Neuro-Oncology Advances

Pediatric brain tumor classification using deep learning on MR images with age fusion

Iulian Emil Tampu[†,©], Tamara Bianchessi[†,©], Ida Blystad[©], Peter Lundberg[©], Per Nyman[©], Anders Eklund[©]and , Neda Haj-Hosseini[©]

All author affiliations are listed at the end of the article
[†]These authors contributed equally to this work.

Corresponding Author: Iulian Emil Tampu, Department of Biomedical Engineering, Campus US, Linköping University, 581 85 Linköping, Sweden (iulian.emil.tampu@liu.se).

Abstract

Purpose. To implement and evaluate deep learning-based methods for the classification of pediatric brain tumors (PBT) in magnetic resonance (MR) data.

Methods. A subset of the "Children's Brain Tumor Network" dataset was retrospectively used ($n = 178$ subjects, female = 72, male = 102, NA = 4, age range [0.01, 36.49] years) with tumor types being low-grade astrocytoma ($n = 84$), ependymoma ($n = 32$), and medulloblastoma ($n = 62$). T1w post-contrast ($n = 94$ subjects), T2w ($n = 160$ subjects), and apparent diffusion coefficient (ADC: $n = 66$ subjects) MR sequences were used separately. Two deep learning models were trained on transversal slices showing tumor. Joint fusion was implemented to combine image and age data, and 2 pre-training paradigms were utilized. Model explainability was investigated using gradient-weighted class-activation mapping (Grad-CAM), and the learned feature space was visualized using principal component analysis (PCA).

Results. The highest tumor-type classification performance was achieved when using a vision transformer model pre-trained on ImageNet and fine-tuned on ADC images with age fusion (Matthews correlation coefficient [MCC]: 0.77 ± 0.14, Accuracy: 0.87 ± 0.08), followed by models trained on T2w (MCC: 0.58 ± 0.11, Accuracy: 0.73 ± 0.08) and T1w post-contrast (MCC: 0.41 ± 0.11, Accuracy: 0.62 ± 0.08) data. Age fusion marginally improved the model's performance. Both model architectures performed similarly across the experiments, with no differences between the pre-training strategies. Grad-CAMs showed that the models' attention focused on the brain region. PCA of the feature space showed greater separation of the tumor-type clusters when using contrastive pre-training.

Conclusion. Classification of PBT on MR images could be accomplished using deep learning, with the top-performing model being trained on ADC data, which radiologists use for the clinical classification of these tumors.

Key Points

- The vision transformer model fine-tuned on ADC data with age fusion outperformed those trained on T2w (second-best) and T1w-Gd data.

- Age information slightly improved classification, while model architecture and pre-training strategies had no significant impact.

- Explainability analysis reveals that models use tumor region data for classification, and age information enhances tumor-type separation.

Tumors in the central nervous system are the second most common type of cancer in children and young adults up to the age of 19, with an estimated age-standardized rate (in 100 000 population) of 1.2 for incidence and 0.60 for mortality worldwide,[1] where brain tumors account for about 57% of the total causes of cancer deaths in this population.[2] Pediatric brain tumors (PBT) can be grouped concerning the location relative to the tentorium, as infratentorial or supratentorial.

Importance of the Study

Deep learning-based classification of pediatric brain tumors can be achieved using single-sequence pre-operative magnetic resonance data, showing the potential of automated decision support tools that can aid radiologists in the primary diagnosis of these tumors.

Tumors in the infratentorial brain region (posterior fossa) are more common in pediatric patients; however, the frequency varies depending on age.[3–5] Brain tumor treatment procedures are usually complicated where tumor detection and preliminary diagnosis are based on magnetic resonance images (MRI), and treatment planning also uses histopathological and molecular analysis of the tissue sample.[6] Diagnosis by radiologists, when comparing the first MRI diagnosis to the final histology diagnosis, varies greatly among tumor types and locations, with an overall sensitivity of 72% for broad tumor-type classification (range 0%–100%), which shows the need for computational methods to improve qualitative assessments.[7] Deep learning algorithms have been successfully applied to several medical image-related tasks and can be trained to assist radiologists in diagnosing brain tumors based on MR. Even though deep learning methods have led to reasonable advancements in adult brain tumor detection, classification, and segmentation,[8–10] their implementation in pediatric cases has been limited[11,12] mainly due to the lack of large and standardized open-access datasets.[13,14] Deep learning models trained on MR images from adults will not perform well on images from children, since PBTs have different diagnostic properties. The "Children's Brain Tumor Network" (CBTN)[15,16] is one of the largest PBT datasets, and could potentially be used in the future similarly to the adult brain tumor segmentation challenge (BraTS),[17–19] as a standard and reference dataset for development and comparison of deep learning methods. This study is, to the best of our knowledge, the first report on the implementation of deep learning on the MR dataset from CBTN for brain tumor classification, and also one of few hitherto published MR-based deep learning studies on any brain tumor pediatric dataset.

This exploratory study aimed to investigate deep learning-based methods for the classification of PBT, considering different pre-operative MRI sequences, and fusing age and image information. A convolutional neural network (ResNet50) and a vision transformer (ViT) were implemented and evaluated, exploring 3 pre-training strategies, and investigating model explainability by visualization of activation maps and the learned feature space.

Materials and Methods

Dataset Cohort

In this retrospective study, the open-access dataset was obtained upon application to and approval from CBTN[15] (accessed in 2021). The downloaded dataset contained 326 subjects, with tumor-type information available for 273 subjects (females = 153, males = 116, not available = 4,

age range [0.01, 36.49] years). Patients older than 18 years ($n = 3$) were included in the dataset given the pediatric tumor-type diagnosis. The tumor types available were low-grade astrocytoma (ASTR) ($n = 132$), medulloblastoma (MB) ($n = 67$), ependymoma (EP) ($n = 45$), atypical teratoid rhabdoid tumor ($n = 20$), diffuse intrinsic pontine glioma (DIPG) ($n = 6$), ganglioglioma ($n = 1$), germinoma ($n = 1$), and teratoma ($n = 1$). Due to the limited number of subjects for DIPG, ganglioglioma, germinoma, and teratoma categories, these tumor types were excluded from the subsequent analysis. From the remaining tumor-type groups, T1w-Gd, T2w and diffusion-weighted (DW) MR data were collected and used in the analysis.

Data Selection and Exclusion

An automated selection based on the image quality followed by a visual assessment was performed. Quality selection for T1w-Gd and T2w data was based on the voxel resolution, removing those with an axial in-plane resolution larger than 1 mm, and with less than 50 axial slices. These values were chosen to avoid artifacts due to a low image resolution. For the DW data, scans with less than 6 diffusion-encoding directions were excluded. Visual assessment of all individual images was performed and data were excluded if: (1) images were acquired post-operatively, (2) images showed the spine only, (3) the tumor was not visible, (4) the transversal plane had been clipped, and (5) image artifacts (motion, metal, induced by neurosurgical clips) were present. By visual assessment, the tumor location was saved as a boundary box.

Pre-Processing of Image and Age Data

The DW-MR data were processed using MRtrix3 software[20] to obtain diffusion tensors from which the ADC map was calculated. Brain extraction was performed, followed by data harmonization, using a per-sequence voxel intensity normalization and interpolation down to 1 mm isotropic resolution. These steps were performed since the CBTN dataset was collected on a variety of MR scanners (manufacturer, field strength, gradient performance, etc.).[16] The final volumes were reshaped to have 224 × 224 pixels in the transverse plane. Transversal 2D slices positioned within 20%–80% of the tumor boundary box were extracted from the volumetric data to ensure that images showing only small portions of the tumor were not included. Transversal slices were used instead of the volumetric data due to the limited number of subjects available for training 3D deep learning models. A detailed description of the pre-processing steps and software used is available in the Supplementary materials. The age in days of

Table 1. Per tumor type and MR sequence summary of the dataset. The age information was obtained from the earliest scan available for each subject. The number of extracted slices reflects the slices within 20-80% of the tumor boundary box. m: mean, std: standard deviation, M/F: male/female, NA: not available, ASTR: astrocytoma, EP: ependymoma, MED: medulloblastoma

Tumor type	Subjects (infra/supra/both)	Sex (M/F/NA)	Age in years median [range] (m ± std)	Subjects (infra/supra/both) *Extracted slices*		
				T2-w	T1w-Gd	ADC
ASTR	84 (38/43/3)	48/34/2	7.76 [0.53, 20.08] (8.13 ± 4.84)	77 (34/41/2) *1643*	38 (21/16/1) *628*	38 (16/19/3) *726*
EP	32 (19/10/3)	19/12/1	4.44 [0.00, 22.85] (6.35 ± 5.74)	30 (17/10/3) *800*	17 (9/6/2) *263*	9 (5/3/1) *287*
MB	62 (62/0/0)	35/26/1	8.95 [0.24, 36.49] (9.02 ± 5.84)	53 (53/0/0) *969*	39 (39/0/0) *622*	19 (19/0/0) *370*
Total	178 (119/53/6)	102/72/4	7.24 [0.00, 36.49] (8.13 ± 5.44)	160 (104/51/5) *3412*	94 (69/22/3) *1513*	66 (40/22/4) *1383*

each subject from the earliest available scan was obtained from the CBTN portal, converted in years, and normalized using z-score normalization using the [0.5th, 99.5th] value range. The final composition of the dataset with age and sex information is summarized in Table 1.
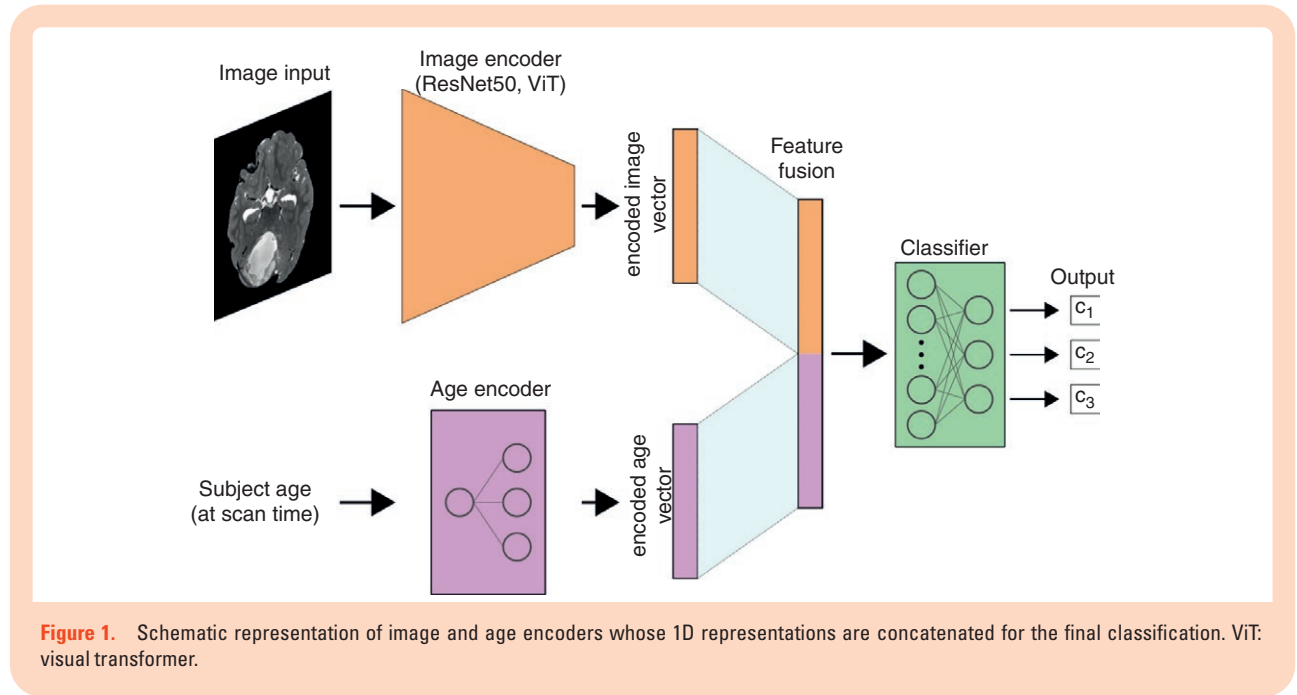
Network Architecture and Training

Two deep learning model architectures extensively used in literature were employed in this study, distinguished by their feature extraction approach: ResNet50[21] and Vision Transformer[22] (ViT), in its base 16 version. ResNet50[21] is a deep convolutional neural network that uses stacked 2D convolutional layers and residual connections to extract image features. ViT[22] is a transformer-based model free from convolution operations that uses self-attention to learn local and global relations between non-overlapping patches in an image. Both methods serve as image feature extractors, producing a 1D representation of an input image suitable for classification, with ViT showing to perform better than ResNet-like models on natural image classification tasks as well as being more robust to image perturbations when trained on sufficient data.[23] Given the limited training data available, transfer learning was used with the image encoding models fine-tuned on the target CBTN dataset starting from pre-trained weights. Three distinct pre-training strategies were investigated: supervised pre-training on *out-of-domain* data (ImageNet1K[24]), self-supervised pre-training on *close-to-domain* data (BraTS[17–19]), and self-supervised pre-training on *in-domain* (CBTN) data. For the self-supervised pre-training, the SimCLR[25] framework was employed (see *Pre-training* section in the Supplementary materials for details). We also investigated the integration of image and age through a joint fusion approach. In this case, ResNet50 and ViT models were used to encode the image data, while a tabular network encoded the age information. Figure 1 shows a schematic representation of the network architecture when trained on image and age information. For the details on the implementation, model pre-training and fine-tuning, and data augmentation, see the Supplementary materials.

Evaluation Metrics and Statistical Methods

A 10-times repeated 5-fold stratified cross-validation scheme was employed in all experiments to account for the small size of the dataset. For each of the repetitions, subject-wise splitting was performed to obtain training, validation, and testing sets. Models' performance was evaluated volume-wise in terms of Matthew's correlation coefficient (MCC) [–1, 1][26] since it is a more stable metric in case of class imbalance. Accuracy and area under the receiver operating characteristic curve were also computed to allow comparison with previous studies. Class-wise F1 score, precision, and recall were additionally computed. Volume-wise predictions were obtained by soft voting aggregation of the models' predicted probability for each of the slices in a volume. The Wilcoxon signed-rank test (two-sided) was used to investigate if there were differences in classification performance between models trained on image data alone or fused with age information, or when using different pre-training strategies. The Wilcoxon rank-sums (two-sided) test was instead used to compare models when trained on different MR sequences. A *p*-value < .05 was considered significant with Bonferroni correction applied when multiple comparisons were performed.

Model Explainability and Learned Feature Space Visualization

In this work, Grad-CAM[27,28] were computed for the last convolutional layer of the ResNet50 models, and the last attention block in the ViT models, for the ground truth class. This method was chosen among the many other salience mapping methods given its overall better faithfulness, localization stability, and sensitivity.[29] Grad-CAMs were employed to ensure that the models focused on relevant regions of the input image for classification, rather than to elucidate the specific reasons or features used by models for prediction. Additionally, principal component analysis (PCA) was performed on the image feature vectors obtained from the trained models to visualize the effect of pre-training and image-age fusion.

Figure 1. Schematic representation of image and age encoders whose 1D representations are concatenated for the final classification. ViT: visual transformer.

Results

Classification Performance

The highest classification performance was achieved by the ViT model pre-trained on ImageNet and fine-tuned on ADC data with age fusion (MCC: 0.77 ± 0.14, Accuracy: 0.87 ± 0.08). This was significantly higher than the best-performing models trained on either T2w (MCC: 0.58 ± 0.11) or T1w-Gd (MCC: 0.45 ± 0.16) data. Class-wise performance (see Supplementary materials, Table S1), showed that the classification of EP is the most challenging across settings, with an average F1 score of 0.37 ± 0.28 over all the experiments, while ASTR and MED obtained 0.74 ± 0.18 and 0.76 ± 0.15, respectively. Looking at the overall effect of fusing image and age information, the models' performance did not significantly change compared to models trained on image data only, except for the ResNet50 model pre-trained on ImageNet and fine-tuned on ADC data where the addition of age information significantly decreased classification performance. The effect of the 3 pre-training strategies was not consistent across MR sequences, model architectures, or input configuration. Moreover, there was no clear benefit between pre-training on *close-to-domain* or *in-domain* data. Of notice, the ViT models trained on ADC data had a significantly better performance when fine-tuned from ImageNet pre-trained weights compared to contrastive pre-training. Finally, looking at the different model architectures, ResNet50 and ViT models performed similarly when considering pre-training strategies and input configuration. A summary of the classification performance for all the experiments is presented in Figure 2 and Table 2.

Qualitative Analysis of the Image Feature Space

Scatter plots for features extracted using the ViT model from T2w and ADC images are shown in Figure 3. On the training set, the features of the different tumor types are grouped in distinct clusters for both T2w and ADC. The different pre-training strategies did not substantially impact the feature space, with the SimCLR pre-training on CBTN showing a marginally improved cluster separation. The fusion of the age information with the image data resulted in a larger separation between the tumor type clusters compared to when only image information was used. On the test set, clusters were less distinct, with EP features largely overlapping with ASTR and MED, reflecting the lower F1 score for this class. Scatter plots for the ResNet50 models and T1w-Gd sequence are available in the Supplementary materials, Figures S1 and S2.

Grad-CAMs

Representative Grad-CAMs for models trained on the available MR sequences, with and without age fusion, and for the 3 pre-training strategies are presented in Figure 4. Results are shown for a transversal slice of a test subject for which all MR sequences were available. For the ResNet50 models, the Grad-CAMs focused primarily on the brain region with those of models trained on T2w data showing a better localization of the tumor compared to T1w-Gd or ADC. The Grad-CAMs of the ViT models highlighted the whole brain region with no discrimination of the tumor region, which is a consequence of the short and long relations between the image regions that these models learn. There is no overall difference in the Grad-CAMs between pre-training strategies, and when using age and image
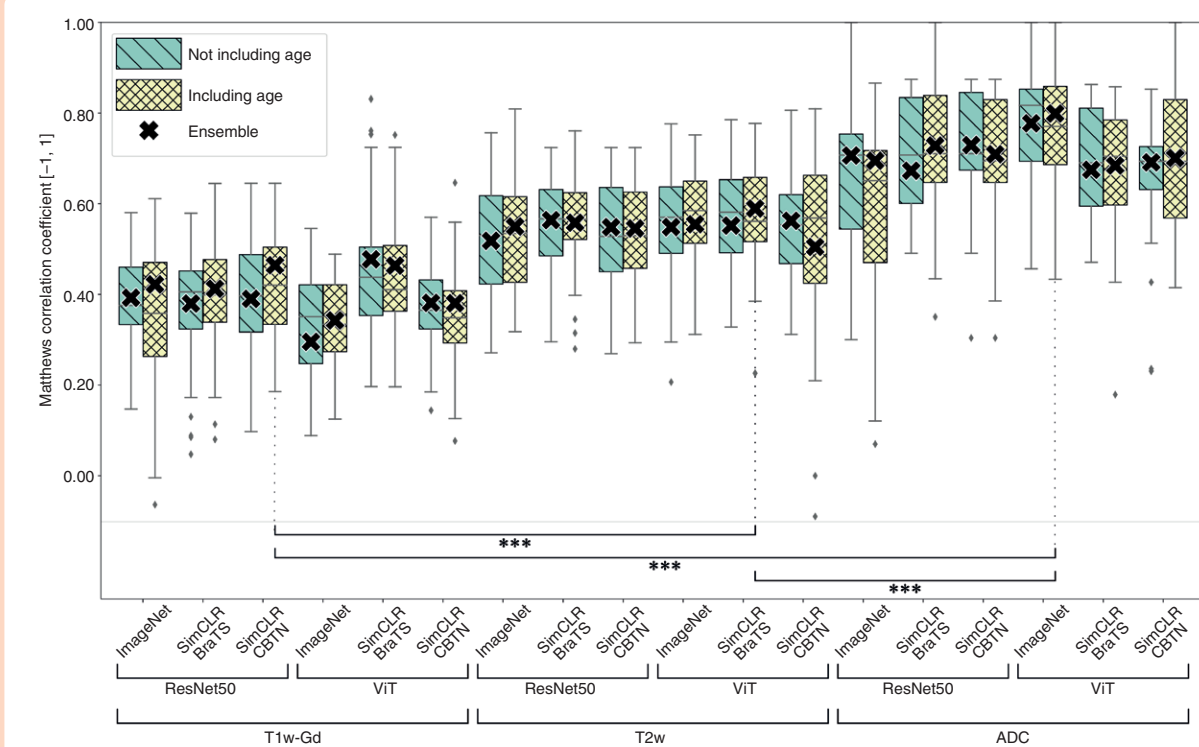
Figure 2. Subject-wise classification performance on the test set for all the available magnetic resonance (MR) sequences (with and without age fusion) and investigated model architectures and pre-training strategies. Each box plot summarizes the Matthew's correlation coefficient for the 50 models trained through a 10-times repeated 5-fold cross-validation scheme. Outliers are shown as diamonds (♦). Statistical significance is shown for the best-performing models on each MR sequence (***2-sided P-value < .0001 using Wilcoxon rank-sums test adjusted with post hoc Bonferroni correction). See Table 2 for the performance details of the best models for each MR sequence.

Table 2. Subject-wise classification performance for the best-performing models on all the available MR sequences (with and without age fusion). The overall best-performing model is highlighted in bold text. Models fine-tuned on ADC data perform significantly better than models fine-tuned on either T2w data or T1w-Gd. The addition of the age information did not significantly improve the models' performance. SimCLR: self-supervised contrastive pre-training strategy, MCC: Matthew's correlation coefficient, AUC: area under the receiver operating characteristic curve (macro-average)

MR sequence	Model specification	Use age	MCC [mean ± std]	Accuracy [mean ± std]	AUC [mean ± std]
T1w-Gd	ViT SimCLR pre-training on TCGA	✗	0.454 ± 0.156	0.659 ± 0.099	0.770 ± 0.103
		✓	0.438 ± 0.125	0.579 ± 0.082	0.693 ± 0.088
T2w	ViT SimCLR pre-training on TCGA	✗	0.569 ± 0.099	0.723 ± 0.062	0.838 ± 0.055
		✓	0.580 ± 0.114	0.731 ± 0.075	0.846 ± 0.056
ADC	ViT Pre-training on ImageNet	✗	0.767 ± 0.132	0.872 ± 0.074	0.931 ± 0.077
		✓	**0.773 ± 0.144**	**0.874 ± 0.084**	**0.940 ± 0.073**

information. This was true for both the models, except for the ViT, where examples of activation being around the brain region can be found. Additional Grad-CAMs are available in the Supplementary materials, Figure S3.

Discussion

In this study, deep learning methods were implemented for the classification of PTBs based on pre-operative MR images from the CBTN dataset. The effect of network

architectures, pre-training, MR sequence, and fusion of patient age were investigated.

Network Architecture

The ResNet50 model was chosen given previous reports in the literature for similar tasks on both pediatric[11,12] and adult brain tumor datasets.[30,31] The ViT model was selected as an alternative to convolution-based deep learning models given its success on natural image tasks[32] and its increasing adoption in medical imaging-related tasks.[33]
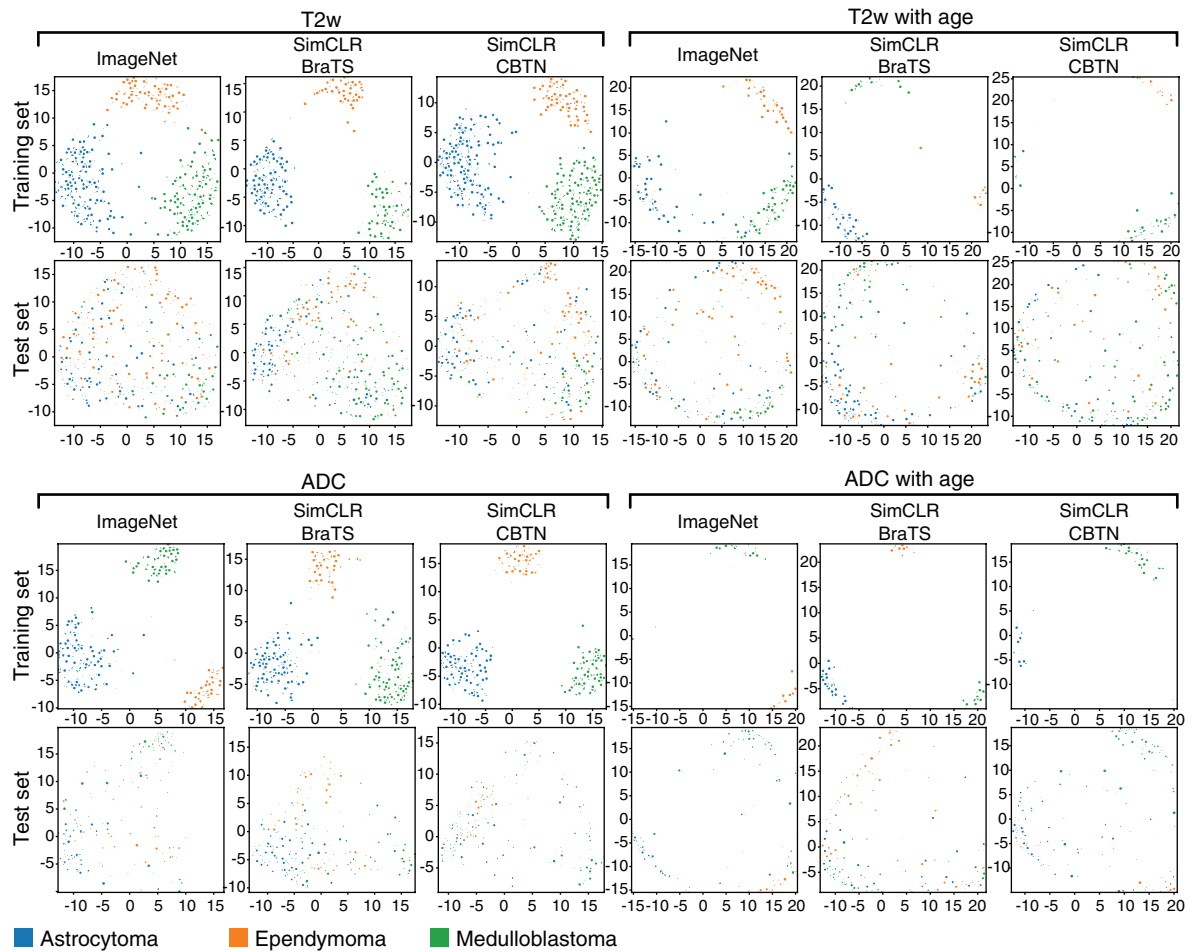
Figure 3.   Principal component analysis (PCA) of image features extracted by ViT models fine-tuned on T2w or ADC data, with and without age information, using ImageNet or SimCLR pre-trained weights. The first and second principal components are presented, for both training and testing sets. Classes are color-coded. The addition of the age information stretches the feature space and helps, in the training set, in clustering the tumor types separately. On the test set, ependymoma samples are scattered and overlapping with the other 2 classes, confirming the low F1 score for this class. ADC: apparent diffusion coefficient, SimCLR: self-supervised contrastive pre-training.

Both model architectures are available in most of the deep learning frameworks with and without supervised pre-trained weights on ImageNet, which is beneficial when training data are scarce. However, supervised pre-training does not always benefit the downstream task, with the pre-training dataset and objective having an impact on the final performance of the fine-tuned model.[34] For this reason, self-supervised contrastive pre-training[25] was employed to bridge the gap between the pre-training and down-stream dataset and objective. Overall, classification results did not benefit from the contrastive pre-training, with the best-performing model being fine-tuned from ImageNet weights. One reason the anticipated benefits of contrastive learning were not observed could be that models were trained to learn shared information from augmented views of the same transversal slice, discarding the fact that this information should be shared by all the slices in a subject. Nonetheless, PCA of the learned feature space showed a better distinction of the different tumor types when using contrastive pre-training, especially for ResNet50 models.

MR Sequences

Among the MR sequences, ADC achieved the highest overall classification performance and class-wise F1 scores for both ASTR and MED tumor types, whereas models trained on T2w data achieved the highest F1 score for the EP class (see Table S1 in the Supplementary materials). This can be attributed to the small number of EPs having ADC data ($n = 9$) compared to those having T2w data ($n = 30$). The results on ADC data were in agreement with those using deep learning-based methods,[12] as well as intensity analysis,[35] and consistent with the information neuro-radiologists use when assessing tumor cellularity and possible tumor grade, during the primary diagnosis work-up.

Age Information

A joint fusion approach was used to combine the image and age information, allowing the image and age encoder to be jointly trained. Results show that the addition of age
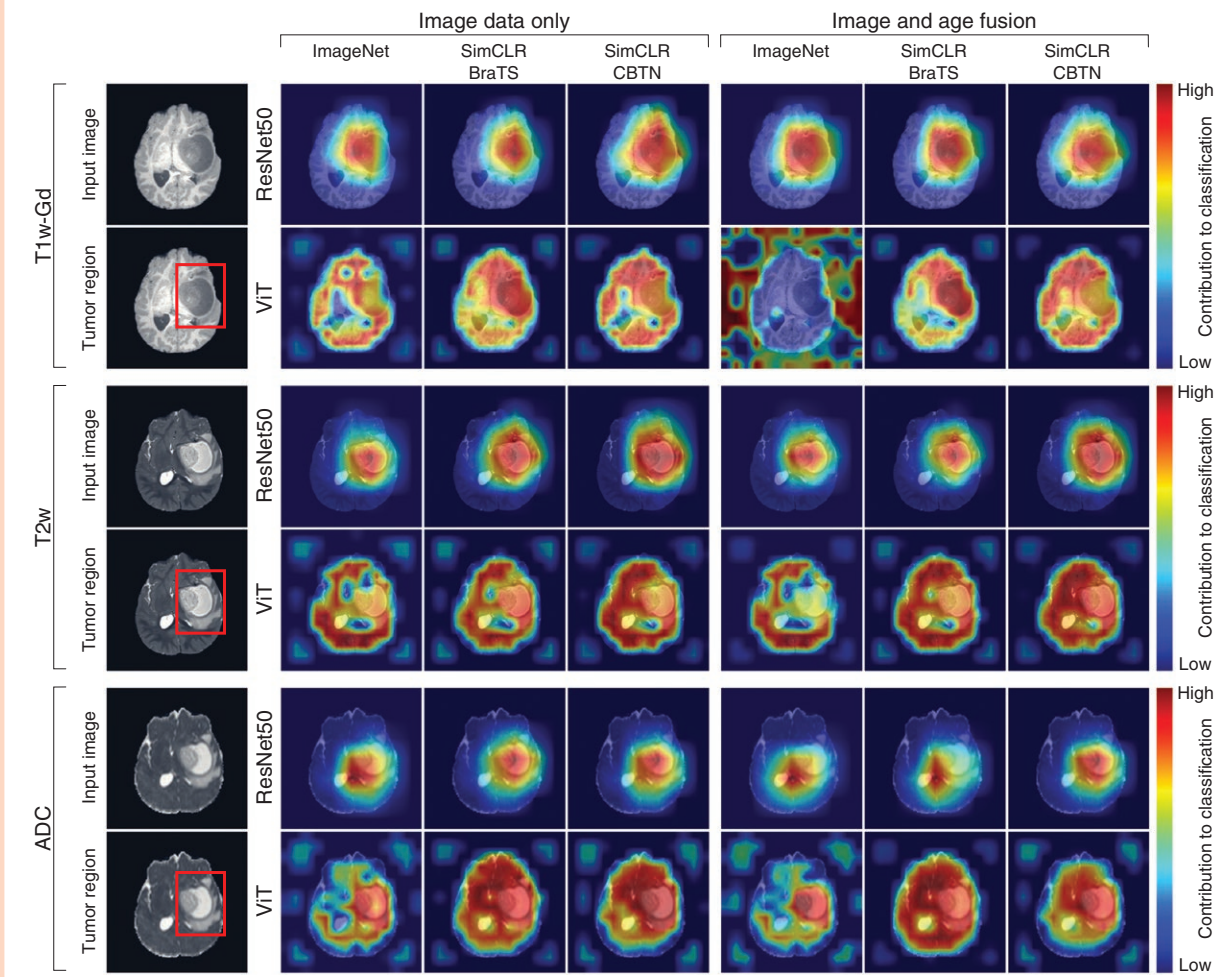
Figure 4. Grad-CAMs for the models trained on different MR sequences, with and without age fusion, and for the 3 pre-training strategies investigated. Grad-CAMs are computed with respect to the ground truth class and for the same subject (the transversal slice was taken to be as close as possible in all MR modalities). The bounding box in the *tumor region* panel delineates the tumor. In the Grad-CAMs images, the high contributing regions identify the parts of the input image that mostly influenced the model classification. ADC: apparent diffusion coefficient, SimCLR: self-supervised contrastive pre-training. ViT: visual transformer.

information did not improve classification performance across the different MR sequences, model architectures, and pre-training strategies. Preliminary investigations also explored the number of encoding layers in the age encoder, with no variation in outcome. This can be attributed to the overlapping age distribution of the different classes as well as to the choice of data fusion approach. By contrast, in a similar experimental setting the combination of image and age information improved model classification performance.[12] Thus, this leaves open the question of whether the benefits of combining age and image information for PBT classification are restricted to specific subject populations or if a more general method for image and age fusion needs to be explored which can be broadly and successfully applied.

Model Explainability

To qualitatively assess the regions used for predictions, class-activation mapping was implemented to highlight the regions in the image used by the model for prediction. Given the depth and complexity of both ResNet50 and ViT models, activation maps do not have sufficient spatial resolution to target the tumor region only and provide a visualization of tumor regions and/or image features that are relevant for the classification. Nevertheless, the models' activation maps showed that the information used for classification fell within the brain region. Interestingly, the effect of the pre-training strategy and the input configuration seen in the feature space visualizations does not reflect on the class-activation maps, suggesting that the models use the same brain regions for classification but rely on a somewhat different set of features.

Comparison with Related Work

The findings align with the few previously reported studies on deep learning-based PBT type classification, with Quon *et al.*,[11] reporting a 0.92 accuracy (F1 score of

0.80 on T2w data classifying 4 tumor types and controls) and Artzi *et al.,*[12] of 0.87 (F1 score 0.82 on diffusion Trace data classifying 3 tumor types and controls). This study's class-wise results for ASTR and MB are in agreement to those previously reported, with EP having the lowest score in all studies. It should be noted that the comparison of the performance metrics among the studies can only be considered in very general terms due to the differences in tumor types included in the analysis and the evaluation protocol. Additionally, previous studies did not provide statistical analysis to assess the impact of specific MR sequences and/or age fusion on models' performance.

Limitations

This study has some limitations, particularly concerning 2 major aspects: the amount and quality of data and the use of 2D models. The dataset, while being one of the largest accessible, has a relatively small size and the distribution of tumor types is unbalanced. Not all tumor types available in the dataset could be included in the analysis, limiting the applicability of the trained models in a real-world clinical scenario. The image quality varied greatly between each scan and subject. This large variability in data quality is advantageous for the development of a robust classification method but also adds complexity in determining which factors to adjust to optimize model performance. Moreover, the information regarding the site where the scans were acquired was missing. If available, it could have been used to stratify the subjects in training and testing sets based on MR-site. External datasets such as the BraTS-2024 pediatric dataset[36] could be considered for model development and evaluation. However, due to the absence of information on tumor types and the potential overlap of subjects with our dataset, we opted not to use this resource. Additionally, we have only reported results when using each MR sequence separately since preliminary results on multi-sequence experiments using early fusion did not show any performance benefits. However, additional experiments on multi-sequence fusion methods should be considered. Finally, although the choice of 2D deep-learning models enables experimentation with ImageNet pre-trained models, it led to the loss of information regarding the spatial relationship between the slice and the tridimensionality of the tumor structure. This information, however, was found to be relevant for pediatric tumor type and grade classification when using radiomic features.[37–39] If more data were available, using the MR volumes as model input could address this limitation. Alternatively, while still using 2D models, aggregation methods that incorporate the slice position for computing volume-wise predictions could be developed or multiple slices could be arranged in a multi-channel format and used as model input. Considering the limitations, the results of this study are preliminary and highlight the need for further research to develop methods that can match the diagnostic accuracy of radiologists that, for the investigated brain tumors, is 91.7% (range [85.1, 96.7]%) when using T1-w with and without contrast, T2-w, FLAIR, and ADC.[7]

Conclusions

In this proof-of-concept study, the classification of PBT based on MR images was achieved using deep learning methods. The vision transformer model pre-trained on ImageNet and fine-tuned on ADC data obtained the highest classification performance, with models trained on T2w data also achieving reasonable performance. Image and age fusion did improve classification performance, but not significantly. In future studies, the combination of multiple MR sequences along with more detailed clinical information and further refinements of the network architectures, pre-training, and data fusion are warranted to aid radiologists in the clinical assessment of these tumors.

Supplementary material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords:

age | data fusion | deep learning | MRI | pediatric brain tumor

Conflict of interest statement

The authors declare no conflict of interest in this work.

Authorship statement

I.E.T.: methodology, algorithm, validation, formal analysis, investigation, data curation, visualization, original draft, and

reviewing. T.B.: methodology, investigation, data curation, original draft, and manuscript review. I.B.: conceptualization, data curation, and manuscript reviewing. P.L.: manuscript review. P.N.: conceptualization, data curation, and manuscript reviewing. A.E.: methodology, resources, and manuscript reviewing. N.H.H.: conceptualization, methodology, resources, supervision, and manuscript drafting and reviewing.

Data availability

The open-access dataset used in this study was obtained from The Children's Brain Tumor Network (CBTN) at https://cbtn.org. The code linked to this manuscript is available at https://github.com/IulianEmilTampu/PediatricBrainTumorClassification-MRI.

Affiliations

Department of Biomedical Engineering, Linköping University, Linköping, Sweden (I.E.T., T.B., A.E., N.H.-H.); Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden (I.E.T., T.B., I.B., P.L., P.N., A.E., N.H.-H.); Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden (T.B.); Department of Radiology and Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden (I.B.); Department of Radiation Physics and Department of Medical and Health Sciences, Linköping University, Linköping, Sweden (P.L.); Crown Princess Victoria Children's Hospital and Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden (P.N.); Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden (A.E.)

References

1. Ferlay J, Ervik M, Lam F, et al. Global cancer observatory: cancer today. https://gco.iarc.fr/today/ home 2022. Accessed 2023.
2. Sharma R. A systematic examination of burden of childhood cancers in 183 countries: estimates from GLOBOCAN 2018. *Eur J Cancer Care.* 2021;30(5):e13438.
3. Ostrom Quinn T, Patil N, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan Jill S. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro-oncology.* 2020;22(7):iv1–iv96.
4. Corti C, Urgesi C, Massimino M, Gandola L, Bardoni A, Poggi G. Effects of supratentorial and infratentorial tumor location on cognitive functioning of children with brain tumor. *Child's Nerv Syst.* 2020;36(3):513–524.
5. Pollack Ian F. Brain tumors in children. *N Engl J Med.* 1994;331(22):1500–1507.
6. Louis David N, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology.* 2021;23(8):1231–1251.
7. Dixon L, Jandu Gurpreet K, Sidpra J, Mankad K. Diagnostic accuracy of qualitative MRI in 550 paediatric brain tumours: evaluating current practice in the computational era. *Quant Imaging Med Surg.* 2022;12(1):131.
8. Ali S, Li J, Pei Y, Khurram R, Rehman Khalil U, Mahmood T. A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal MR image. *Arch Comput Methods Eng.* 2022;29(7):4871–4896.
9. Amin J, Sharif M, Haldorai A, Yasmin M, Nayak Ramesh S. Brain tumor detection and classification using machine learning: a comprehensive survey. *Complex Intell Syst.* 2021;8(4):1–23.
10. Tandel Gopal S, Biswas M, Kakde Omprakash G, et al. A review on a deep learning perspective in brain cancer classification. *Cancers.* 2019;11(1):111.
11. Quon JL, Bala W, Chen LC, et al. Deep learning for pediatric posterior fossa tumor detection and classification: a multiinstitutional study. *Am J Neuroradiol.* 2020;41(9):1718–1725.
12. Artzi M, Redmard E, Tzemach O, et al. Classification of pediatric posterior fossa tumors using convolutional neural network and tabular data. *IEEE Access.* 2021;9:91966–91973.
13. Shaari H, Kevric J, Jukić S, Bešić L, Jokić D, Ahmed N, Rajs V. Deep learning-based studies on pediatric brain tumors imaging: narrative´ review of techniques and challenges. *Brain Sci.* 2021;11(6):716.
14. Huang J, Shlobin Nathan A, Lam Sandi K, DeCuypere M. Artificial intelligence applications in pediatric brain tumor imaging: a systematic review. *World Neurosurg.* 2022;157:99–105.
15. The Children's brain tumor network. https://cbtn.org/. Accessed: 2021.
16. Lilly Jena V, Rokita Jo L, Mason Jennifer L, et al. The children's brain tumor network (CBTN)-Accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia.* 2023;35:100846.
17. Menze Bjoern H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging.* 2014;34(10):1993–2024.
18. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data.* 2017;4(1):1–13.
19. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS Challenge. *CoRR.* 2018;abs/1811.02629. https://doi.org/10.48550/arXiv.1811.02629.
20. Tournier J-D, Smith R, Raffelt D, et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage.* 2019;202:116137.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.
22. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations*, 2021.
23. Bhojanapalli S, Chakrabarti A, Glasner D, Li D, Unterthiner T, Veit A. Understanding robustness of transformers for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, 10231–10241.
24. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, 248–255.
25. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning PMLR, 2020, 1597–1607.
26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21(1):1–13.

27. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 2921–2929.

28. Selvaraju Ramprasaath R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. GradCAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017, 618–626.

29. Li X-H, Shi Y, Li H, Bai W, Cao Caleb C, Chen L. An experimental study of quantitative evaluations on saliency methods. In: Feida Z, Chin OB, Chunyan M, eds. KDD'21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021, ACM, 2021, 3200–3208.

30. Chelghoum R, Ikhlef A, HameurLaine A, Jacquir S. Transfer learning using convolutional neural network architectures for brain tumor classification from MRI images. In Proceeding of Artificial Intelligence Applications and Innovations, 2020, 189–200.

31. Rehman A, Naz S, Razzak Muhammad I, Akram F, Imran M. A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits Syst Signal Process.* 2020;39(2):757–775.

32. Khan S, Naseer M, Hayat M, Zamir Syed W, Khan Fahad S, Shah M. Transformers in vision: a survey. *ACM Comput Surv.* 2022;54(10):1.

33. Shamshad F, Khan S, Zamir Syed W, Khan Muhammad H, Hayat M, Khan Fahad S, et al. Transformers in medical imaging: a survey *Med Image Anal.* 2023;88:102802.

34. Zoph B, Ghiasi G, Lin TY, et al. Rethinking pre-training and self-training. *Adv Neural Inf Process Syst* 2020;33:3833–3845.

35. Tanyel T, Nadarajan C, Duc Nguyen M, Keserci B. Deciphering machine learning decisions to distinguish between posterior fossa tumor types using MRI features: what do the data tell us? *Cancers.* 2023;15(16):4015.

36. Kazerooni Anahita F, Khalili N, Gandhi D, et al. The brain tumor segmentation in pediatrics (BraTS-PEDs) challenge: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). *CoRR.* 2024;abs/2404.15009. https://doi.org/10.48550/ARXIV.2404.15009.

37. Iv M, Zhou M, Shpanskaya K, et al. MR imaging–based radiomic signatures of distinct molecular subgroups of medulloblastoma. *Am J Neuroradiol.* 2019;40(1):154–161.

38. Grist James T, Withey S, MacPherson L, et al. Distinguishing between paediatric brain tumour types using multi-parametric magnetic resonance imaging and machine learning: a multi-site study. *NeuroImage: Clin.* 2020;25:102172.

39. Dong J, Li L, Liang S, et al. Differentiation between ependymoma and medulloblastoma in children with radiomics approach. *Acad Radiol.* 2021;28(3):318–327.