Article

Human-Level Differentiation of Medulloblastoma from Pilocytic Astrocytoma: A Real-World Multicenter Pilot Study

Benedikt Wiestler [1,2,3,*] , Brigitte Bison [3,4,5,6] , Lars Behrens [3,4,5,6], Stefanie Tüchert [3,7], Marie Metz [1,3], Michael Griessmair [1,3], Marcus Jakob [3,4,8], Paul-Gerhardt Schlegel [3,4,9], Vera Binder [3,4,10], Irene von Luettichau [3,4,11], Markus Metzler [3,4,12] , Pascal Johann [3,4,13], Peter Hau [3,14] and Michael Frühwald [3,4,13]

1   Department of Neuroradiology, School of Medicine and Health, Technical University of Munich, 81675 Munich, Germany; michael.griessmair@tum.de (M.G.)
2   TranslaTUM, Center for Translational Cancer Research, Technical University of Munich, 81675 Munich, Germany
3   Study Groups on CNS Tumors Within the Bavarian Cancer Research Center (BZKF)
4   KIONET, Kinderonkologisches Netzwerk Bayern
5   Diagnostic and Interventional Neuroradiology, Faculty of Medicine, University Hospital Augsburg, 86156 Augsburg, Germany; brigitte.bison@uk-augsburg.de (B.B.); lars.behrens@uk-augsburg.de (L.B.)
6   Neuroradiological Reference Center for the Pediatric Brain Tumor (HIT) Studies of the German Society of Pediatric Oncology and Hematology, Faculty of Medicine, University Hospital Augsburg, 86156 Augsburg, Germany
7   Department of Diagnostic and Interventional Radiology, University Hospital Augsburg, 86156 Augsburg, Germany
8   Department of Pediatric Hematology, Oncology and Stem Cell Transplantation, University of Regensburg, 93053 Regensburg, Germany; marcus.jakob@ukr.de
9   Department of Pediatric Hematology, Oncology and Stem Cell Transplantation, University Children's Hospital Würzburg, 97080 Würzburg, Germany; schlegel_p@ukw.de
10 Department of Pediatrics, Dr. Von Hauner Children's Hospital, University Hospital, LMU Munich, 80539 Munich, Germany; vera.binder@med.uni-muenchen.de
11 Division of Pediatric Hematology and Oncology, Department of Pediatrics, Kinderklinik München Schwabing, Children's Cancer Research Center, TUM School of Medicine and Health, Technical University of Munich, 80333 Munich, Germany; irene.teichert-vonluettichau@mri.tum.de
12 Pediatric Oncology and Hematology, Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), 91054 Erlangen, Germany; markus.metzler@uk-erlangen.de
13 Swabian Children's Cancer Center, Pediatrics and Adolescent Medicine, University Hospital Augsburg, 86156 Augsburg, Germany; pascal.johann@uk-augsburg.de (P.J.); michael.fruehwald@uk-augsburg.de (M.F.)
14 Department of Neurology and Wilhelm Sander-NeuroOncology Unit, University Hospital Regensburg, 93053 Regensburg, Germany; peter.hau@klinik.uni-regensburg.de
*   Correspondence: b.wiestler@tum.de

Simple Summary: Reliable preoperative differentiation of pediatric brain tumors can be challenging. While deep learning models have made significant progress in radiology, their use in pediatric populations is limited, typically through limited data availability. In this proof-of-concept study, we investigated the potential of a deep learning classifier trained on a multicenter data set of 195 children to learn to differentiate between pilocytic astrocytoma and medulloblastoma, the two most common infratentorial pediatric brain tumors, which in general present with overlapping imaging features. Our model is validated against the assessment of five independent readers of varying expertise. The final models performed strongly (AUC 0.986) on the unseen test set, correctly predicting the tumor diagnosis in 62 of 64 patients (97%). Compared to human readers, the classifier performed significantly better than relatively inexperienced readers and was on par with pediatric neuroradiologists with specific expertise in pediatric neuro-oncology. Our work highlights the potential of deep learning even in this challenging population and warrants future studies, including different tumor types and diverse acquisition protocols.

Abstract: Medulloblastoma and pilocytic astrocytoma are the two most common pediatric brain tumors with overlapping imaging features. In this proof-of-concept study, we investigated using

a deep learning classifier trained on a multicenter data set to differentiate these tumor types. We developed a patch-based 3D-DenseNet classifier, utilizing automated tumor segmentation. Given the heterogeneity of imaging data (and available sequences), we used all individually available preoperative imaging sequences to make the model robust to varying input. We compared the classifier to diagnostic assessments by five readers with varying experience in pediatric brain tumors. Overall, we included 195 preoperative MRIs from children with medulloblastoma ($n = 69$) or pilocytic astrocytoma ($n = 126$) across six university hospitals. In the 64-patient test set, the DenseNet classifier achieved a high AUC of 0.986, correctly predicting 62/64 (97%) diagnoses. It misclassified one case of each tumor type. Human reader accuracy ranged from 100% (expert neuroradiologist) to 80% (resident). The classifier performed significantly better than relatively inexperienced readers ($p < 0.05$) and was on par with pediatric neuro-oncology experts. Our proof-of-concept study demonstrates a deep learning model based on automated tumor segmentation that can reliably preoperatively differentiate between medulloblastoma and pilocytic astrocytoma, even in heterogeneous data.

Keywords: brain; pediatric brain tumor; MRI; artificial intelligence; deep learning

1. Introduction

Tumors of the CNS constitute the largest group of solid neoplasms in children and adolescents [1]. Medulloblastomas, comprising 15–20% of all CNS tumors, constitute the most common malignant CNS neoplasm in this age group. Low-grade gliomas, however, are by far the most common pediatric CNS tumors, accounting for up to 40% of all CNS tumors in childhood. Among these, pilocytic astrocytomas are the single most common entity in children and young adults [2].

Modern imaging techniques have significantly improved the differentiation of low- and high-grade lesions. Several guidelines providing detailed information for standard imaging approaches are in place [3,4]. Usually, on MRIs, medulloblastomas appear iso- to hypointense on T1w images and the T2w signal is variable and often heterogenous, ranging from hyperintense to hypointense. They show restricted diffusion, (which may help differentiate medulloblastoma from pilocytic astrocytoma) and, depending on the subtype, variable enhancement and edema. Intralesional cysts can be found. MR spectroscopy can depict a high choline peak at 3.2 ppm and a taurine peak at 3.4 ppm [5–7]. Nevertheless, especially in very young children, atypical localizations for the respective tumor subtype and in the relapse/recurrence situation, imaging features might be less distinct, and neuropathological diagnosis following neurosurgical interventions remains the mainstay of diagnosis.

Distinguishing medulloblastomas from pilocytic astrocytomas is already preoperatively clinically relevant for planning additional staging diagnostics, such as MRI of the neuroaxis and CSF puncture and therapeutic procedures like the extent of neurosurgical resection. While in both tumors a maximum safe resection is generally advised, for the malignant entity of medulloblastoma it is even more prognostically imperative to completely resect the tumor, as residual tumors have repeatedly been shown to be of prognostic importance [8]. CSF cytology and spinal MRI are required to accurately assess the extent of disease in medulloblastoma (as per Chang stages [9]). Both CSF and spinal MRI analyses are postoperatively at risk for false-positive findings (for example, due to hemorrhage or unspecific postoperative change), and are therefore better scheduled preoperatively than postoperatively [10].

Recently, deep learning has enabled unprecedented advances in how clinicians can use imaging data of CNS tumor patients to improve the diagnosis and prognosis of these patients. Modern algorithms enable accurate volumetric segmentation of gliomas across the clinical course of the disease [11,12], which allows for more objective response assessment [13,14], which is now consequently also codified in the respective diagnostic criteria [15–17]. Apart from objective volumetry, segmentation is another basis for subse-

quent image analysis strategies [18], which have yielded important image-based biomarkers for molecular subtyping [19,20] and the prognostication [21] of gliomas. In pediatric neuro-oncology, fewer studies have used preoperative MRI to predict tumor biology [22,23]. Further, some of these studies were comparatively small (<100 patients), hindering the development (and evaluation) of deep learning classifiers and highlighting the need for multi-centric analysis.

Here, we aimed to develop a deep learning pipeline for automated tumor segmentation and classification into medulloblastoma and pilocytic astrocytoma in a challenging, multicenter data set with high variability in imaging sequences (and their availability) as a pilot study. We further evaluated this classifier against a group of radiologists with varying expertise in pediatric brain tumor assessment.

2. Materials and Methods

2.1. Data Set

The German HIT network for children with tumors of the CNS variation is responsible for the neuroradiology reference evaluations for most patients with CNS tumors recruited to the different clinical trials and registries within the community. As such, most diagnostic images of German patients affected by CNS tumors are remotely and continuously evaluated by the German neuroradiology reference center located in Augsburg. As young adults, including individuals aged up to 21 years, are frequently treated at pediatric sites, the reference network also includes MRI images from this age group. The six University Medical Centers in Bavaria, Germany, are organized within the Bavarian Cancer Research Center (BZKF). The pediatric branch of the BZKF is the KIONET, which comprises the six pediatric hematology oncology units at the six University Centers (Augsburg, Erlangen, TU München, LMU München, Regensburg, and Würzburg). In total, about 400 to 450 of the 2010–2200 yearly diagnosed German children with malignancies are located in Bavaria. Up to 65–70 of these are tumors of the CNS variation.

Most of the imaging studies included were completed according to the reference panel recommendations of the HIT network (with some deviations as detailed below). Patients were treated uniformly according to the different clinical trials initiated by the German HIT network or outside clinical trials.

In general, MR images and, whenever available, CTs of patients with CNS tumors are registered on a common data platform. In general, medical data and images are exchanged and stored in the HIT network via the MDPE (medical data and picture exchange) server. This server is operated by the central data management (ZDM) of the GPOH. The data protection concept of this server allows the use of anonymized data for research purposes. The patients and/or their legal guardians gave written informed consent.

For the purpose of this study, all images of affected patients diagnosed within the previous 10 years within the KIONET were available for analysis in an anonymized way.

2.2. Image (Pre)Processing

Available preoperative MR sequences (T1w $-/+$ contrast, T2w, FLAIR, and ADC maps) were rigidly coregistered and transformed into SRI space [24] using NiftyReg [25]. Following skull-stripping using HD-BET [26], we normalized images into [0; 1] within the brain mask and performed automated tumor segmentation using the ensemble strategy implemented in BraTS.Toolkit [27]. BraTS.Toolkit performs tumor segmentation (into necrosis/cysts, contrast-enhancing tumor, and peritumoral edema) using several top-performing algorithms from the BraTS (Brain Tumor Segmentation) challenge [28] and fuses these candidate segmentations into a single consensus segmentation. For tumor segmentation, four input sequences are necessary (T1w $-/+$ contrast, T2w, and FLAIR). Missing sequences were imputed using a GAN-based strategy [29]. Note, however, that we used these synthetic images only for segmentation, but not for downstream classification. An attending neuroradiologist with over 10 years of experience in brain tumor imaging (BW) checked all resulting segmentations. From the center of mass of the automatically

segmented tumor core (i.e., the union of necrotic/cystic and contrast-enhancing tumor areas), we extracted $96 \times 96 \times 96$ patches for downstream classification.

2.3. Model Development

We implemented a DenseNet Deep Learning model [30] to predict the tumor entity. In brief, a DenseNet is characterized by dense connections within a layer, where each block receives direct input from all blocks preceding it. This architecture helps to exploit feature re-use to efficiently learn image features using comparatively small filter banks. The reference implementation of DenseNet121 in Keras (version 2.6; https://www.tensorflow.org/versions/r2.6/api_docs/python/tf/keras/applications/densenet/DenseNet121, accessed on 1 June 2023), which consists of 4 layers with 6, 12, 24, and 16 blocks, respectively, was used for this study. Given the three-dimensional nature of our input images, we changed the architecture to 3D convolutions and pooling operations and switched to a single (binary) output neuron with sigmoid activation.

The input to our network consisted of $64 \times 64 \times 64$ sized patches, where all available imaging sequences were concatenated along the last axis. Missing sequences were replaced using blank masks. During training, we random-cropped the $96 \times 96 \times 96$ patches from above as a data augmentation strategy, while for testing, we center-cropped patches (given that the tumor core's center of mass is the center of each patch). Besides this random cropping, we also implemented random gamma adjustment, random Gaussian noise, and random axis flipping as intensity or geometric augmentation strategies. In addition, to improve the robustness of our network to missing sequences, we randomly blanked out one input sequence. The network was trained using the Adam optimizer with a base learning rate of 1×10^{-3} and cosine annealing schedule, and a batch size of 42 for a total of 250 epochs using binary cross-entropy loss on an Nvidia Quadro RTX 8000 GPU with 48 GB of RAM.

2.4. Statistical Evaluation and Comparison

We assessed classifier performance in a hold-out test set (n = 64 patients) not used during training. In addition, we provided MR images from these test set patients to five pediatric radiologists and neuroradiologists with varying levels of expertise in pediatric gliomas, asking them to classify tumors as either medulloblastoma or pilocytic astrocytoma. To compare the proportion of samples correctly predicted between the classifier and the human raters, we calculated the Z statistic as follows [31]:

$$Z = \frac{p1 - p2}{\sqrt{2p(1-p)/n}}$$

where p_1 is the proportion of the correctly predicted n samples for the model (x_1/n), p_2 is the respective proportion for the human rater, and p is their mean ($(x_1 + x_2)/2 \times n$).

In addition, we plotted the feature representation (from the global average pooling layer) for the test set data after tSNE (T-distributed Stochastic Neighbor Embedding) dimensionality reduction (employing "cosine" distance) using the scikit-learn (version 1.2.2) implementation.

3. Results

3.1. Patient Characteristics

Our cohort comprised a total of 195 pediatric and adolescent patients with either medulloblastoma (n = 69) or pilocytic astrocytoma (n = 126). The age distribution was similar in both groups: the median age for patients with medulloblastoma was 8.5 years (interquartile range 4.8–12.9 years) and 9.1 years (interquartile range 4.9–13.5 years) for patients with pilocytic astrocytoma (p = 0.93, Mann–Whitney U test). Of all patients, seventeen had a missing ADC map, six patients had missing T2w images, and five patients had no non-enhanced T1w images. Also, with respect to imaging parameters, we observed a high variability across patients. While contrast-enhanced T1w images tended

to be acquired in an isotropic fashion with voxel sizes $< 2 \times 2 \times 2$ mm^3, the remaining sequences were mainly acquired in 2D, i.e., with a through-plane resolution usually exceeding 4 mm. We performed a stratified split of this group into a training cohort of 131 patients ($n = 46$ medulloblastomas and $n = 85$ pilocytic astrocytomas) and an independent test cohort of 64 patients ($n = 23$ medulloblastomas and $n = 41$ pilocytic astrocytomas).

3.2. Deep Learning Results

The entire processing runtime (including registration, skull-stripping, segmentation, and classification) amounted to less than 5 min per sample on a standard workstation with a GPU (12 GB VRAM). In the independent test cohort, the developed classifier showed a very high area under the receiver operating characteristic curve of 0.986 (Figure 1).
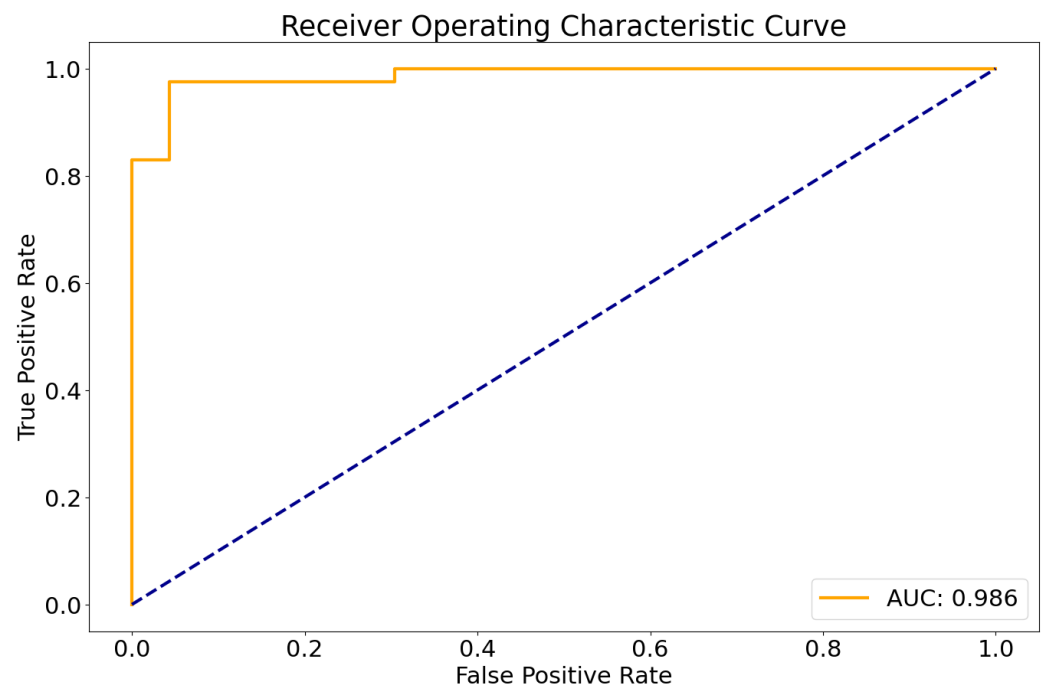

Figure 1. Receiver operating characteristic curve of the DenseNet model in the test set. AUC, Area under the curve.

Using a pre-defined decision threshold of 0.5 to binarize the predictions, 62 out of 64 samples (97%) were correctly classified as either medulloblastoma or pilocytic astrocytoma in comparison to neuropathological diagnosis according to the WHO classification, which was used as the gold standard. The model misclassified one pilocytic astrocytoma and one medulloblastoma. For these two cases, representative central slices are shown in Figure 2. For correctly predicting medulloblastoma, this translated into a sensitivity of 0.96 and a specificity of 0.97. For pilocytic astrocytoma, sensitivity was 0.97 and specificity 0.96, consequently. In total, the resulting classifier had an averaged F1 score of 0.96 and a Matthews correlation coefficient of 0.93.

To investigate the learned representations of the two different tumor types, we additionally plotted the features (taken from the global average pooling layer immediately before the classification head) after dimensionality reduction with tSNE (T-distributed Stochastic Neighbor Embedding), as shown in Figure 3. For the vast majority of samples, the two tumor types show a clearly distinct clustering; only for one case each (which are the two misclassified cases as shown in Figure 2) were the representations not distinctive.
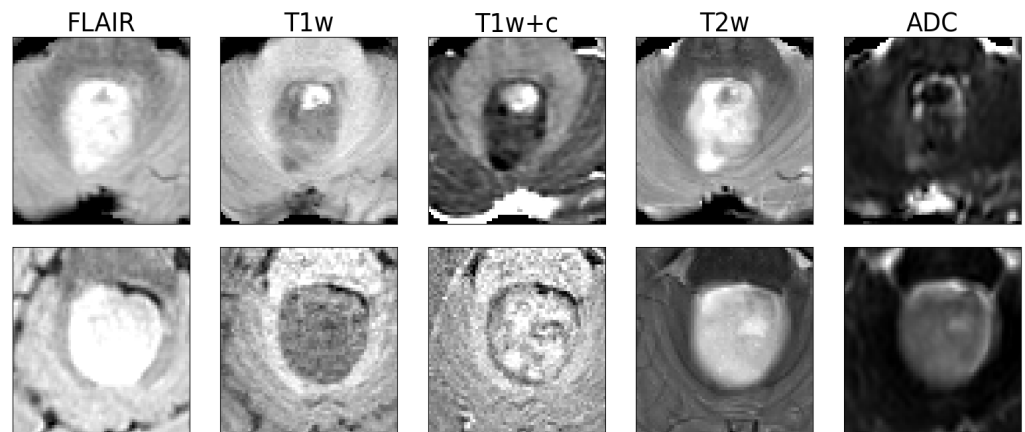
Figure 2. The two examples misclassified by the model. The top row is a medulloblastoma, and the bottom row is a pilocytic astrocytoma. Shown are axial slices of the 64^3 input patches presented to the network (the human rater received whole-brain 3D volumes).
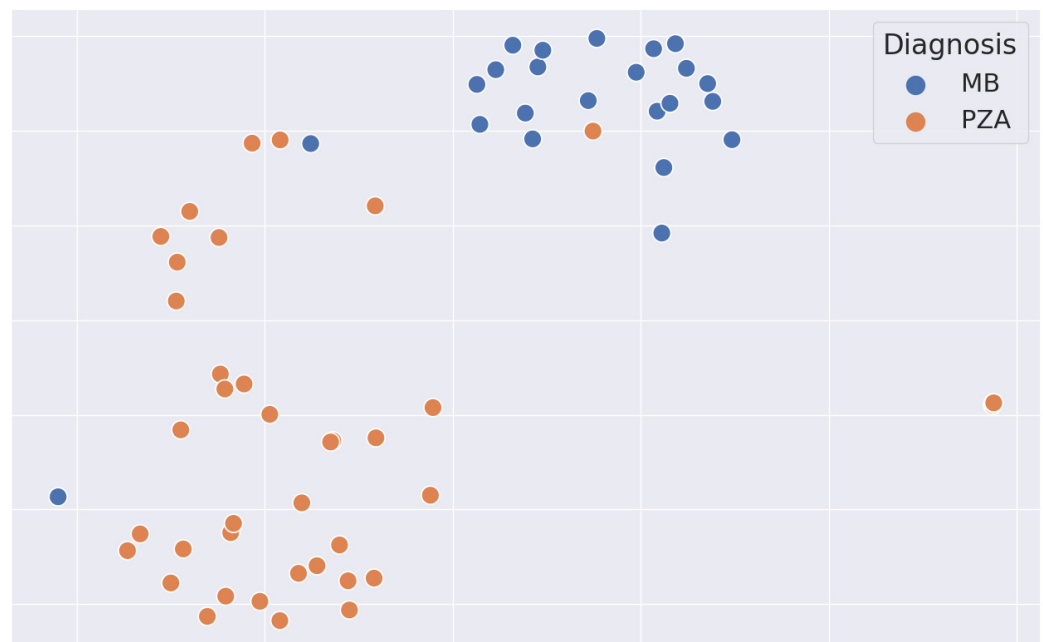
Figure 3. tSNE plot of the learned feature representations in the test set. Note the clear distinction between medulloblastoma (blue) and pilocytic astrocytoma (orange), highlighting the robust separations in the learned representations. tSNE, T-distributed Stochastic Neighbor Embedding; MB, medulloblastoma; PZA, pilocytic astrocytoma.

3.3. Sequence Importance

To better understand the importance of the different sequences for tumor classification (and to evaluate the robustness of our model to missing data), we performed an additional experiment where we intentionally blanked out each input sequence (FLAIR, T1w, T1w+c, T2w, and ADC) in turn and re-calculated the AUC for the test set. Coming from an AUC of 0.986 when using all available data, the test set performance remained stably high when omitting ADC (0.983), FLAIR (0.963), T1w (0.984), or even T1w+c (0.987). Only upon the exclusion of T2w images did we observe a noticeable drop in performance (AUC 0.92).

3.4. Expert Comparison

To compare our model to human raters, we asked five radiologists with varying levels of expertise in pediatric brain tumor imaging to classify the 64 test set cases. The results are summarized in Table 1. For the two experts from the Neuroradiological Reference

Center for the pediatric brain tumor (HIT) studies of the German Society of Pediatric Oncology and Hematology, the classification performance was very similar (in one case identical) to our model. The two expert mistakes were different cases than the classifier. For the remaining three raters, the classification accuracy of our deep learning model was higher, in particular for the two neuroradiological residents with expertise in adult brain tumor imaging but without relevant prior experience in pediatric brain tumors. Here, the proportion of correctly classified cases was significantly higher for the deep learning model compared to both readers ($p < 0.05$ each).

Table 1. Results for the DL-based model and the human raters.

	Accuracy	F1	MCC	*
DL Model	0.97	0.96	0.93	
Expert Rater 1	1	1	1	
Expert Rater 2	0.97	0.96	0.93	
Pediatric Radiologist	0.92	0.91	0.82	
Resident 1	0.87	0.86	0.72	*
Resident 2	0.84	0.81	0.66	*

* denotes a significantly ($p < 0.05$) higher accuracy for the DL-based model. MCC, Matthews Correlation Coefficient.

4. Discussion

Reliable preoperative differentiation of medulloblastoma and pilocytic astrocytoma can be challenging. Deep learning models may meaningfully support clinicians in this task. However, the development of these tools is typically limited by small sample sizes, particularly in single-center studies [32]. Here, we performed a proof-of-concept study demonstrating how a deep learning pipeline encompassing segmentation and classification can leverage a highly heterogeneous imaging data set to train a reliable classifier with a strong performance (AUC 0.986). We further demonstrate that our model performs rather positively when compared to highly specialized and experienced pediatric neuroradiologists from the Neuroradiological Reference Center for the pediatric brain tumor (HIT) studies of the German Society of Pediatric Oncology and Hematology and outperforms neuroradiology residents. Our results pave the way for larger, multicenter studies, including further pediatric tumor entities, to train generalizable image classifiers for clinical applications.

An increasing number of deep learning-based approaches for tumor detection and segmentation or classification in adult neuro-oncology is on record. At the same time, remarkably fewer studies exist in pediatric neuro-oncology, as highlighted in a recent review by Madhogarhia and colleagues [32]. These authors identify limited sample sizes (particularly in single institutional studies) as one major challenge for developing such models in pediatric patients. Many of the studies in this review contained data sets of fewer than 100 patients, which impairs the training of robust deep learning models. In contrast, we curated a large, heterogeneous data set from the six University Medical Centers in Bavaria, Germany, organized within the Bavarian Cancer Research Center (BZKF). This "exposure" of the DenseNet model during training to a large variety of patients and MR scanners—to account for technical variations—translated into a highly efficient classifier. Notably, the cases the model misclassified are different from the cases in which the human raters erred. This "complementarity" of errors highlights the potential for a high-level joint assessment by a deep learning model and radiologists to correct each other's mistakes, which would lead to 100% accuracy, at least in our cohort. In addition, deep learning models promise to incorporate additional information, for example, from clinical data, other imaging modalities such as PET or CSF cytology, and genomic analyses, supporting clinicians in diagnostic (and therapeutic) decisions even preoperatively: Tumor resection following neoadjuvant chemotherapy has been a long-standing hope in pediatric neuro-

oncology, i.e., the prospect of complete resection of a smaller lesion with the potential of fewer permanent neurological deficits stands out. Knowing in advance the histology (and potentially also the molecular background) of a lesion employing AI imaging in conjunction with liquid biopsy and potentially other imaging technologies such as PET may lead such a project to success.

Some studies into the imaging-based differentiation of pediatric brain tumors investigated the importance of individual MR sequences. Among the commonly acquired sequences, T2w images [33] and ADC maps from diffusion-weighted imaging [34,35] have been identified as particularly helpful in this task. Consequently, we included these sequences wherever available. Upon experimental evaluation of sequence importance (through round-robin omission), we found that T2w stood out for its relevance for correct classification in our setting.

As with any real-world data set, we observed missing (or corrupted, e.g., by motion) sequences. Instead of excluding such cases (as is usually the case), we specifically opted to make our classifier robust to missing data. Handling missing data is an active area of research, and several strategies have been devised to deal with this. With recent developments in generative AI, models have been developed to synthesize missing sequences from existing data. The use of these models has, for example, been demonstrated for tumor segmentation [29], and we subsequently employed this strategy for automated segmentation as well, given the availability of a pre-trained model. For the classification, however, we specifically opted for a different strategy, as to our knowledge, no task-specific generative network is available. Here, we adopted a random drop-out strategy, i.e., we randomly deleted input sequences during training as part of our augmentation pipeline. Recently, such a strategy has also been demonstrated to achieve state-of-the-art results in brain tumor segmentation [36] and provide an attractive additional strong augmentation paradigm for training. Further, when including multiple input sequences, care must be taken not to overfit the classifier by adding excessive imaging "noise", with potentially harmful consequences for generalizability [37]. The random drop-out strategy we employed provides additional regularization to avoid overfitting.

In a prior study, Zhou et al. developed a machine learning classifier to differentiate pediatric posterior fossa tumors on MRIs [38]. They report high AUC values for the classification of medulloblastoma and pilocytic astrocytoma. Similar to our findings, they report that non-expert radiologists have lower accuracy than their classifier model. A key difference to our model is their reliance on hand-crafted radiomics features extracted from manually drawn tumor masks: prior studies have shown that differences in segmentation critically affect the feature stability and hence, the reproducibility of results [39]. We thus chose a patch-based approach (centered around the center of mass of the tumor segmentation) paired with an automatic segmentation module, enabling a fully-automated image analysis without the need for manual interference. Coupled with our random cropping augmentation, our model should, therefore, be robust against minor differences in seed voxel location, i.e., as long as the seed voxel is placed well inside the tumor, downstream classification stability should not be affected. This robustness against input variations is a strength of our approach.

Quon et al. report on developing a 2D slice-wise deep learning classifier for differentiating medulloblastoma, pilocytic astrocytoma, diffuse midline glioma, and ependymoma in a large cohort of 617 children [40]. Their final ensemble classifier had an F1 score of 0.8, albeit in a more challenging multi-class classification. Again, in line with our results, they found that when comparing model performance with radiologists, in particular, less experienced readers have lower diagnostic accuracy. As opposed to our work, only axial T2w slices were used due to the 2D slice-wise design. Conceptually, our 3D concatenation approach allows the deep learning model to capture relevant synergies between the different modalities and offers another explanation for the higher performance we observe in our model. This improved differentiation of medulloblastoma and pilocytic astrocytoma can also be seen when comparing the representations in feature space. tSNE

plotting shows a clear separation of the two tumor types, where medulloblastomas cluster particularly tightly.

We acknowledge some limitations of our work. First, our analysis focused on medulloblastoma and pilocytic astrocytoma, excluding other entities such as ependymoma or diffuse midline glioma. While these are the two most common infratentorial tumors in children, and their preoperative differentiation has clinical relevance, this creates a specific context for this study and clearly labels our study an experimental proof-of-concept study. Also, owing to the retrospective nature of this study, with several cases diagnosed in the early 2010s, molecular diagnoses as outlined in the 2021 WHO classification of brain tumors (CNS5) [41] were only available for some cases. This precluded training classifiers for molecular alterations (such as BRAF for pilocytic astrocytomas or the medulloblastoma subgroups [22]), which, of course, is an attractive future research direction given the importance of these markers for diagnosis and also, in part, therapeutic stratification. Second, segmentation and downstream patch-based classification are separate tasks in our pipeline. Despite efforts to improve segmentation, particularly for pediatric brain tumors in the 2023 BraTS challenge [42], jointly optimizing both tasks holds promise for further improving performance. With the broader availability of manually annotated pediatric brain tumor data sets, end-to-end optimized deep learning approaches for joint segmentation/detection and classification of pediatric brain tumors are therefore an attractive follow-up extension to our study. Lastly, when considering the possibilities of joint human and AI assessment of pediatric brain tumors, explainable AI strategies are another attractive avenue of research. Recent advances in joint learning representations from text (e.g., radiology reports) and medical images, such as BioMedCLIP [43], offer intriguing opportunities for improving the interaction between radiologists and deep learning models: by allowing the latter to offer textual explanation, the otherwise seemingly intractable decision made by a deep learning classifier can be retraced by a radiologist.

5. Conclusions

In summary, we developed and validated a robust deep learning model for the automated precise differentiation of medulloblastoma from pilocytic astrocytoma in a large, heterogeneous data set of pediatric brain tumor patients. Based on our proof-of-concept study, demonstrating how reliable, fully automated classifiers (including tumor segmentation) can be trained from heterogeneous multicenter data, our work highlights the potential of deep learning models to make this expert knowledge broadly available. In future studies, we will implement additional entities and prospective data sets to further validate our approach.

Author Contributions: Conceptualization, B.W., B.B., P.H. and M.F.; methodology, B.W. and B.B.; software, B.W.; formal analysis, B.W.; investigation, B.W., B.B., L.B., S.T., M.M. (Marie Metz), M.G. and M.F.; resources, B.W. and M.F.; data curation, M.J., P.-G.S., V.B., I.v.L., M.M. (Markus Metzler), P.J. and M.F.; writing—original draft preparation, B.W.; writing—review and editing, B.W., B.B., P.J., P.H. and M.F.; supervision, M.F. All authors have read and agreed to the published version of the manuscript.

References

1. Frühwald, M.C.; Rutkowski, S. Tumors of the Central Nervous System in Children and Adolescents. *Dtsch. Arzteblatt Int.* **2011**, *108*, 390–397. [CrossRef] [PubMed]

2. Pollack, I.F.; Agnihotri, S.; Broniscer, A. Childhood Brain Tumors: Current Management, Biological Insights, and Future Directions. *J. Neurosurg. Pediatr.* **2019**, *23*, 261–273. [CrossRef] [PubMed]

3. Avula, S.; Peet, A.; Morana, G.; Morgan, P.; Warmuth-Metz, M.; Jaspan, T.; European Society for Paediatric Oncology (SIOPE)-Brain Tumour Imaging Group. European Society for Paediatric Oncology (SIOPE) MRI Guidelines for Imaging Patients with Central Nervous System Tumours. *Childs Nerv. Syst. ChNS* **2021**, *37*, 2497–2508. [CrossRef] [PubMed]

4. Franceschi, E.; Hofer, S.; Brandes, A.A.; Frappaz, D.; Kortmann, R.-D.; Bromberg, J.; Dangouloff-Ros, V.; Boddaert, N.; Hattingen, E.; Wiestler, B.; et al. EANO-EURACAN Clinical Practice Guideline for Diagnosis, Treatment, and Follow-up of Post-Pubertal and Adult Patients with Medulloblastoma. *Lancet Oncol.* **2019**, *20*, e715–e728. [CrossRef] [PubMed]

5. Mittal, P. Magnetic Resonance Spectroscopy Findings in Non-Enhancing Desmoplastic Medulloblastoma. *Ann. Indian Acad. Neurol.* **2011**, *14*, 200–202. [CrossRef] [PubMed]

6. Fruehwald-Pallamar, J.; Puchner, S.B.; Rossi, A.; Garre, M.L.; Cama, A.; Koelblinger, C.; Osborn, A.G.; Thurnher, M.M. Magnetic Resonance Imaging Spectrum of Medulloblastoma. *Neuroradiology* **2011**, *53*, 387–396. [CrossRef] [PubMed]

7. De Menezes Jarry, V.; Pereira, F.V.; Dalaqua, M.; Duarte, J.Á.; França Junior, M.C.; Reis, F. Common and Uncommon Neuroimaging Manifestations of Ataxia: An Illustrated Guide for the Trainee Radiologist. Part 2—Neoplastic, Congenital, Degenerative, and Hereditary Diseases. *Radiol. Bras.* **2022**, *55*, 259–266. [CrossRef] [PubMed]

8. Liu, Y.; Xiao, B.; Li, S.; Liu, J. Risk Factors for Survival in Patients With Medulloblastoma: A Systematic Review and Meta-Analysis. *Front. Oncol.* **2022**, *12*, 827054. [CrossRef] [PubMed]

9. Chang, C.H.; Housepian, E.M.; Herbert, C. An Operative Staging System and a Megavoltage Radiotherapeutic Technic for Cerebellar Medulloblastomas. *Radiology* **1969**, *93*, 1351–1359. [CrossRef]

10. Meyers, S.P.; Wildenhain, S.L.; Chang, J.-K.; Bourekas, E.C.; Beattie, P.F.; Korones, D.N.; Davis, D.; Pollack, I.F.; Zimmerman, R.A. Postoperative Evaluation for Disseminated Medulloblastoma Involving the Spine: Contrast-Enhanced MR Findings, CSF Cytologic Analysis, Timing of Disease Occurrence, and Patient Outcomes. *Am. J. Neuroradiol.* **2000**, *21*, 1757–1765.

11. Pati, S.; Baid, U.; Edwards, B.; Sheller, M.; Wang, S.-H.; Reina, G.A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. Federated Learning Enables Big Data for Rare Cancer Boundary Detection. *Nat. Commun.* **2022**, *13*, 7346. [CrossRef] [PubMed]

12. Rudie, J.D.; Calabrese, E.; Saluja, R.; Weiss, D.; Colby, J.B.; Cha, S.; Hess, C.P.; Rauschecker, A.M.; Sugrue, L.P.; Villanueva-Meyer, J.E. Longitudinal Assessment of Posttreatment Diffuse Glioma Tissue Volumes with Three-Dimensional Convolutional Neural Networks. *Radiol. Artif. Intell.* **2022**, *4*, e210243. [CrossRef] [PubMed]

13. Kickingereder, P.; Isensee, F.; Tursunova, I.; Petersen, J.; Neuberger, U.; Bonekamp, D.; Brugnara, G.; Schell, M.; Kessler, T.; Foltyn, M.; et al. Automated Quantitative Tumour Response Assessment of MRI in Neuro-Oncology with Artificial Neural Networks: A Multicentre, Retrospective Study. *Lancet Oncol.* **2019**, *20*, 728–740. [CrossRef] [PubMed]

14. Vollmuth, P.; Foltyn, M.; Huang, R.Y.; Galldiks, N.; Petersen, J.; Isensee, F.; van den Bent, M.J.; Barkhof, F.; Park, J.E.; Park, Y.W.; et al. Artificial Intelligence (AI)-Based Decision Support Improves Reproducibility of Tumor Response Assessment in Neuro-Oncology: An International Multi-Reader Study. *Neuro Oncol.* **2023**, *25*, 533–543. [CrossRef] [PubMed]

15. Wen, P.Y.; van den Bent, M.; Youssef, G.; Cloughesy, T.F.; Ellingson, B.M.; Weller, M.; Galanis, E.; Barboriak, D.P.; de Groot, J.; Gilbert, M.R.; et al. RANO 2.0: Update to the Response Assessment in Neuro-Oncology Criteria for High- and Low-Grade Gliomas in Adults. *J. Clin. Oncol.* **2023**, *41*, JCO2301059. [CrossRef]

16. Erker, C.; Tamrazi, B.; Poussaint, T.Y.; Mueller, S.; Mata-Mbemba, D.; Franceschi, E.; Brandes, A.A.; Rao, A.; Haworth, K.B.; Wen, P.Y.; et al. Response Assessment in Paediatric High-Grade Glioma: Recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) Working Group. *Lancet Oncol.* **2020**, *21*, e317–e329. [CrossRef] [PubMed]

17. Fangusaro, J.; Witt, O.; Hernáiz Driever, P.; Bag, A.K.; de Blank, P.; Kadom, N.; Kilburn, L.; Lober, R.M.; Robison, N.J.; Fisher, M.J.; et al. Response Assessment in Paediatric Low-Grade Glioma: Recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) Working Group. *Lancet Oncol.* **2020**, *21*, e305–e316. [CrossRef] [PubMed]

18. van der Voort, S.R.; Incekara, F.; Wijnenga, M.M.J.; Kapsas, G.; Gahrmann, R.; Schouten, J.W.; Nandoe Tewarie, R.; Lycklama, G.J.; De Witt Hamer, P.C.; Eijgelaar, R.S.; et al. Combined Molecular Subtyping, Grading, and Segmentation of Glioma Using Multi-Task Deep Learning. *Neuro Oncol.* **2023**, *25*, 279–289. [CrossRef]

19. Eichinger, P.; Alberts, E.; Delbridge, C.; Trebeschi, S.; Valentinitsch, A.; Bette, S.; Huber, T.; Gempt, J.; Meyer, B.; Schlegel, J.; et al. Diffusion Tensor Image Features Predict IDH Genotype in Newly Diagnosed WHO Grade II/III Gliomas. *Sci. Rep.* **2017**, *7*, 13396. [CrossRef] [PubMed]

20. Zhang, B.; Chang, K.; Ramkissoon, S.; Tanguturi, S.; Bi, W.L.; Reardon, D.A.; Ligon, K.L.; Alexander, B.M.; Wen, P.Y.; Huang, R.Y. Multimodal MRI Features Predict Isocitrate Dehydrogenase Genotype in High-Grade Gliomas. *Neuro Oncol.* **2016**, *19*, 109–117. [CrossRef]

21. Kickingereder, P.; Neuberger, U.; Bonekamp, D.; Piechotta, P.L.; Götz, M.; Wick, A.; Sill, M.; Kratz, A.; Shinohara, R.T.; Jones, D.T.W.; et al. Radiomic Subtyping Improves Disease Stratification beyond Key Molecular, Clinical, and Standard Imaging Characteristics in Patients with Glioblastoma. *Neuro Oncol.* **2018**, *20*, 848–857. [CrossRef]

22. Zhang, M.; Wong, S.W.; Wright, J.N.; Wagner, M.W.; Toescu, S.; Han, M.; Tam, L.T.; Zhou, Q.; Ahmadian, S.S.; Shpanskaya, K.; et al. MRI Radiogenomics of Pediatric Medulloblastoma: A Multicenter Study. *Radiology* **2022**, *304*, 406–416. [CrossRef] [PubMed]

23. Wang, Y.; Wang, L.; Qin, B.; Hu, X.; Xiao, W.; Tong, Z.; Li, S.; Jing, Y.; Li, L.; Zhang, Y. Preoperative Prediction of Sonic Hedgehog and Group 4 Molecular Subtypes of Pediatric Medulloblastoma Based on Radiomics of Multiparametric MRI Combined with Clinical Parameters. *Front. Neurosci.* **2023**, *17*, 1157858. [CrossRef]

24. Rohlfing, T.; Zahr, N.M.; Sullivan, E.V.; Pfefferbaum, A. The SRI24 Multichannel Atlas of Normal Adult Human Brain Structure. *Hum. Brain Mapp.* **2009**, *31*, 798–819. [CrossRef] [PubMed]

25. Modat, M.; Cash, D.M.; Daga, P.; Winston, G.P.; Duncan, J.S.; Ourselin, S. Global Image Registration Using a Symmetric Block-Matching Approach. *J. Med. Imaging Bellingham Wash* **2014**, *1*, 024003. [CrossRef] [PubMed]

26. Isensee, F.; Schell, M.; Pflueger, I.; Brugnara, G.; Bonekamp, D.; Neuberger, U.; Wick, A.; Schlemmer, H.; Heiland, S.; Wick, W.; et al. Automated Brain Extraction of Multisequence MRI Using Artificial Neural Networks. *Hum. Brain Mapp.* **2019**, *40*, 4952–4964. [CrossRef]

27. Kofler, F.; Berger, C.; Waldmannstetter, D.; Lipkova, J.; Ezhov, I.; Tetteh, G.; Kirschke, J.; Zimmer, C.; Wiestler, B.; Menze, B.H. BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Front. Neurosci.* **2020**, *14*, 125. [CrossRef] [PubMed]

28. Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S.; et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv* **2021**, arXiv:2107.02314.

29. Thomas, M.F.; Kofler, F.; Grundl, L.; Finck, T.; Li, H.; Zimmer, C.; Menze, B.; Wiestler, B. Improving Automated Glioma Segmentation in Routine Clinical Use Through Artificial Intelligence-Based Replacement of Missing Sequences With Synthetic Magnetic Resonance Imaging Scans. *Investig. Radiol.* **2022**, *57*, 187–193. [CrossRef]

30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.

31. Johnson, R.A.; Miller, I.; Freund, J.E. *Miller & Freund's Probability and Statistics for Engineers*; Prentice Hall: Upper Saddle River, NJ, USA, 2011; ISBN 978-0-321-69498-0.

32. Madhogarhia, R.; Haldar, D.; Bagheri, S.; Familiar, A.; Anderson, H.; Arif, S.; Vossough, A.; Storm, P.; Resnick, A.; Davatzikos, C.; et al. Radiomics and Radiogenomics in Pediatric Neuro-Oncology: A Review. *Neuro-Oncol. Adv.* **2022**, *4*, vdac083. [CrossRef]

33. Arai, K.; Sato, N.; Aoki, J.; Yagi, A.; Taketomi-Takahashi, A.; Morita, H.; Koyama, Y.; Oba, H.; Ishiuchi, S.; Saito, N.; et al. MR Signal of the Solid Portion of Pilocytic Astrocytoma on T2-Weighted Images:Is It Useful for Differentiation from Medulloblastoma? *Neuroradiology* **2006**, *48*, 233–237. [CrossRef] [PubMed]

34. Kurokawa, R.; Kurokawa, M.; Baba, A.; Kim, J.; Capizzano, A.; Bapuraj, J.; Srinivasan, A.; Moritani, T. Differentiation of Pilocytic Astrocytoma, Medulloblastoma, and Hemangioblastoma on Diffusion-Weighted and Dynamic Susceptibility Contrast Perfusion MRI. *Medicine* **2022**, *101*, e31708. [CrossRef] [PubMed]

35. Esa, M.M.M.; Mashaly, E.M.; El-Sawaf, Y.F.; Dawoud, M.M. Diagnostic Accuracy of Apparent Diffusion Coefficient Ratio in Distinguishing Common Pediatric CNS Posterior Fossa Tumors. *Egypt. J. Radiol. Nucl. Med.* **2020**, *51*, 76. [CrossRef]

36. Pemberton, H.G.; Wu, J.; Kommers, I.; Müller, D.M.J.; Hu, Y.; Goodkin, O.; Vos, S.B.; Bisdas, S.; Robe, P.A.; Ardon, H.; et al. Multi-Class Glioma Segmentation on Real-World Data with Missing MRI Sequences: Comparison of Three Deep Learning Algorithms. *Sci. Rep.* **2023**, *13*, 18911. [CrossRef] [PubMed]

37. Dietterich, T. Overfitting and Undercomputing in Machine Learning. *ACM Comput. Surv.* **1995**, *27*, 326–327. [CrossRef]

38. Zhou, H.; Hu, R.; Tang, O.; Hu, C.; Tang, L.; Chang, K.; Shen, Q.; Wu, J.; Zou, B.; Xiao, B.; et al. Automatic Machine Learning to Differentiate Pediatric Posterior Fossa Tumors on Routine MR Imaging. *AJNR Am. J. Neuroradiol.* **2020**, *41*, 1279–1285. [CrossRef]

39. Liu, R.; Elhalawani, H.; Radwan Mohamed, A.S.; Elgohari, B.; Court, L.; Zhu, H.; Fuller, C.D. Stability Analysis of CT Radiomic Features with Respect to Segmentation Variation in Oropharyngeal Cancer. *Clin. Transl. Radiat. Oncol.* **2020**, *21*, 11–18. [CrossRef]

40. Quon, J.L.; Bala, W.; Chen, L.C.; Wright, J.; Kim, L.H.; Han, M.; Shpanskaya, K.; Lee, E.H.; Tong, E.; Iv, M.; et al. Deep Learning for Pediatric Posterior Fossa Tumor Detection and Classification: A Multi-Institutional Study. *AJNR Am. J. Neuroradiol.* **2020**, *41*, 1718–1725. [CrossRef] [PubMed]

41. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro-Oncology* **2021**, *23*, 1231–1251. [CrossRef]

42. Kazerooni, A.F.; Khalili, N.; Liu, X.; Haldar, D.; Jiang, Z.; Anwar, S.M.; Albrecht, J.; Adewole, M.; Anazodo, U.; Anderson, H.; et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). *arXiv* **2023**, arXiv:2305.17033v2.

43. Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; et al. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing. *arXiv* **2023**, arXiv:2303.00915.