Do Medical Foundation Models Generalize on the African Brain?

Kaouther Mouheb, Gonzalo Esteban Mosquera Rojas, Juancito van Leeuwen, Stefan Klein, and Esther E. Bron

Dept. of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
k.mouheb@erasmusmc.nl

Abstract. Medical foundation models (FMs) are increasingly being used for brain MRI analysis. However, their evaluation remains dominated by high-resource datasets, leaving generalization to African cohorts underexplored. We assess whether FMs generalize equally to African and non-African brain MRI data across two tasks: dementia classification using a Nigerian dataset and brain tumor segmentation using BraTS-Africa. We evaluated two generalist FMs (BrainIAC, 3DINO) and two segmentation-specific FMs (MedSAM2, Medical-SAM2) against a from-scratch baseline. For classification, FMs provide limited gains (highest ROC-AUC of 0.86 with BrainIAC), whereas for segmentation they consistently improve performance, reaching up to 0.86 Dice with MedSAM2. Performance differences between African and non-African cohorts are inconsistent in direction suggesting that FMs do not show a systematic bias against African cohorts. This study also highlights the limited availability and diversity of African neuroimaging datasets as the main barrier to robust evaluation and deployment.

Keywords: Foundation Models · African Data · MRI · Neuroimaging.

1 Introduction

Medical image analysis is shifting from specialized fully-supervised models toward foundation models (FMs), which are trained on large-scale heterogeneous data to learn general-purpose representations. This allows FMs to have strong generalization across anatomies, modalities, and downstream tasks [14]. Although few FMs are specifically pretrained on brain scans, both brain-specific and general-purpose FMs are increasingly applied to neuroimaging. Currently, available FMs can be broadly divided into two categories: those trained specifically for segmentation, e.g. Med-SAM2 [10], and those trained as task-independent feature extractors with self-supervision, such as 3DINO [20]. On well-curated benchmarks such as the brain tumor segmentation challenge (BraTS) [11], FMs like BrainSegFounder [18] achieve Dice scores of up to 0.90, which is on par with specialized models, showing the potential of FMs for neuroimage analysis.

However, the training data of medical FMs is predominantly derived from high-resource regions, particularly North America and Europe [3,13]. This imbalance raises concerns about the external validity of the reported performance

and the ability of medical FMs to generalize when applied to underrepresented groups. African populations in particular remain largely absent from both the pretraining datasets and the evaluation benchmarks of current FMs, introducing a critical gap in current evaluation practices. In the context of brain MRI analysis, recent studies on Sub-Saharan African data demonstrate that models trained on standard datasets often exhibit degraded performance when applied to African cohorts [1]. These findings suggest that strong performance on established benchmarks does not necessarily translate to robust generalization on clinical African brain MRI data. To address this gap, recent efforts have introduced brain MRI datasets that reflect African clinical settings. The BraTS-Africa 2023 challenge released a multi-institutional Sub-Saharan MRI dataset for brain tumor segmentation [1] with heterogeneous acquisition protocols and image quality, and Wogu et al. [19] released a Nigerian brain MRI dataset for the study of neurodegenerative diseases. Despite these advances, a systematic evaluation of medical FMs on African brain MRI data is still missing, leaving an open question: *Is there a generalization gap between African and non-African populations when applying medical FMs to neuroimaging tasks?*

In this work, we present a comprehensive evaluation of four medical foundation models, spanning segmentation-specific and generalist architectures, on one classification and one segmentation task, using two Sub-Saharan African brain MRI datasets. We benchmark the performance against two high-resource datasets from North America and Europe. Through this study, we (i) quantify how well medical FMs generalize to African brain MRI data relative to high-resource cohorts, and (ii) underscore the importance of inclusive benchmarking for developing equitable medical FMs.

2 Materials and Methods

2.1 Datasets

To assess the generalization of FMs on African neuroimaging data, we used two publicly available datasets originating from Sub-Saharan African clinical centers. We compare the results to two high-resource datasets with the same task.

Dementia classification: For the classification task, we use a Nigerian brain MRI dataset [19] comprising T1-weighted scans from 50 subjects (22 female; age 48.8 ± 19.9), including dementia (DE, $n = 23$) and healthy control (HC, $n = 27$) cases. For comparison, we use OASIS-4, a U.S. clinical cohort. We include all 47 HC subjects and 50 randomly sampled DE cases to obtain a balanced dataset of 97 subjects (50 female; age 72.14 ± 9.4). All scans are preprocessed using HD-BET for brain extraction [7], N4 bias field correction [16], and FSL FLIRT for registration to the MNI152 space [4].

Brain tumor segmentation: For the segmentation task, we use the BraTS-Africa dataset [1], comprising 146 multi-parametric MRI studies (T1, T1 with Gadolinium, T2, and FLAIR) from Sub-Saharan African medical centers, including 95 glioma and 51 non-glioma cases. Demographic metadata (age, sex) is

not reported. We compare against the Erasmus Glioma Dataset (EGD), a high-resource counterpart from the Netherlands, using a size-matched random subset of 150 cases to control for sample size as a confounder, we refer to this subset as EGD-150. The remaining EGD subjects (n=625) are used to assess how fine-tuning sample size affects the performance of the generalistic FMs. Since BraTS-Africa provides tumor-subtype labels while EGD has only whole-tumor masks, we derive whole-tumor masks for BraTS-Africa by taking the union of its sub-region masks to keep the task consistent. Scans are released already preprocessed (co-registered, skull-stripped, resampled); we apply no further preprocessing.
All experiments use Monte-Carlo cross-validation (MCCV) with 5 random train-test splits stratified by class label, with 50% of the data in each split.

2.2 Models

We benchmark four FMs satisfying two selection criteria: (i) the model operates on 3D inputs, either via a volumetric encoder or by propagating predictions across a stack of 2D slices, and (ii) the model is pre-trained on medical data, including MRI. These comprise two generalist FMs, BrainIAC [15] and 3DINO [20], and two segmentation-specific FMs, MedSAM2 [10] and Medical-SAM2 [21]. Importantly, none of these models is reported to have used the datasets used in this study during pretraining.

2.3 Methods

This section details the pipelines applied in this study. For both tasks, we train conventional CNNs from scratch on the data as baselines using DenseNet121 for classification and a 3D U-Net for segmentation. **Classification:** We evaluate the models on the task of classifying DE vs. HC subjects. For segmentation-specific FMs we use the image encoder as the feature extractor. We explore two approaches of adapting FMs for a classification task:

1. **Linear Probing:** The image encoder is frozen and used to extract features from each scan. For models producing volumetric or slice-wise features, we apply global average pooling over the spatial dimensions to obtain a 1D feature vector. A linear classifier is then trained to perform the classification based on these 1D vectors.
2. **Parameter-efficient fine-tuning:** To assess whether moderate adaptation of the encoder can improve performance compared to linear probing, we apply parameter-efficient fine-tuning (PEFT), which updates only a small subset of parameters instead of the full model. Specifically, we use low-rank adaptation (LoRA) [6], which injects trainable low-rank matrices into each transformer block while keeping the pre-trained weights frozen, and jointly train the linear classifier. We apply LoRA to BrainIAC and 3DINO. MedSAM2 and Medical-SAM2 are excluded, as their large input size exceeded the available GPU memory for end-to-end fine-tuning.

Segmentation: For segmentation, segmentation-specific and generalist models are used in different ways:

1. **Segmentation-specific FMs:** We evaluate the zero-shot performance of SAM-based models in their native promptable setting, without task-specific fine-tuning. Following the prompting scheme described by the model developers, we use the bounding box of the tumor's middle slice as a 2D prompt, which the model's video-style decoder then propagates across slices. We explore two settings: (i) generating a probability map per sequence and averaging them to obtain the final mask, and (ii) using FLAIR only, since it is the most informative sequence for whole-tumor segmentation [17].
2. **Generalist FMs:** for BrainIAC and 3DINO, which are not promptable segmentation models, we attach a UNETR-style decoder [5] on top of the pre-trained encoder. The encoder is kept entirely frozen, including its normalization layers, and only the decoder parameters are optimized. Training uses the combined Dice and cross-entropy loss with the AdamW optimizer [9], a learning rate of 10^{-4}, and 50 epochs. Since the models take only 1 channel as input we use the FLAIR sequence in this experiment. For EGD, we first perform the experiments using the size-matched subset EGD-150 (training size = 75 in each fold). In a second experiment, we add the remaining 625 samples to the training samples of each MCCV split to evaluate whether increasing the training set size could improve performance.

Full implementation details, including data augmentation and all remaining hyperparameters, are available in our public repository[1].

Performance Evaluation: We use area under the receiver operating characteristic curve (ROC-AUC) for classification, and the Dice score (DSC) for segmentation. Given the small number of splits (n=5), which limits the statistical power of formal hypothesis testing, we adopt a descriptive approach: results are summarized as mean ± standard deviation across splits, and differences between models (both relative to the from-scratch baseline and pairwise across FMs) are assessed in terms of mean performance differences.

3 Results

3.1 Classification

Table 1 reports the ROC-AUC for all models on the classification task. On the Nigerian Brain dataset, the DenseNet121 baseline achieves a ROC-AUC of 0.85 ± 0.10. Using linear probing, BrainIAC obtains the highest average ROC-AUC (0.85 ± 0.06), matching the baseline, followed by 3DINO with 0.80 ± 0.03, MedSAM2 with 0.66 ± 0.09, and Medical-SAM2 with 0.59 ± 0.07. With LoRA fine-tuning, the performance improves slightly with 0.86 ± 0.05 for BrainIAC and

[1] https://gitlab.com/radiology/neuro/MFM-African-Brain

Table 1. ROC-AUC of the dementia classification task (mean$_{\pm\text{std}}$ across the test sets of 5 MCCV splits).

Model	Nigerian Brain	OASIS-4
DenseNet121 baseline	$0.85_{\pm 0.10}$	$0.80_{\pm 0.03}$
Linear probing		
BrainIAC	$0.85_{\pm 0.06}$	$0.80_{\pm 0.05}$
3DINO	$0.80_{\pm 0.03}$	$0.83_{\pm 0.06}$
MedSAM2	$0.66_{\pm 0.09}$	$0.73_{\pm 0.06}$
Medical-SAM2	$0.59_{\pm 0.07}$	$0.68_{\pm 0.03}$
Fine-tuning		
BrainIAC + LoRA	$0.86_{\pm 0.05}$	$0.81_{\pm 0.02}$
3DINO + LoRA	$0.82_{\pm 0.05}$	$0.84_{\pm 0.04}$

Table 2. Tumor segmentation Dice score for segmentation-specific FMs obtained with zero-shot prompting. "Combined" refers to the results obtained by averaging the output probability maps of the four sequences. Results are reported as mean$_{\pm\text{std}}$ across the test sets of 5 Monte-Carlo splits.

	BraTS-Africa		EGD-150	
Model	**Combined**	**FLAIR**	**Combined**	**FLAIR**
Baseline (U-Net)	$0.21_{\pm 0.31}$	$0.55_{\pm 0.09}$	$0.34_{\pm 0.20}$	$0.62_{\pm 0.08}$
MedSAM2	$0.86_{\pm 0.01}$	$0.86_{\pm 0.01}$	$0.82_{\pm 0.02}$	$0.82_{\pm 0.01}$
Medical-SAM2	$0.62_{\pm 0.01}$	$0.57_{\pm 0.02}$	$0.71_{\pm 0.01}$	$0.70_{\pm 0.01}$

0.82 ± 0.05 for 3DINO. On OASIS-4, the DenseNet121 baseline achieves a ROC-AUC of 0.80 ± 0.03. With linear probing, 3DINO obtains the highest average ROC-AUC at 0.83 ± 0.06, followed by BrainIAC at 0.80 ± 0.05, MedSAM2 at 0.73 ± 0.06, and Medical-SAM2 at 0.68 ± 0.03. With LoRA fine-tuning, both models show a slight performance increase where 3DINO reaches 0.84 ± 0.04 and BrainIAC reaches 0.81 ± 0.02. Since Nigerian Brain and OASIS-4 are different datasets, the ROC-AUC values cannot be compared one-to-one to claim that a model performs strictly better or worse on one versus the other. However, the gap between the two datasets varies considerably between models. For segmentation-specific FMs, the gap is relatively larger (up to 9%) compared to generalist models.

3.2 Segmentation

Table 2 summarizes the performance of segmentation-specific FMs in terms of DSC across two settings: Combined (using all modalities) and FLAIR. In BraTS-Africa, MedSAM2 achieves a DSC of 0.86 ± 0.01 in both settings, substantially higher than the U-Net baseline of 0.21 ± 0.31 for Combined and 0.55 ± 0.09 for FLAIR. Medical-SAM2 achieves 0.62 ± 0.01 (Combined) and 0.57 ± 0.02 (FLAIR),

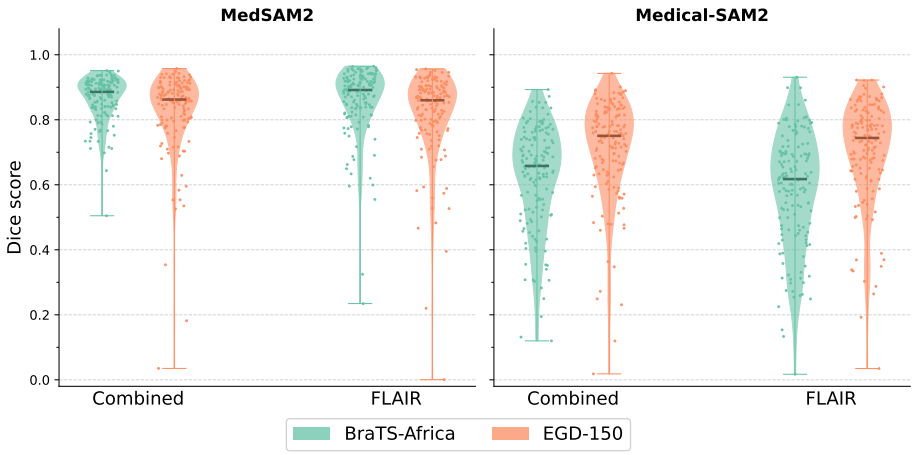

Fig. 1. Per-case Dice score distributions for zero-shot tumor segmentation on BraTS-Africa ($n = 146$) and EGD-150. Combined denotes averaging the output probability maps across all four MRI sequences; FLAIR denotes using the FLAIR sequence only.

Table 3. Tumor segmentation Dice score for generalist FMs, using the FLAIR sequence only (mean$_{\pm\text{std}}$ across 5 Monte-Carlo splits).

Model	BraTS-Africa	EGD-150	EGD-large
Baseline (U-Net)	$0.55_{\pm 0.09}$	$0.62_{\pm 0.08}$	$0.82_{\pm 0.02}$
BrainIAC + Decoder	$0.63_{\pm 0.02}$	$0.62_{\pm 0.03}$	$0.80_{\pm 0.01}$
3DINO + Decoder	$0.73_{\pm 0.02}$	$0.78_{\pm 0.01}$	$0.82_{\pm 0.01}$

also higher than the baseline. In EGD-150, MedSAM2 achieves 0.82 ± 0.02 for Combined and 0.82 ± 0.01 for FLAIR, while Medical-SAM2 achieves 0.71 ± 0.01 for Combined and 0.70 ± 0.01 for FLAIR, both exceeding the baseline performance of 0.34 ± 0.20 (Combined) and 0.62 ± 0.08 (FLAIR). Both FMs show similar performance between Combined and FLAIR configurations within each dataset, with differences of at most 0.05 DSC. The U-Net baseline, however, demonstrates pronounced sensitivity to settings, showing substantially lower DSC in Combined versus FLAIR: 0.21 ± 0.31 vs. 0.55 ± 0.09 in BraTS-Africa and 0.34 ± 0.20 vs. 0.62 ± 0.08 in EGD-150. Figure 1 shows zero-shot DSC distributions on the full BraTS-Africa dataset ($n = 146$) and EGD-150. For MedSAM2, BraTS-Africa displays tightly concentrated distributions in both settings, with most cases scoring above 0.80 and few low-scoring outliers, whereas EGD-150 shows wider spread with a longer lower tail toward 0 despite similar median values. This pattern is reversed for Medical-SAM2: EGD-150 cases are distributed around a higher median (~ 0.75) than BraTS-Africa.

The segmentation results for generalist FMs are reported in Table 3. On BraTS-Africa, both generalist models outperform the U-Net baseline (0.55 ± 0.09)

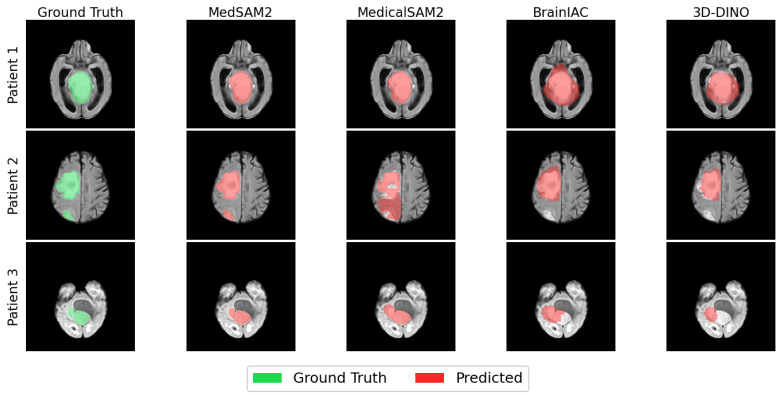

Fig. 2. Ground truth segmentation and model outputs for three example cases from BraTS-Africa. Results are obtained using the FLAIR sequence as input.

with BrainIAC reaching 0.63 ± 0.02 and 3DINO reaching 0.73 ± 0.02. On EGD-150, BrainIAC (0.62 ± 0.03) performs comparably to the baseline (0.62 ± 0.08), while 3DINO (0.78 ± 0.01) outperforms it. Comparing the two datasets in the size-matched setting, the performance is similar between BraTS-Africa and EGD-150 for all three models, with differences of at most 0.05 DSC. With the larger EGD training set, all models improve substantially, reaching 0.80 ± 0.01 for BrainIAC and 0.82 ± 0.01 for 3DINO. In this setting, both models perform comparably to the from-scratch baseline (0.82 ± 0.02). Qualitative results on BraTS-Africa (Fig. 2) show that all four models generally localize the tumor region correctly, but the predictions tend to be less accurate compared to the ground-truth. MedSAM2 produces the closest match to the ground-truth lesion boundary among the four models across the three cases.

4 Discussion

In this work, we evaluated whether the generalization of medical foundation models extends equally to African and non-African neuroimaging populations, across one classification and one segmentation task.

When assessing the benefit of FMs over training from scratch, we observed different patterns across tasks. In classification, FMs generally do not offer large improvements compared to training from scratch, whereas in segmentation all FMs outperformed the U-Net baseline. This suggests that pre-trained features transfer well to spatial localization but less to whole-volume classification of subtle pathologies such as dementia. This conclusion holds equally for non-African data. Previous benchmarks have shown similar results; for instance, Li et al. found that FMs do not generalize to tasks with subtle pathologies in X-ray analysis [8]. LoRA fine-tuning yielded only modest gains over linear probing for both African and non-African data. This could be due to the small size of the fine-

tuning data. For both African and non-African data, SAM2-based models showed strong segmentation performance yet performed poorly at classification, possibly because they are optimized for spatial localization rather than for learning globally discriminative semantic representations. Recent work has similarly reported that SAM2 representations may be entangled with localized task-specific cues that limit higher-level semantic understanding [2].

When comparing African and non-African cohorts under matched sample-size, the performance gap between datasets varied across models, but its direction was not consistent. While in some experiments FMs exhibited higher performance on non-African datasets, in most experiments we observed similar patterns between the two populations suggesting that the tested FMs do not show a systematic bias toward the high-resource population. Increasing the EGD training set size substantially improved segmentation performance, suggesting that the advantage of high-resource datasets may partly reflect larger sample sizes rather than higher data quality. However, these factors cannot be disentangled in our study, as they co-vary across cohorts. Further research could use explainable AI methods to disentangle performance differences related to data quality from those related to population differences. We further showed that with a larger training set, FMs no longer outperformed the from-scratch baseline, indicating that their benefit is greatest in limited-data settings. This is a valuable result in this context given the scarcity of African datasets. Promptable segmentation-specific FMs are well suited to this setting, as their strong zero-shot performance allows them to be used without fine-tuning. Nevertheless, this class of models is highly dependent on prompt quality. In this work, we relied on ground-truth-derived bounding boxes, which may not be available in clinical practice. When some annotated data are available, training a segmentation decoder on top of generalist FMs offers a strong prompt-free alternative.

Some limitations of this study should be considered. First, experiments were conducted on an H100 GPU, which is not representative of typical compute resources in African settings. However, the large parameter count of FMs does not necessarily imply high deployment costs, as adapting a frozen encoder requires training only a lightweight probe or decoder. As a result, FM adaptation can be more computationally efficient than training smaller task-specific models from scratch while achieving comparable performance [12]. Second, the African datasets used are small, reflecting the broader scarcity of publicly available African neuroimaging data, which limits the strength of comparative claims and motivated a descriptive analysis. Third, we observed a relatively high classification performance on the Nigerian Brain MRI dataset, although its sample size (n=50) is smaller than that of OASIS-4 (n=97). This performance should be interpreted with caution due to a substantial age gap between the dementia and healthy groups in this dataset, raising the possibility that the models rely in part on age-related features rather than pathology-specific ones. More broadly, African datasets often exhibit design limitations, including demographic confounders and missing metadata, which restrict both subgroup evaluation and the interpretability of model behavior.

In conclusion, a more rigorous evaluation of the true value of FMs still requires addressing the data limitations we identified in this work, through larger, better-annotated, and more equitably designed African neuroimaging datasets. However, this study highlights the potential of medical FMs to generalize effectively to African populations, with the largest gains observed in the low-data regime that characterizes the African medical imaging landscape.

Acknowledgments. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-15942.

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.D., Gbadamosi, A., Zhang, D., Raymond, C., Ajigbotoshso, J., Toyobo, O., Aguh, K., Omidiji, O., Akinola, R., et al.: The brats-africa dataset: expanding the brain tumor segmentation data to capture african populations. Radiology: Artificial Intelligence **7**(4), e240528 (2025)
2. Cuttano, C., Trivigno, G., Averta, G., Masone, C.: Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation. Advances in Neural Information Processing Systems **38**, 118923–118957 (2026)
3. Danaei, S., Dehghanian, Z., Meftah, E., Naderi, N., Safavi-Naini, S.A.A., Khorasanizade, F., Rabiee, H.R.: State of abdominal ct datasets: A critical review of bias, clinical relevance, and real-world applicability. arXiv preprint arXiv:2508.13626 (2025)
4. Fischer, B., Modersitzki, J.: Flirt: A flexible image registration toolbox. In: International workshop on biomedical image registration. pp. 261–270. Springer (2003)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
7. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al.: Automated brain extraction of multisequence mri using artificial neural networks. Human brain mapping **40**(17), 4952–4964 (2019)
8. Li, F., Dapamede, T., Chavoshi, M., Jeon, Y.S., Khosravi, B., Dere, A., Brown-Mulry, B., Isaac, R.S., Mansuri, A., Sanyika, C., et al.: Feature quality and adaptability of medical foundation models: A comparative evaluation for radiographic classification and segmentation. arXiv preprint arXiv:2511.09742 (2025)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
10. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)

11. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
12. Mouheb, K., Elbatel, M., Papma, J., Biessels, G.J., Claassen, J., Middelkoop, H., van Munster, B., van der Flier, W., Ramakers, I., Klein, S., et al.: Federated fine-tuning of sam-med3d for mri-based dementia classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 69–79. Springer (2025)
13. Musa, A., Prasad, R., Onwualu, P., Hernandez, M.: A systematic review of cross-population shifts in medical imaging analysis with deep learning. Big Data and Cognitive Computing **10**(3), 76 (2026)
14. Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A.: Foundation models in radiology: what, how, why, and why not. Radiology **314**(2), e240597 (2025)
15. Tak, D., Garomsa, B.A., Zapaishchykova, A., Chaunzwa, T.L., Climent Pardo, J.C., Ye, Z., Zielke, J., Ravipati, Y., Pai, S., Vajapeyam, S., et al.: A generalizable foundation model for analysis of human brain mri. Nature Neuroscience pp. 1–12 (2026)
16. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE transactions on medical imaging **29**(6), 1310–1320 (2010)
17. van der Voort, S.R., Incekara, F., Wijnenga, M.M., 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-oncology **25**(2), 279–289 (2023)
18. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: a vision foundation model for general-purpose segmentation on volumetric medical images. IEEE Transactions on Neural Networks and Learning Systems (2025)
19. Wogu, E., Filima, P., Caron, B., Deabler, D., Herholz, P., Leal, C., Mehboob, M.F., Kim, S., Gosain, A., Flexwala, A., et al.: A labeled clinical-mri dataset of nigerian brains. Scientific Data **12**(1), 518 (2025)
20. Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025)
21. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)