" – was a separate prediction of tumor type for each specific slice in 17,730 images (an average of approximately 25 slices for each of 707 patients). The dataset appeared to be almost balanced across the glioma WHO grades I, II, III, IV (189, 133, 127, and 258 cases, respectively).

Data preprocessing

In the 3D classification approach, we applied the following data processing methods to normalize and unify each series in quantity. The selected DICOM images were read and formatted as numeric arrays. These numbers reflected the pixel values of the source images. The resulting values were normalized by subtracting the average of the array elements and dividing by the standard deviation:

$$arr = \frac{arr - arr.mean()}{arr.std()}$$

The resolution of each MR slice was scaled to 512 x 512. For all patients, the number of slices was adjusted to 32 using the Area interpolation algorithm. Thus, patients with fewer slices received additional images, while for patients with extra slices, some adjacent sections were averaged.

In the case of 2D classification, image preprocessing was different due to further use. The MR image from one slice was also read into a numeric array. The original single-channel image was transformed to three-channel, with the initial values duplicated into each channel. That approach did not violate the image structure but was required for feeding into the learning pipeline. Transformations were applied to adjust images to a single size 512 × 512, and normalize them with mean = (0.485, 0.456, 0.406), and standard deviation = (0.229, 0.224, 0.225). These values were derived from the ImageNet dataset to apply for pre-trained models. We used several augmentation techniques with a probability of 0.2: rotation by an angle <= 10 degrees, scaling transformation 0.9-1.1 of the original image, and image mirroring to increase the train size artificially.

Training and testing split

In the 3D and 2D cases, the data were divided into training, validation, and testing samples using the split function with a class balance maintained. The training dataset was kept as 80% of the original. Both the validation and texting samples were 10% of the total sampled data.

Model training

The main deep learning model in the case of 3D was the DenseNet architecture adapted to process 3D images. In the 2D case, the Resnest200e architecture was used. Adam optimizer with a learning rate of 1e-4 was applied for all the models. Cross-entropy was chosen as the loss function, which implemented the following formula, where P was the distribution of true response, Q was the probability distribution of model predictions, x was the patient's image:

$$H(P,Q) = -\sum_{x} P(x) log Q(x)$$

The training was carried out both for the classification of glial tumors into all 4 types of malignancy, and for the binary case, when only low (I and II) and high (III and IV) grades were separated.

The deep learning was performed on 8 NVIDIA A100-SXM4-40GB GPU in both 3D and 2D cases. The calculations were parallelized. The pipelines were scripted with Python programming language (version 3.8) using PyTorch library (version 1.7).

Results

The results of binary classification into LGG and HGG with 3D and 2D approaches are shown in Table 1.

Table 1 – The binary classification quality metrics for 3D and 2D approaches

Metric	3D (DenseNet)	2D (Resnest200e)
Accuracy	67%	61%
ROC AUC	76%	73%
Sensitivity	58%	44%
Specificity	78%	81%

Four-level classification resulted in higher indicators for the 3D approach demonstrated in Table 2. The overall accuracy was calculated as the ratio of correct predictions to the number of objects classified (MRI studies in 3D design and slices in 2D case). ROC AUC was the average of the ROC AUCs for all pairwise classifications "one against all." The ROC AUC was first calculated for the binary task: (I vs. II, III, IV), then (II vs. I, III, IV), etc. The resulting ROC AUCs were averaged.

Table 2 – The multinomial classification quality metrics for 3D and 2D approaches

Metric	3D (DenseNet)	2D (Resnest200e)
Accuracy	83%	50%
ROC AUC	95%	72%

The by-class classification metrics for 3D and 2D paradigms are summarized in Tables 3 and 4, respectively.

Table 3 – multinomial classification quality metrics for 3D approach

WHO grade	Precision	Recall	F1-Score
I	0.79	1.00	0.88
II	0.97	0.63	0.76
III	0.50	1.00	0.67
IV	0.95	0.85	0.90

Table 4 – multinomial classification quality metrics for 2D approach

WHO grade	Precision	Recall	F1-Score
I	0.60	0.56	0.58
II	0.11	0.45	0.17
III	0.02	0.32	0.04
IV	0.85	0.47	0.61

Discussion

Defining the type of glioma is important for an early treatment plan design and prognosis assessment. However, the exact diagnosis determining the treatment modality comes from an invasive biopsy which is unsafe to a certain extent. The substitution of surgical procedures with non-invasive diagnostic methods should significantly improve patient safety and reduce time to deliver the most effective care, which is also crucial.

N.N. Burdenko Neurosurgery Center has a large collection of glioma surgery cases documented in medical information systems. More than a thousand operations on glial tumors are performed annually. That is why the expertise in glioma surgery is rich at our Center, and the data we accumulated for two decades may be secondarily used for deep learning experiments. In this study, we present our initial experience in automatic differentiating of WHO grades for glial tumors – the task intuitively performed by clinicians as a preliminary judgment.

The efficiency of machine learning in subtyping gliomas reported in the literature does not contradict our preliminary results. H. Cho et al. (2018) have shown the potential of three models (logistic regression, support vector machine and random forest) to distinguish LGGs from HGGs with an average area under the curve (AUC) of 0.9030 for the test cohort and an average of 88% accuracy, 95% sensitivity and 70% specificity [1,2]. The authors analysed MRI data from 285 patients with brain tumors using T1-weighted, T1-contrast enhanced, T2-weighted and FLAIR MRI. Feature extraction was done with minimal redundancy maximum relevance algorithm.

C. Su et al. (2019) explored the feasibility and diagnostic performance of radiomics based on anatomical, diffusion and perfusion MRI in differentiating glioma subtypes and predicting tumour proliferation [7]. The best AUC reached 0.896 for grades II-III, 0.997 for grades II-IV, and 0.881 for grades III-IV.

The above-mentioned studies were primarily focused on feature engineering. Z. Ning et al. (2021) attempted to combine local MRI features with deep feature extraction using a convolutional neural network (CNN) to develop a noninvasive glioma grading model [6]. The authors report the AUC, sensitivity, and specificity of the model based on a combination of radiomics and deep features were 0.88 (95% CI: 0.84, 0.91), 88% (95% CI: 80%, 93%), and 81% (95% CI: 76%, 86%), respectively, for the testing cohort. They stated that the developed model outperformed the models based only on either radiomics or deep features ($p<0.001$), and was also comparable to the clinical radiologists.

In our studies, a sole deep learning approach with well-proven models for image classification was utilized.

According to the literature and our initial experience, the separation of LGG from HGG should be more solvable than differentiation WHO grades inside LGG or HGG. That is reasonable since an experienced physician usually sees the differences between benign and malignant neoplasms of the brain on MRI data in typical cases. Table I demonstrates the confidence of the model in the diagnosis of WHO grade I and IV, and less accuracy in WHO grade II-III subtyping. That is consistent with the clinician's experience in assessing the types of gliomas "by eye."

Nowadays, we gain evidence that the application of machine learning to preoperative MRI demonstrates promising results for predicting IDH mutation, MGMT methylation, and 1p/19q codeletion in glioma [3]. We hope that our pilot study serves as a basis to contribute to research in computer-aided glioma diagnosis.

Nevertheless, we believe data scientists and neuroscientists do not have enough arguments to expect a non-invasive "biopsy" coming from only one neuroimaging modality. It is much more likely that the highest quality of non-invasive diagnostics with deep learning can be obtained with simultaneous usage of various neuroimaging modalities: different MRI modes, including spectroscopy, CT, including perfusion, and PET. Collecting such a complex dataset presents significant challenges. However, the development of non-invasive tumor subtyping technologies within individual modalities will create the basis for future complex solutions and enable a better understanding of which methods appear promising, which tumor signatures are divided better. At least, that information may be supportive in clinical decision-making. This is how the authors of this article see the prospects for deep learning development in neuroimaging diagnostics.

The limitation of our study was still a sample size, data inconsistency in terms of format and quantity. However, the experimental settings were close to what practitioners are faced with in real-world routine. In a 2D approach, no manual by-slice labeling was applied, so the slices with no tumor on them were inevitably and erroneously labeled with the WHO grade a patient has. That could certainly influence the quality of classification. However, we accept this limitation in the pilot study since the manual labeling requires much more effort and may be shifted to the next stage. It was principally important for us to verify our neuroimaging data's separability by WHO grade and relate it with the results of other researchers, especially when obtained with a different methodology. Tumor delineation may also be considered as an additional preprocessing step.

Conclusions

Our preliminary results prove the principal separability of MR T1 axial images with contrast enhancement by WHO grade using DL models. The diagnostic accuracy of DL in glioma subtyping is expected to improve through adding modalities, testing new methodologies and image pre- and post-processing methods.

Acknowledgements

The dataset preparation was done with financial support from the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-15-2020-809 (alt. 13.1902.21.0030). Data preprocessing was supported by Russian Foundation for Basic Research (grant 18-29-01052).

References

[1] H.-H. Cho, S.-H. Lee, J. Kim, and H. Park, Classification of the glioma grading using radiomics analysis, (2018). doi:10.7717/peerj.5982.

[2] A. Habib, N. Jovanovich, M. Hoppe, M. Ak, P. Mamindla, R.R. Colen, and P.O. Zinn, MRI-Based Radiomics and Radiogenomics in the Management of Low-Grade Gliomas: Evaluating the Evidence for a Paradigm Shift, *J. Clin. Med.* (2021) 10. doi:10.3390/jcm10071411.

[3] A. Jian, K. Jang, M. Manuguerra, S. Liu, J. Magnussen, A. Di Ieva, and M. Neurosurgery, Machine Learning for the Prediction of Molecular Markers in Glioma on Magnetic Resonance Imaging: A Systematic Review and Meta-Analysis, *Neurosurgery*. (2021). doi:10.1093/neuros/nyab103.

[4] D.N. Louis, H. Ohgaki, · Otmar, D. Wiestler, W.K. Cavenee, P.C. Burger, A. Jouvet, B.W. Scheithauer, and P. Kleihues, The 2007 WHO Classification of Tumours of the Central Nervous System, *Acta Neuropathol*. **114** (2007) 97–109. doi:10.1007/s00401-007-0243-4.

[5] D.N. Louis, A. Perry, · Guido Reifenberger, A. Von Deimling, D. Figarella-Branger, · Webster, K. Cavenee, H. Ohgaki, · Otmar, D. Wiestler, P. Kleihues, · David, and W. Ellison, The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary, *Acta Neuropathol.* (n.d.). doi:10.1007/s00401-016-1545-1.

[6] Z. Ning, J. Luo, Q. Xiao, L. Cai, Y. Chen, X. Yu, J. Wang, and Y. Zhang, Multi-modal magnetic resonance imaging-based grading analysis for gliomas by integrating radiomics and deep features, *Ann Transl Med.* **9** (2021). doi:10.21037/atm-20-4076.

[7] C. Su, J. Jiang, S. Zhang, J. Shi, K. Xu, N. Shen, J. Zhang, L. Li, L. Zhao, J. Zhang, Y. Qin, Y. Liu, and W. Zhu, Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour, *Eur. Radiol.* **29** (2019) 1986–1996. doi:10.1007/s00330-018-5704-8.

Address for correspondence

Corresponding author: Gleb Danilov, e-mail: glebda@yandex.ru; phone: +7 (916) 775-11-48