The Color of Time: Detecting Glioma IDH Mutation Status in MRI Through Pseudo-Colored Transfer Learning

Hamish MacKinnon[1,2,3], Silvia Paracchini[3], David Harris-Birtill[2], John Hipwell[1], and Keith Goatman[1]

[1] Canon Medical Research Europe, Edinburgh, Scotland, UK
[2] University of St Andrews School of Computer Science, St Andrews, Scotland, UK
[3] University of St Andrews School of Medicine, St Andrews, Scotland, UK

Abstract.

Background: Glioma is the most common brain cancer and is conventionally diagnosed with MR imaging. Its prognosis and treatment depend on the tumor genetic subtype. However, tumor genotyping is invasive, requiring a sample of tumor tissue; a noninvasive method to determine glioma subtype from an image would be a valuable addition to the oncology toolbox. Necessary restrictions on access to clinical data make developing medical applications challenging. Radiogenomics is especially challenging, since it requires paired imaging and genotype data.

Aims: We investigate whether classification models, pre-trained on natural scene images before being finetuned on MR images to determine glioma subtype, can outperform models trained from scratch on larger private medical datasets. We investigate the most effective way of applying the MR sequences to the color model.

Methods: The T1, contrast enhanced T1, T2 and FLAIR sequences (defined by their different repetition, echo and inversion times) are used as inputs to the color channels, allowing the use of preexisting natural scene models. A hyperparameter search determined the optimum parameters. Two pretrained CNN models (VGG16 and ResNext) were finetuned and compared across 24 psudo-color permutations and 4 gray monocolor configurations to explore effects on performance from combinations of MR sequence and color channel.

Results: Our best model exceeds the baseline from literature, achieving 88.1% accuracy, 0.935 AUC and 0.819 F1 score on a held out test set.

Conclusions: Classification of genetic markers in volumetric images can be undertaken effectively and efficiently with models pretrained on 2D natural scene images finetuned for the imaging genomics task. Crafting a custom 3D volumetric model from scratch is not always necessary.

Keywords: Radiogenomics · Transfer Learning · Glioma · IDH mutation · pseudo-color · MRI.

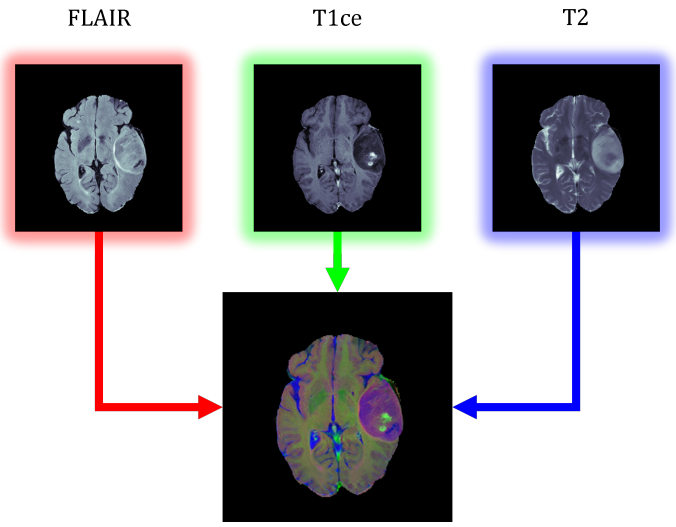

Fig. 1: Illustration of the process creating pseudo-colored composite images from different MRI sequences. This configuration corresponds to permutation 21 in figure 2.

1 Introduction

Glioma is the most common brain cancer. It consists of several subtypes, the identification of which is an important step in determining prognosis: glioblastoma (previously known as *high grade glioma* and *glioblastoma multiforme* or GBM) has a higher mortality rate than astrocytoma or oligodendroglioma (together previously known as *low grade glioma*). Though somewhat visible on a medical image, until recently the differentiation of these subtypes was poorly understood and been subject to changes in nomenclature as the delineation became clearer. The glioma subtypes are strongly associated with genetic level differences within the mutating tumor cells, to the point that these genetic differences are now the official definition of the subtypes in the World Health Organisation (WHO) guide [13]. The specific genetic differences of interest are the mutuation status (i.e. wild-type or mutant) of Isocitrate Dehydrogenase (IDH) — this protein has two paralogue genes, IDH1 and IDH2, of which IDH1 is the gene of interest — and the deletion of both the short arm of chromosome 1 and the long arm of chromosome 19 (succinctly denoted as 1p19q codeletion). In addition to these type-defining differences, research suggests that O6-Methylguanine-DNA-methyltransferase promoter methylation (MGMT methylation) influences treatment response: tumors with methylated MGMT being less likely to respond to standard treatment[9].

Genetic subtyping requires a sample of brain tissue. The standard care pathway for glioma uses the Stupp protocol, which recommends complete surgical resection of the tumor at an early stage, irrespective of subtype. Resected tissue can then

be used for molecular sequencing. However, there are options for acquiring the genotyping information without the full tumor resection:

Biopsy Biopsies are invasive and, despite improved methods such as sterotactic biopsies guided by plans derived from medical imaging, the operation is still associated with around a 1% mortality rate[7].

Blood test Fragments of tumor cells circulating in the blood can be detected by blood tests targeting the IDH1 gene[22]. Recent clinical validation achieved an AUC of 0.84, sensitivity of 70.0% , positive predictive value of 90.9%, and a F1 score of 0.791[5]. This workflow has the advantage that it can be completed in under four hours from sample to result.

MRI Image Analysis Human expert identification of IDH-mutant glioma exists with sophisticated techniques [15][11], but is not treated as a front-line approach and in practice is subordinate to histopathological tissue analysis. Automated IDH1 mutation prediction has been investigated [8][6][21], and is described in more detail below. Image analysis is quick and non-invasive. However, data availability is a major challenge. Although public brain tumor data exists, the genetic data associated with it is far more scarce due to tighter confidentiality requirements. In addition, how detailed does the genetic data have to be to be useful — are genome sequences required, or do mutation panels suffice?

Given this limited availability of relevant data, it is vital the data that is available is used as effectively and efficiently as possible. We propose that transfer learning, the use of models pretrained on different data and objectives to the task at hand, is undervalued in medical image analysis and can be applied effectively with some creativity to achieve better performing models without additional data collection or model size costs.

2 Related Work

Several studies have investigated the identification of derived characteristics from brain MRI images, including IDH mutation, 1p/19q codeletion, and MGMT methylation state [8][6][21]. However, it is difficult to compare published results due to the absence of common evaluation benchmarks: each study uses different evaluation data and data splits. Our work uses [21] as a baseline due to the availablity of the model and code, and that study uses the publicly available TCGA-LGG and TCGA-GBM as its held out test set. These TCGA datasets are publicly available as a subset of the Brain Tumor Segmentation (BraTS) dataset[2][14][3].

The use of RGB color channels as a method of input for MRI sequences to a model has been used for visualization [4] and for detection[12]. However, while these works have been useful for visualizing features of MRI, they have not investigated the practical possibilities of using this method as an accurate classifier.

3 Data

This work follows an example set by van der Voort et al.[21], who reserved public datasets (TCIA glioblastomas and low-grade gliomas, also known as GBM and LGG)[16][17] for their held-out test set. Training, validation, and test data splits are created from the Erasmus Glioma Database (EGD)[20], a dataset released by the same research group[21]. Training, validation, and test splits were created from the 467 out of 768 patients in the EGD collection that had paired genetic IDH and MRI data, with a 70:15:15 split matching that of van der Voort et al., giving 327 training, 70 validation, and 70 test patients.

4 Methods

The method of van der Voort et al.[21] as provided via released code and models is used as a state of the art baseline. This baseline method has been trained on more data from a wider array of datasets than is described in section 3. Some of that is private data, while other data is from datasets that do not include IDH1 mutation data; van der Voort et al. [21] describe a multitask model that can learn for its other tasks when IDH1 data is absent. Where that model used multi-task training to allow more data to be used, our method instead relies on transfer learning from a natural image model, leveraging the pretraining to make up for using a smaller, single task dataset.

The training and evaluation pipeline was built with pytorch-lightning version 2.4.0 [10] using torch 2.5.1 with CUDA version 12.4. A hyperparameter search was performed using Optuna [1]. Since the selection of hyperparameters depends strongly on the status of the experimental interventions, these interventions are added as hyperparameters and allowed to be explored by the hyperparameter search algorithm. Once all experiments have concluded, the best model for each of the desired comparisons is selected and evaluated on the held out test set.

Model To test the effectiveness of pretraining on non-medical, natural scene images we use two models as a base for both training from scratch and finetuning with medical images after natural scene training. The model architectures investigated were ResNext[23] and VGG-16 [18], both from the TIMM pretrained image model repository hosted by Huggingface. VGG is a simple deep convolutional neural network (`timm/vgg16.tv_in1k`); ResNext is a model with residual connections (`timm/resnext101_32x16d.fb_swsl_ig1b_ft_in1k`) pretrained on a very large natural image dataset [24].

Data Processing To prepare images for input to the model, we use the segmentation masks included with the EGD dataset. From the full volume, a single z-axis axial slice is selected from the midpoint of the range of slices annotated as including tumor. The slices at this position are selected from the MRI sequences and composed into a pseudo-color image as in figure 1, with the choice and order of the MR image sequences determined by a model parameter.

Hyperparameter Search The hyperparameter search was performed using the Optuna library[1], with the F1 score for the performance metric. Samples were selected in the hyperparameter space following the default tree-structured Parzen estimator algorithm. The hyperparameters explored, the ranges of these spaces, and the optimal results are shown in table 1. Some hyperparameters were fixed and not varied by the Optuna system: batch size was set to 16; the choice of optimizer was AdamW; and random seeds were 42, 43, and 44.

Permutations The optimal mapping of the four MRI sequences (T1, contrast-enhanced T1, T2, and FLAIR) to the three color channels (RGB) was investigated by comparing every permutation of selecting 3 sequences from 4 (${}^4P_3$). This results in $\frac{4!}{(4-3)!} = 24$ permutations of sequences to evaluate. In addition to these 24 permutations, we are interested in the performance of "monochrome" images, i.e. where the same MRI sequence is given to all three color channels. Together with the permutations this gives a total of 28 configurations.

Hardware The hyperparameter search was undertaken with an NVIDIA RTX A4000 with 16GB VRAM, 128GB RAM, and an Intel Xeon w5-3435X CPU. Permutation training and final test evaluation used a collection of eight NVIDIA L4 GPUs with 24GB VRAM, each permutation using one card to allow parallelized training with 40GB RAM and four 32 core CPUs each. Training takes 4 hours to complete a job of training models for three different random seeds for a single permutation configuration.

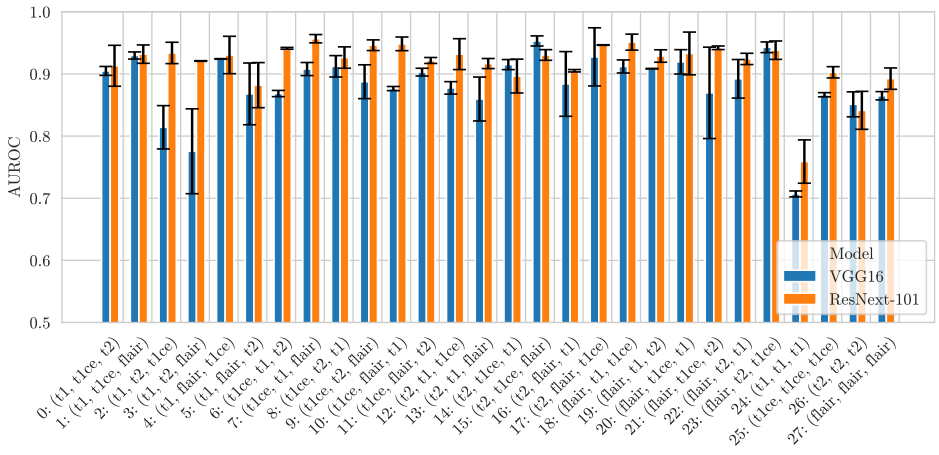

Fig. 2: Classification performance of models finetuned with each of the possible MRI sequence to color channel permutations and the 4 gray combinations. Bars are the average of 3 runs, with error bars showing the standard deviation.

Hyperparameter Search Space

Hyperparameter	Range	Optimized Value
Initial **learning rate**. Logarithmic sampling used to enable exploration of learning rate sensitivity in small ranges.	[6e-6, 6e-4]	1.08e-4
Use **pretrained** weights. If false, use freshly initialized weights.	[True, False]	True
Use the OneCycleLR learning rate scheme (**superconvergence**[19]) if true. Otherwise, use a constant learning rate.	[True, False]	True
Scheduled peak learning rate fraction (i.e. this number times max_epochs) where the learning rate peaks before falling.	[0.01, 0.5]	0.437
Superconvergent **learning rate update interval**, the resolution of learning rate updates. On step is a smoother curve, on epoch treats batches equally.	["step", "epoch"]	epoch
Max epochs for superconvergence, so the lower bound is much lower as faster training is more likely to be viable. Note early stopping is on, so training may end sooner than this.	[10, 100]	40
Max epochs for the **constant LR** scheme. Note early stopping is on, so training may end sooner than this. Used when Superconvergence is off.	[100, 150]	-
Mask non-tumor regions: Use tumor segmentation masks to remove the rest of the head from the image fed to the model.	[True, False]	False

Table 1: The search space defined for the hyperparameter search. Optimized values (rounded) were found after 297 trials using ResNeXt-101 and color map 21 (FLAIR, T1ce, T2).

5 Results

The optimal hyperparameters were retrieved from the hyperparameter search for the permutation (FLAIR, T1ce, T2 — number 21 in figure 2). These hyperparameters were then fixed and the 24 permutations of the MRI sequence-color mapping (plus four gray mappings) were trained and validated. The results for each permutation are shown in figure 2. In both figures 2 and 3 the gray (single sequence) configurations perform worse than the pseudocolored (multi-sequence) configurations. In particular, the all-T1 configuration performs notably worse than all other models.

A winning permutation was selected for test set evaluation by calculating the sum of the mean AUROC for the two models. This method was selected to allow the final models to be compared using a permutation that they were both successful with: this is permutation 15 (T2, T1ce, FLAIR) with a sum-AUROC of 1.88. A test set comparison of the best models versus other methods from literature is shown in table 2.

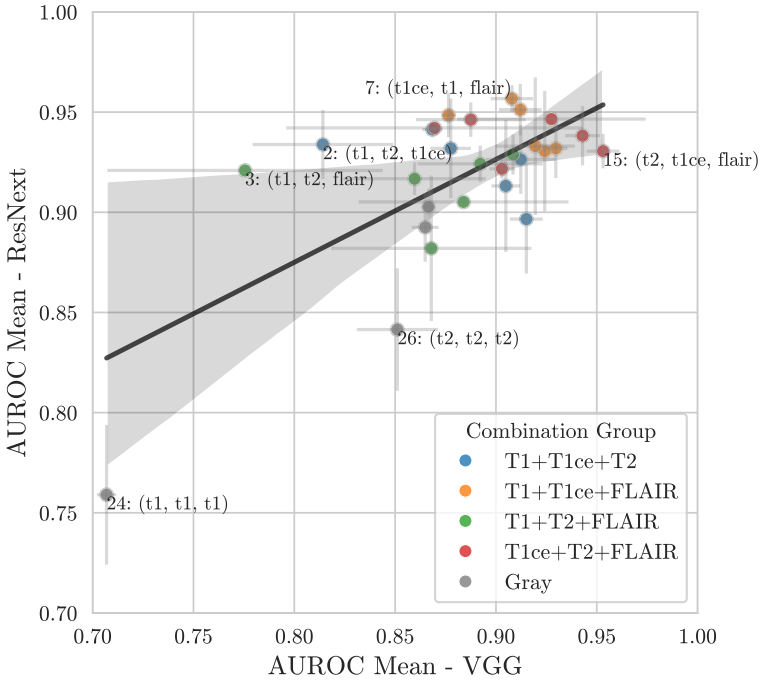

Fig. 3: Validation data comparison between the two models' classification score across permutations. Point colors are grouped according to their constituent sequences. Gray combinations are 24, 25, 26, 27 from figure 2). Data points at the edge of the cluster are labeled. Error bars show the standard deviation from 3 random seeds.

IDH Classification Results

Method	AUROC	Acc	F1	Prec.	Rec./ Sens.	Spec.	Dataset
VGG16 (no pretraining)	0.756	0.690	0.592	0.586	0.603	0.742	EGD test split
Resnext (no pretraining)	0.718	0.710	0.526	0.657	0.449	0.864	EGD test split
(ours) Finetuned VGG16	0.915	0.867	0.797	0.917	0.705	**0.962**	EGD test split
(ours) Finetuned ResNext	**0.935**	**0.881**	**0.819**	**0.939**	**0.731**	0.864	EGD test split
3D Multitask Network [21]	0.9	0.84	0.79†	0.88†	0.72	0.93	TCGA-LGG +TCGA-GBM
Circulating DNA test [5]	0.84	*	0.791†	0.909	0.7	*	[5]

† These results were derived with calculation from published results.
* These values were not possible to calculate from released results.

Table 2: Results for comparable methods of classifying IDH mutation. Our results are reporting on the test set we separated from the EGD dataset described in section 3. Our models are trained using permutation 15 (T2, T1ce, FLAIR) from figures 2 and 3. Acc. is accuracy, Prec. is precision, Rec. is recall, Sens. is sensitivity, Spec. is specificity.

6 Discussion

The results in table 2 confirm some established knowledge: that pretraining is more effective than training from scratch for the provided models for binary classification tasks, and that the finetuned ResNext models outperform finetuned VGG models. However, the results extend this idea to scenarios where the pretraining is performed on a very different problem domain, specifically using natural images as pretraining for pseudo-color MRI. The difference in information between the 16-bit MR sequences and the 8-bit natural image color channels in the models could have been a significant obstacle to training a classifier, but this does not seem to be the case. The comparison between the models with pretraining versus no pretraining indicates the ResNext performing better, which may be a benefit from the much larger pretraining dataset though this could also be due to the ResNext architecture learning more general transfers. The permutation used for hyperparameter search (21) was not the same as the top performing permutation (15), indicating that there may be opportunity for further optimisation. Future research may be able to apply this method in a way that requires less preprocessing of the MR volumes before analysis, which could bring significant time and computation savings. The results also suggest a future direction for medical classification models, away from specialized medical solutions towards methods that are able to leverage wider developments in natural image processing.

Acknowledgments. The authors would like to acknowledge the support of the UKRI Center for Doctoral Training in Applied Photonics.

The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., et al.: Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2623–2631. Association for Computing Machinery (Jul 2019). https://doi.org/10.1145/3292500.3330701, https://doi.org/10.1145/3292500.3330701
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., et al.: RSNA-ASNR-MICCAI-BraTS-2021 (2023). https://doi.org/10.7937/JC8X-9874, https://www.cancerimagingarchive.net/analysis-result/rsna-asnr-miccai-brats-2021/
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (Sep 2017). https://doi.org/10.1038/sdata.2017.117
4. Banerjee, S., Mitra, S., Shankar, B.U., Hayashi, Y.: A Novel GBM Saliency Detection Model Using Multi-Channel MRI. PLOS ONE **11**(1), e0146388 (Jan 2016). https://doi.org/10.1371/journal.pone.0146388, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146388, publisher: Public Library of Science
5. Batool, S.M., Escobedo, A.K., Hsia, T., Ekanayake, E., et al.: Clinical utility of a blood based assay for the detection of IDH1.R132H-mutant gliomas. Nature Communications **15**(1), 7074 (Aug 2024). https://doi.org/10.1038/s41467-024-51332-7, https://www.nature.com/articles/s41467-024-51332-7, publisher: Nature Publishing Group
6. Chakrabarty, S., LaMontagne, P., Shimony, J., Marcus, D.S., et al.: MRI-based classification of IDH mutation and 1p/19q codeletion status of gliomas using a 2.5D hybrid multi-task convolutional neural network. Neuro-Oncology Advances **5**(1), vdad023 (Jan 2023). https://doi.org/10.1093/noajnl/vdad023, https://doi.org/10.1093/noajnl/vdad023
7. Chen, C.C., Hsu, P.W., Erich Wu, T.W., Lee, S.T., et al.: Stereotactic brain biopsy: Single center retrospective analysis of complications. Clinical Neurology and Neurosurgery **111**(10), 835–839 (Dec 2009). https://doi.org/10.1016/j.clineuro.2009.08.013, https://www.sciencedirect.com/science/article/pii/S0303846709002285
8. Decuyper, M., Bonte, S., Deblaere, K., Van Holen, R.: Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma. Computerized Medical Imaging and Graphics **88**, 101831 (Mar 2021). https://doi.org/10.1016/j.compmedimag.2020.101831, https://www.sciencedirect.com/science/article/pii/S0895611120301269
9. Do, D.T., Yang, M.R., Lam, L.H.T., Le, N.Q.K., et al.: Improving MGMT methylation status prediction of glioblastoma through optimizing radiomics features using genetic algorithm-based machine learning approach. Scientific Reports **12**(1), 13412 (Aug 2022). https://doi.org/10.1038/s41598-022-17707-w, https://www.nature.com/articles/s41598-022-17707-w, publisher: Nature Publishing Group
10. Falcon, W., Borovec, J., Harris, E., Painter, M., et al.: PyTorch Lightning (2019), https://github.com/PyTorchLightning/pytorch-lightning
11. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., et al.: RadGraph: Extracting Clinical Entities and Relations from Radiology Reports (Aug 2021), http://arxiv.org/abs/2106.14463, arXiv:2106.14463 [cs]

12. Kalaiselvi, T., Kumarashankar, P., Sriramakrishnan, P., Karthigaiselvi, S.: Brain Tumor Detection from Multimodal MRI Brain Images using Pseudo Coloring Processes. Procedia Computer Science **165**, 173–181 (2019). https://doi.org/10.1016/j.procs.2020.01.094, https://linkinghub.elsevier.com/retrieve/pii/S1877050920301022
13. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., et al.: The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology **23**(8), 1231–1251 (Aug 2021). https://doi.org/10.1093/neuonc/noab106, https://doi.org/10.1093/neuonc/noab106
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (Oct 2015). https://doi.org/10.1109/TMI.2014.2377694, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833122/
15. Patel, S.H., Poisson, L.M., Brat, D.J., Zhou, Y., et al.: T2–FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project. Clinical Cancer Research **23**(20), 6078–6085 (Oct 2017). https://doi.org/10.1158/1078-0432.CCR-17-0560, https://doi.org/10.1158/1078-0432.CCR-17-0560
16. Pedano, N., Flanders, A.E., Scarpace, L., Mikkelsen, T., et al.: The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) (2016). https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK, https://www.cancerimagingarchive.net/collection/tcga-lgg/
17. Scarpace, L., Mikkelsen, T., Cha, S., Rao, S., et al.: The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM) (2016). https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9, https://www.cancerimagingarchive.net/collection/tcga-gbm/
18. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Apr 2015). https://doi.org/10.48550/arXiv.1409.1556, http://arxiv.org/abs/1409.1556, arXiv:1409.1556 [cs]
19. Smith, L.N., Topin, N.: Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates (May 2018). https://doi.org/10.48550/arXiv.1708.07120, http://arxiv.org/abs/1708.07120, arXiv:1708.07120 [cs, stat]
20. van der Voort, S.R., Incekara, F., Wijnenga, M.M.J., Kapsas, G., et al.: The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. Data in Brief **37**, 107191 (Aug 2021). https://doi.org/10.1016/j.dib.2021.107191, https://www.sciencedirect.com/science/article/pii/S2352340921004753
21. van der Voort, S.R., Incekara, F., Wijnenga, M.M.J., Kapsas, G., et al.: Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. Neuro-Oncology **25**(2), 279–289 (Feb 2023). https://doi.org/10.1093/neuonc/noac166, https://doi.org/10.1093/neuonc/noac166
22. Wang, J., Zhao, Y.y., Li, J.f., Guo, C.c., et al.: IDH1 mutation detection by droplet digital PCR in glioma. Oncotarget **6**(37), 39651–39660 (Oct 2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741852/
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., et al.: Aggregated Residual Transformations for Deep Neural Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5987–5995 (Jul 2017). https://doi.org/10.1109/CVPR.2017.634, https://ieeexplore.ieee.org/document/8100117
24. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., et al.: Billion-scale semi-supervised learning for image classification (May 2019). https://doi.org/10.48550/arXiv.1905.00546, http://arxiv.org/abs/1905.00546, arXiv:1905.00546 [cs]