This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Benchmarking Explainable AI for Brain Tumour Segmentation and Classification in African MRI

Ayotola Hannah Olabode[1][0009-0003-9257-0077], Aondona Moses Iorumbur[2,3,4][0009-0006-5621-0981], Confidence Raymond[2,5,6], and Olufade Williams Onifade[1]

[1] Department of Computer Science, University of Ibadan, Ibadan, Nigeria
[2] Medical Artificial Intelligence Laboratory, Lagos, Nigeria
[3] Department of Medical Physics, College of Medicine, University of Lagos, Lagos, Nigeria
[4] Department of Physics, Federal University of Technology Minna, Minna, Nigeria
[5] Montreal Neurological Institute, McGill University, Montreal, Canada
[6] Department of Biomedical Engineering, McGill University, Montreal, Canada

Abstract. Deep learning for brain tumour analysis holds significant promise for Sub-Saharan Africa, where specialist neuroradiologists are scarce and MRI infrastructure is limited. However, models trained on high-income country datasets are rarely evaluated for explainability on African MRI data, raising concerns about clinical trustworthiness in resource-constrained settings. We present a systematic evaluation of explainable AI (XAI) across a domain shift from BraTS 2021 (GLI, source) to BraTS-Africa (SSA, target), using a matched 146-subject design. We evaluate three 2D classifiers (ResNet18, EfficientNet-B0, ViT-Small) and four 3D segmentation models (UNet, Attention UNet, TransUNet, Swin-UNETR) under zero-shot and domain-adapted conditions, with XAI assessed via Grad-CAM, RISE, and Grad-CAM++ using overlap, faithfulness, and pointing game metrics. Transformer-based architectures consistently outperform convolutional models in cross-domain performance and explanation quality: ViT-Small achieves a pointing game accuracy of 0.655 on SSA despite comparable classification AUC to CNN models, and Swin-UNETR attains the best adaptation DSC (0.8411 ± 0.027) with the most coherent XAI profile, being the only model whose attributions are simultaneously faithful and anatomically aligned. Critically, high faithfulness does not imply trustworthy explanations: CNN models produce confounded attributions that are genuinely sensitive to the highlighted regions, yet those regions are not the lesion. These findings argue for mandatory XAI evaluation in clinical AI pipelines for African healthcare settings.

Keywords: Explainable AI · Brain Tumour Segmentation · Domain Adaptation · African MRI · BraTS-Africa · Grad-CAM · Trustworthy AI

1 Introduction

Brain tumours, particularly gliomas, remain difficult to diagnose, and Magnetic Resonance Imaging (MRI) is the primary imaging modality for their assessment owing to its soft-tissue contrast. However, manual interpretation is tedious and lacks consistency across readers [16], motivating automated deep learning approaches that now match expert performance on several diagnostic tasks [6,7]. Yet most models remain black boxes [13,14]. To address this, Explainable Artificial Intelligence (XAI) methods have been developed to provide insight into model decisions. Gradient-based approaches such as Grad-CAM [12] generate heatmaps indicating important image regions, while model-agnostic techniques such as SHAP [8] and LIME [11] explain predictions based on feature contributions. These methods have been widely applied in brain tumour MRI analysis to improve transparency [15], yet concerns remain regarding the faithfulness and reliability of the explanations they produce [5,14]. A further gap is the field's reliance on Western datasets such as BraTS, which under-represent the global variability of clinical MRI. African MRI data often differ in scanner generation, spatial resolution, and acquisition quality; curated resources such as BraTS-Africa [2,1] were introduced precisely to expose this imbalance. Strong performance on such data reportedly demands additional strategies including fine-tuning, data mixing, and ensembling [9], indicating that model robustness under African acquisition conditions remains an open challenge. Critically, this prior work targets segmentation accuracy alone; whether XAI methods retain their reliability under the same distribution shift is unexamined. We address this gap by quantitatively benchmarking explanation overlap, faithfulness, and localisation across architectures and cohorts. This study evaluates the accuracy and explainability of deep learning for brain tumour classification and segmentation under geographic domain shift, comparing CNN and Transformer architectures across global and Sub-Saharan African MRI cohorts. Our contributions are as follows:

- We establish a matched 146-subject benchmark pairing BraTS-GLI and BraTS-Africa, enabling direct cross-domain performance and XAI comparison.
- We demonstrate that transformer-based architectures consistently outperform CNN-based models in cross-domain generalisation, and that ViT-Small and Swin-UNETR yield the most coherent explanations for classification and segmentation respectively.
- We show that high faithfulness scores can be clinically misleading: CNN models exhibit confounded attributions where high model sensitivity co-occurs with near-zero anatomical overlap on SSA data.
- We provide actionable recommendations for XAI-aware model selection and deployment in resource-constrained African clinical settings.

2 Methods

2.1 Datasets

BraTS-GLI (Source Domain). We use the BraTS 2021 dataset [3], subsampling 146 of the 1,251 training subjects to match the target domain size for balanced cross-domain comparison. Each subject provides four co-registered modalities (T1, T1-Gd, T2, T2-FLAIR) with whole tumour (WT), tumour core (TC), and enhancing tumour (ET) annotations. The 146 subjects are partitioned into training and validation splits for source-domain model development.

BraTS-Africa (Target Domain). The target domain comprises 146 subjects from BraTS-Africa [1] (95 Glioma, 51 Other Neoplasms) from Sub-Saharan African sites, reflecting lower-field-strength scanners, variable protocols, and greater morphological heterogeneity than high-income cohorts. As BraTS-Africa lacks ET labels, all ET metrics are excluded from SSA evaluation. Subjects are split into five stratified cross-validation folds (117/29 train/validation, except fold 4 at 116/30).

For slice-level classification, we extract a manifest of representative axial, coronal, and sagittal 2D slices (438 per domain), preserving native class imbalance (GLI: ET = 354, ED = 73, NCR = 11; SSA: ED = 391, NCR = 47). SSA labels are binarised (ED = 1, NCR = 0) given annotation differences.

2.2 Preprocessing

Volumes are resampled to 1 mm^3 isotropic resolution, skull-stripped, z-score normalised per-modality over the brain mask foreground, then centre-cropped and stored as 4-channel NIfTI arrays. All 146 GLI and 146 SSA subjects are processed successfully, with zero failures. For 2D slice extraction, the single most informative slice per anatomical plane per subject is selected by maximum tumour mask area, normalised to $[0, 1]$, and resized to 224×224.

2.3 Classification Models

Three 2D slice-level classifiers, ResNet18, EfficientNet-B0, and ViT-Small, are trained on GLI training slices under a zero-shot cross-domain evaluation paradigm, using cross-entropy loss and the Adam optimiser (lr $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$) for up to 100 epochs with early stopping (patience = 10) on the GLI validation set. No fine-tuning on SSA data is performed; models are evaluated directly on all 438 SSA slices to assess zero-shot domain transfer.

2.4 Segmentation Models

Four 3D volumetric segmentation models, UNet, Attention UNet (Att-UNet), TransUNet, and Swin-UNETR, take 4-channel MRI inputs and produce 3-class segmentation outputs (WT, TC, ET). Training uses a composite Dice and cross-entropy loss, the Adam optimiser (lr $= 1 \times 10^{-4}$), and a batch size of 2, for up

to 200 epochs with early stopping (patience = 20) on GLI. Domain adaptation to SSA employs full fine-tuning of GLI-pretrained weights across all five cross-validation folds, with early stopping (patience = 20) and a maximum of 150 epochs per fold. Since SSA lacks ET annotations, all SSA adaptation and XAI metrics are reported over WT and TC only, with ET entries marked as not applicable.

2.5 XAI Methods

Classification XAI. Two post-hoc saliency methods are applied to each classifier: Grad-CAM [12] (gradient-weighted feature activations at a target layer) and RISE [10] (pixel importance from random binary mask probing). Target layers are the final residual block (ResNet18), last MBConv block (EfficientNet), and final transformer norm layer (ViT-Small). Three per-slice metrics are computed: **Overlap** (IoU between the top-50% saliency map and tumour mask), **Faithfulness** (Pearson correlation between saliency-ranked perturbation and confidence drop), and **Pointing Game** accuracy (proportion of slices where peak saliency falls within the tumour mask).

Segmentation XAI. Grad-CAM and Grad-CAM++ [4] are applied to all four segmentation models under zero-shot (GLI-pretrained) and adapted (SSA fine-tuned) conditions. Target layers are the network bottleneck block for CNN-based models, and the final transformer encoder norm layer (TransUNet) or last Swin stage norm layer (Swin-UNETR) for transformer-based models. XAI evaluation uses two metrics averaged across the 29 fold-0 SSA validation subjects: **Overlap** (IoU between the binarised saliency map and the segmentation mask), and **Faithfulness** (normalised change in Dice score when the top-k% salient voxels are occluded).

3 Experimental Setup

All experiments run on the Narval cluster using NVIDIA A100 SXM4 40 GB GPUs under SLURM scheduling, with Python 3.10, PyTorch 2.1, and MONAI 1.3 as the core software stack. Classification backbones are initialised with ImageNet-pretrained weights via `timm`; segmentation models are trained from scratch on GLI without external pretraining, and domain-adaptation runs are initialised from these GLI-trained checkpoints. Segmentation hyperparameters are fixed at learning rate 1×10^{-4}, batch size 2, Adam (default betas), composite Dice plus cross-entropy loss, and early stopping patience of 20 epochs. Classification uses learning rate 1×10^{-4}, batch size 32, and patience of 10 epochs. All seeds are fixed at 42 for reproducibility, and segmentation XAI metrics are computed on the fold-0 validation split (29 subjects).

4 Results

4.1 Zero-Shot Classification Transfer

Table 1 reports zero-shot classification performance for three models trained on GLI slices and evaluated on SSA slices without target-domain adaptation.

Table 1. Zero-shot classification performance on SSA (438 slices, binary ED vs. NCR).

Model	GLI Val AUC	SSA AUC	SSA Bal. Acc.	NCR Recall
ResNet18	1.000	0.653	0.500	0/47
EfficientNet-B0	0.982	0.641	0.499	0/47
ViT-Small	1.000	0.642	**0.597**	**41/47**

All three models achieve near-perfect GLI validation performance, confirming successful within-domain training. Cross-domain transfer to SSA reveals a dramatic collapse for CNN-based models: ResNet18 and EfficientNet predict ED for every slice, yielding 0.500 balanced accuracy and zero NCR recall. ViT-Small achieves substantially higher balanced accuracy (0.597) and recovers 41 of 47 NCR cases, suggesting its attention-based inductive bias generalises more robustly across the domain shift than convolutional feature extractors.

4.2 Segmentation on GLI

Table 2 reports segmentation performance of all four models trained and evaluated on the GLI domain.

Table 2. Segmentation performance on GLI validation set.

Model	DSC Mean	DSC WT	DSC TC	DSC ET	Time (h)
UNet	0.8386	0.9000	0.8469	0.7689	1.42
Att-UNet	0.8299	0.8911	0.8330	0.7656	0.92
TransUNet	**0.8889**	**0.9300**	**0.8915**	0.8373	1.65
Swin-UNETR	0.8859	0.9294	0.8542	**0.8742**	2.34

TransUNet achieves the highest mean DSC (0.8889), leading on WT and TC; Swin-UNETR is competitive overall (0.8859) and leads on ET (0.8742), likely reflecting the spatial inductive bias of hierarchical shifted-window attention for small high-signal regions. CNN-based models trail the transformer variants by roughly 0.05 mean DSC.

4.3 Domain Adaptation to SSA

Table 3 reports 5-fold cross-validation DSC (mean over WT and TC only) after full fine-tuning of GLI-pretrained weights on SSA data.

Table 3. SSA adaptation results (full fine-tuning, 5-fold CV, WT+TC mean DSC). ET excluded (no SSA annotation).

Model	F0	F1	F2	F3	F4	Mean ± Std
UNet	0.7284	0.7021	0.7493	0.7082	0.7768	0.7330 ± 0.028
Att-UNet	0.7322	0.7341	0.7247	0.7179	0.7482	0.7314 ± 0.011
TransUNet	0.8287	0.8004	0.8370	0.7389	0.8662	0.8142 ± 0.045
Swin-UNETR	0.8281	0.8294	0.8585	0.8104	0.8791	**0.8411 ± 0.027**

Swin-UNETR achieves the strongest adaptation performance (0.8411 ± 0.027), followed by TransUNet (0.8142 ± 0.045); CNN-based models plateau near 0.73, indicating a structural limitation in adapting convolutional representations to SSA with limited data. Fold 3 is consistently hardest across all models, and TransUNet's higher fold-to-fold variance (± 0.045 vs. ± 0.027 for Swin-UNETR) suggests greater sensitivity to subject sampling per fold.

4.4 Classification XAI

Table 4 reports XAI quality metrics for the three classifiers over all 438 SSA slices.

Table 4. Classification XAI summary (zero-shot on SSA, 438 slices).

Model	Method	Overlap	Faithfulness	Pointing Game
ResNet18	Grad-CAM	0.101 ± 0.105	−0.000 ± 0.000	0.030
ResNet18	RISE	0.114 ± 0.087	−0.001 ± 0.012	0.055
EfficientNet	Grad-CAM	0.196 ± 0.117	0.002 ± 0.032	0.244
EfficientNet	RISE	0.110 ± 0.084	0.002 ± 0.032	0.055
ViT-Small	Grad-CAM	**0.536 ± 0.256**	**0.242 ± 0.312**	**0.655**
ViT-Small	RISE	0.306 ± 0.251	0.191 ± 0.227	0.434

ViT-Small dominates all XAI metrics by a substantial margin, with a Grad-CAM pointing game accuracy of 0.655 (versus 0.030 for ResNet18), meaning the most salient pixel falls within the tumour mask for nearly two thirds of slices. This reveals a dissociation between predictive accuracy and explanation quality: CNN models predict using features unrelated to tumour anatomy, whereas

ViT attends to tumour-relevant regions even under uncertain classification confidence. ResNet18's near-zero faithfulness confirms its saliency maps are functionally uninformative, since perturbing highlighted regions produces no meaningful change in output.

4.5 Segmentation XAI

Table 5 reports Grad-CAM and Grad-CAM++ overlap and faithfulness for all four segmentation models under zero-shot and adapted conditions, evaluated on 29 fold-0 SSA subjects. ET is excluded throughout.

Table 5. Segmentation XAI results (fold-0 SSA validation, 29 subjects). ET excluded (no SSA annotation).

Model	Condition / Method	WT Ovl.	WT Faith	TC Ovl.	TC Faith
UNet	zero-shot, Grad-CAM	0.0001	0.788	0.0000	0.268
	zero-shot, Grad-CAM++	0.0083	0.144	0.0002	0.005
	adapted, Grad-CAM	0.0000	0.844	0.0000	0.381
	adapted, Grad-CAM++	0.0139	0.039	0.0000	0.005
Att-UNet	zero-shot, Grad-CAM	0.0063	0.575	0.0000	0.208
	zero-shot, Grad-CAM++	0.0043	0.031	0.0003	0.003
	adapted, Grad-CAM	0.0004	0.599	0.0000	0.316
	adapted, Grad-CAM++	0.0782	0.245	0.0094	0.086
TransUNet	zero-shot, Grad-CAM	0.0295	0.028	0.0036	0.008
	zero-shot, Grad-CAM++	0.0295	0.028	0.0036	0.008
	adapted, Grad-CAM	0.0267	0.040	0.0044	0.042
	adapted, Grad-CAM++	0.0267	0.040	0.0044	0.042
Swin-UNETR	zero-shot, Grad-CAM	0.0127	0.080	0.0032	0.004
	zero-shot, Grad-CAM++	0.0206	0.071	0.0054	0.010
	adapted, Grad-CAM	0.0227	**0.144**	0.0091	**0.150**
	adapted, Grad-CAM++	0.0144	0.124	0.0089	0.172

Three patterns emerge. First, overlap scores are universally low (0.000 to 0.078), reflecting the spatial coarseness of Grad-CAM maps projected from deep 3D feature spaces. Second, UNet and Att-UNet show an anomalous combination of near-zero overlap and high faithfulness under Grad-CAM (UNet adapted WT faithfulness: 0.844), indicating sensitivity to perturbations in highlighted regions that do not overlap the ground-truth mask; this decoupling suggests these models may exploit anatomical context cues rather than lesion-intrinsic features, potentially amplified by the SSA domain shift. Third, Swin-UNETR adapted achieves the most coherent XAI profile, with moderate and anatomically aligned overlap (WT: 0.0227, TC: 0.0091) paired with moderate faithfulness (WT: 0.144, TC: 0.150), the only model where both metrics are simultaneously meaningful. We

further note that, for TransUNet, Grad-CAM and Grad-CAM++ yield identical overlap and faithfulness values under both conditions.

5 Discussion

Transformer-based models consistently outperform CNN-based models in both predictive performance and explanation quality across the GLI-to-SSA shift: Swin-UNETR's hierarchical attention captures multi-scale context that transfers better than UNet's convolutional representations. Classification results underscore this: despite near-identical SSA AUC (0.641 to 0.653), only ViT-Small produces explanations spatially coherent with tumour anatomy (pointing game: 0.655 vs. 0.030 to 0.244), showing that accuracy alone is an insufficient criterion for clinical suitability. UNet and Att-UNet's high Grad-CAM faithfulness should be read with caution: it reflects sensitivity to the highlighted regions, not that those regions are the tumour, a form of confounded attribution where the decision boundary anchors to spurious features that co-vary with tumour presence in GLI but do not generalise to SSA. A clinician shown such an overlay would receive a plausible but anatomically misleading explanation. Fold 3's consistent difficulty across models may reflect fold-to-fold variation in diagnostic composition (Glioma versus Other Neoplasms); future work should stratify folds by diagnostic category. Key limitations include the sole use of Grad-CAM variants for segmentation attribution, since Score-CAM, SHAP, and LIME were excluded for computational cost; a perturbation-based baseline is needed to cross-validate these findings, especially the anomalous CNN faithfulness behaviour. TransUNet's identical Grad-CAM and Grad-CAM++ values suggest the two methods coincide at the queried transformer norm layer; querying intermediate attention layers is a direction for future work.

6 Conclusion

We evaluated explainable AI methods across a domain shift from BraTS-GLI to BraTS-Africa MRI. Transformer-based architectures showed superior cross-domain performance and more trustworthy explanations than CNN-based models in both tasks: ViT-Small achieved the highest classification XAI quality despite comparable predictive accuracy, and Swin-UNETR achieved the best adaptation DSC (0.8411 ± 0.027, WT+TC) alongside the most coherent segmentation attributions. Critically, high faithfulness does not imply interpretability, as UNet and Att-UNet produce confounded attributions where high sensitivity co-occurs with near-zero anatomical overlap on SSA data. These findings argue that XAI evaluation should be mandatory in clinical AI validation for low-resource settings, where the consequences of misplaced model trust are most severe.

7 Impact in Resource-Constrained Settings

Sub-Saharan Africa faces a critical shortage of specialist neuroradiologists, with MRI scanners concentrated in urban centres often without specialist interpreta-

tion support. Here, deep learning carries both promise and risk: reporting only DSC or AUC is insufficient for clinical deployment. A UNet adapted to SSA reaches 0.73 mean DSC yet produces saliency maps attending to non-tumour anatomy, whereas Swin-UNETR reaches a higher 0.84 DSC together with attributions spatially coherent with the lesion. DSC ranking alone would not reveal this difference in explanation quality, an actively misleading gap for a non-specialist reviewing an AI-generated report. We advocate three steps for RCS deployment: mandatory XAI audits on target-population data, since the degradation we observe would not surface from GLI-only evaluation; preference for uncertainty-aware models where a second opinion is unavailable; and architecture selection for XAI coherence under fine-tuning, not benchmark DSC alone, given Swin-UNETR's combination of strong adaptation and interpretable attributions. More broadly, BraTS-Africa remains underutilised for studying domain adaptation in low-resource contexts; our matched 146-subject protocol offers a reproducible baseline for future work extending to additional sites, modalities, and outcome data.

Acknowledgments. Computing resources were provided by the Digital Research Alliance of Canada (alliancecan.ca), through the Multimodal Imaging of Neurodegenerative Diseases (MiND) Lab, Montreal Neurological Institute, McGill University, Montreal, Canada.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adewole, M., Rudie, J., Gbadamosi, A., Zhang, D., Raymond, C., Ajigbotoshso, J., Toyobo, O., Aguh, K., Omidiji, O., Akinola, R., Suwaid, M., Emegoakor, A., Ojo, N., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Onuwaje, M., Betiku, B., Saluja, R., Menze, B., Baid, U., Bakas, S., Dako, F., Fatade, A., Anazodo, U.: Expanding the brain tumor segmentation (brats) data to include african populations (brats-africa) (2024). `https://doi.org/10.7937/v8h6-8x67`, version 1, Dataset
2. Adewole, M., Rudie, J., Gbadamosi, A., et al.: The brats-africa dataset: expanding the brain tumor segmentation data to capture african populations. Radiol. Artif. Intell. **7**(4), e240528 (2025)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F., Pati, S., Prevedello, L., Rudie, J., Sako, C., Shinohara, R., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Freymann, J., Kirby, J., Wiestler, B., Crivellaro, P., Colen, R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.A., Mahajan, A., Menze, B., Flanders, A., Bakas, S.: Rsna-asnr-miccai-brats-2021 dataset (2023). `https://doi.org/10.7937/jc8x-9874`
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018)

5. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
6. Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
7. Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sánchez, C.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
8. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 4765–4774 (2017)
9. Musah, T., et al.: Brain tumor segmentation in sub-sahara africa with advanced transformer and convnet methods: fine-tuning, data mixing and ensembling. arXiv preprint arXiv:2508.10905 (2025)
10. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC) (2018)
11. Ribeiro, M., Singh, S., Guestrin, C.: "why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
12. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
13. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. J. Imaging **6**(6), 52 (2020)
14. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Trans. Neural Netw. Learn. Syst. **32**(11), 4793–4813 (2020)
15. Yan, F., Chen, Y., Xia, Y., Wang, Z., Xiao, R.: An explainable brain tumor detection framework for mri analysis. Appl. Sci. **13**(6), 3438 (2023)
16. Zeineldin, R., Karar, M., Elshaer, Z., Coburger, J., Wirtz, C., Burgert, O., Mathis-Ullrich, F.: Explainability of deep neural networks for mri analysis of brain tumors. Int. J. Comput. Assist. Radiol. Surg. **17**(9), 1673–1683 (2022)