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.

Evaluating the Generalization of Neuroimaging Foundation Models on African Brain MRI

Oluwatobi Iyanuoluwa Akinmuleya*[1], Olatokun Shamsudeen Akano*[2], Samuel Danquah Ankapong[3], Olamide Lawal[4], and Toufiq Musah[5]

[1] Department of Medicine, Shenyang Medical College, Shenyang, China
[2] College of Medicine, University of Ibadan, Ibadan, Nigeria
[3] University of Ghana, Accra, Ghana
[4] Obafemi Awolowo University, Ile-Ife, Nigeria
[5] Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Abstract. Neuroimaging foundation models pretrained on large, predominantly western cohorts are increasingly proposed as general-purpose backbones for brain MRI analysis. Yet, their ability to generalize to underrepresented clinical populations remains largely untested. We evaluate four recent foundation models (BrainIAC, NeuroJEPA, NeuroVFM, and Primus) on a three-way diagnostic classification task (Control, Dementia, Parkinson's disease) using a cohort of 88 subjects from a Nigerian clinical brain MRI dataset, across four modality configurations (T1w, T2w, T1w+T2w, FLAIR), and compare against an end-to-end trained ViT3D baseline. All four frozen backbones collapse to majority-class predictions, while NeuroJEPA on FLAIR shows modest but still limited discrimination. In contrast, the end-to-end trained ViT3D achieves higher accuracy and MCC on every task (up to 53.4% accuracy, MCC=0.27) and is the only model with non-trivial recall. Our findings suggest that these frozen neuroimaging foundation models are insufficient for fine-grained diagnostic classification in small, non-western clinical cohorts, motivating parameter-efficient adaptation and broader multi-site external validation for equitable deployment in global health settings.

Keywords: Brain MRI · Foundation Models · Domain Generalization · Deep learning · Explainability

1 Introduction

Diagnostic infrastructure for brain imaging remains critically underdeveloped across the continent [10, 17]. Fewer than 1% of global dementia research data originate from Africa [2, 11]. At the same time, brain MRI foundation models are increasingly pretrained on large datasets drawn predominantly from well-resourced, non-African healthcare systems, and positioned as general-purpose representations for downstream clinical tasks. Whether these representations

* Equal contribution

generalize to the heterogeneous acquisition conditions and clinical populations encountered in African settings therefore remains unclear.

Automated detection of neurodegenerative disease has become an active application area for AI in neuroimaging [4]. Alzheimer's disease, Parkinson's disease, and other dementia subtypes share overlapping structural signatures, including cortical and subcortical atrophy patterns that vary gradually and unevenly across regions [3]. These overlapping patterns can make early or borderline cases difficult to distinguish from routine clinical imaging alone. Machine learning approaches have been proposed to learn discriminative patterns directly from neuroimaging data [4]. However, conventional supervised approaches typically require task-specific training data and can be difficult to develop reliably in small, imbalanced clinical cohorts [1].

This limitation is especially acute in African clinical contexts, where neuroimaging datasets remain small and unevenly distributed across diagnostic centers [17]. Data are also frequently incomplete, with missing sequences, heterogeneous field strengths, and demographic imbalance across diagnostic groups [16]. Under these conditions, training a dedicated deep learning model from scratch for each new clinical task or site is often impractical. Foundation models pretrained on large, predominantly non-African cohorts have been proposed as an alternative: a frozen backbone paired with a lightweight classifier could, in principle, transfer useful representations to a new population using only a small labeled set [12]. However, this assumption has rarely been tested outside the well-curated datasets on which models such as BrainIAC [13] and NeuroJEPA [5] were developed. Whether these learned representations transfer to heterogeneous, resource-constrained clinical data from African populations therefore remains an open question.

In this work, we evaluate whether representations learned by four pretrained brain MRI foundation models transfer to an African clinical cohort under a frozen-backbone, low-data setting. We compare them with a 3D ViT trained from scratch, evaluate performance across T1w, T2w, FLAIR, and multimodal inputs, and examine spatial attribution maps to characterize model behavior beyond aggregate classification metrics. Our findings provide insights into the strengths and limitations of these foundation models, highlighting their applicability and potential challenges in underrepresented clinical populations.

2 Method

2.1 Dataset and Preprocessing

We used the Nigerian Brain MRI dataset [16] which contains a total of 761 MRI sequences collected from 88 participants across three diagnostic centers in Nigeria: 31 participants with dementia (65.26 ± 14.11 years, range: 36–86), 22 participants with Parkinson's disease (61.14 ± 14.60 years, range: 36–79), and 35 healthy controls (34.14 ± 12.75 years, range: 12–75). Data were acquired on scanners with field strengths ranging from 0.3 T to 1.5 T. Not all participants

had all three MRI sequences (T1w, T2w, FLAIR) available. Eleven control participants had one or more missing MRI sequences; consequently, the number of participants varied by sequence: T1w ($n = 79$), T2w ($n = 80$), T1w+T2w ($n = 75$), and FLAIR ($n = 80$).

The dataset contains multiple intrasession acquisitions for some subjects. To ensure consistent evaluation, we used BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [8], a no-reference image quality assessment metric, to identify and select the acquisistion with the highest BRISQUE score in the evaluation. The dataset was skull-stripped using HD-BET [6]. Additional preprocessing was applied as required by the specific foundation model used for evaluation.

2.2 Modeling

We evaluate four pretrained 3D foundation models under a frozen-backbone protocol, alongside ViT3D as a from-scratch baseline. Each model processes the input volume into a sequence of patch tokens that serve as input to the probing head. The pretrained weights are loaded from published checkpoints and remain frozen throughout training. This isolates the contribution of each model's learned representations from the capacity added by the probe.

NeuroVFM [7] is a ViT-B pretrained on 5.24 million brain CT and MRI volumes using Vol-JEPA, a volumetric JEPA objective. It learns a shared latent representation across CT and MRI through masked volumetric prediction, providing a large-scale multimodal pretraining paradigm for downstream neuroimaging tasks.

Neuro-JEPA [5] is a ViT-B pretrained on 1,551,862 multimodal MRI scans (T1, T2, and FLAIR) using a JEPA objective, with mixture-of-experts routing. It produces patch-level representations without a CLS token and uses a sparse MoE architecture to encode multimodal brain MRI.

BrainIAC [13] is a ViT-B pretrained on 32,015 multiparametric brain MRIs from 16 datasets spanning 10 neurological conditions using SimCLR contrastive learning. It produces 768-dimensional representations from the CLS token for downstream adaptation.

Primus [14] is a custom transformer encoder–convolutional decoder architecture pretrained on the OpenMind dataset [15] using a Volume Contrastive (VoCo) [18] learning objective. We use only the pretrained encoder and discard the decoder.

ViT3D serves as the from-scratch baseline and is trained end-to-end rather than frozen.

For single-modality inputs the forward pass is straightforward: the volume goes through the backbone's patch embedding, through the transformer encoder, and outputs a tensor of patch tokens. For multi-modal inputs, the first convolutional layer of the patch embedding must accept more than one input channel. The original weights were learned on single-channel data. To extend them to multi-channel inputs, we replicate the weight tensor across the new channels and divide by the channel count. This adaptation is applied for every multi-modal configuration before training begins.

2.3 Linear Probing Protocol

For each pretrained backbone, we train a linear classifier on the frozen features. The probing head consists of dropout followed by a single linear projection: Dropout($p = 0.3$) $\rightarrow$ Linear(hidden_dim, 3). The hidden dimension depends on the model: 768 for ViT-based backbones (Neuro-JEPA, NeuroVFM, BrainIAC), 864 for Primus, and 384 for ViT3D-B/16.

All heads use the same training setup. We optimize with AdamW at a learning rate of 0.001 and weight decay of 1×10^{-4}. The learning rate warms up linearly from 1% to 100% over five epochs, then decays following a cosine schedule to 1% of the initial rate. The loss is cross-entropy with label smoothing of 0.1, which acts as a regularizer against overconfidence. Training runs for a maximum of 50 epochs with a batch size of 4 in a 5-fold cross validation configuration.

2.4 Evaluation

We report five metrics per fold. Accuracy captures overall correctness, while balanced accuracy and the Matthews correlation coefficient (MCC) account for the moderate class imbalance in our cohort. Macro F1 averages per-class performance without weighting by support, while weighted F1 weights by class size. The macro area under the receiver-operating characteristic curve (AUC-OVR) measures per-class separability independent of any decision threshold, computed in a one-versus-rest fashion.

Five models are evaluated under identical conditions. Neuro-JEPA [5], NeuroVFM [7], BrainIAC [13], and Primus [14] are used as frozen backbones with a linear probe trained on top, while ViT3D is trained end-to-end from scratch as a within-architecture baseline.

2.5 Explainability

We applied gradient-weighted class activation mapping (Grad-CAM) to obtain spatial attribution maps for model predictions. For each subject, gradients of the predicted-class logit with respect to the model's spatial feature representations were used to estimate feature importance, which was projected back to the corresponding patch locations and upsampled to the original voxel resolution. For multimodal inputs, each modality was evaluated independently with the remaining modalities masked to zero, producing modality-specific attribution maps. Subject-level attribution maps were averaged within each class and subsequently min–max normalized.

3 Experimental Results

3.1 Classification performance

Table 1 reports five-fold classification performance across the four modality configurations. ViT3D, trained end-to-end, generally outperforms the frozen

foundation-model backbones across the evaluated tasks and metrics. Its strongest overall performance is observed on the T1w+T2w fusion setting, achieving an accuracy of 0.533, balanced accuracy of 0.482, macro-F1 of 0.434, MCC of 0.272, and AUC of 0.672. Neuro-JEPA attains a higher AUC than ViT3D on T1w+T2w (0.676 vs. 0.672) and FLAIR (0.684 vs. 0.675). On FLAIR, Neuro-JEPA achieves the strongest overall performance among the frozen models, with an accuracy of 0.448, balanced accuracy of 0.409, macro-F1 of 0.324, MCC of 0.128, and AUC of 0.684. The remaining frozen foundation-model backbones generally show weaker classification performance.

Table 1: **Classification Performance on African Brain MRI.** Mean over 5 subject-level cross-validation folds. **Bold** = best, underline = second best. Cell shading indicates relative performance.

(a) Single-Modality MRI

Model	T1w ($n = 79$)					T2w ($n = 80$)				
	Acc	BalAcc	F1	MCC	AUC	Acc	BalAcc	F1	MCC	AUC
BrainIAC [13]	0.381	0.326	0.207	-0.022	0.587	0.375	0.322	0.194	-0.025	0.559
Neuro-JEPA [5]	0.382	0.331	0.205	-0.017	0.548	0.402	0.351	0.233	0.032	0.515
NeuroVFM [7]	0.392	0.333	0.188	0.000	0.535	0.361	0.333	0.176	0.000	0.566
Primus [14]	0.393	0.339	0.230	-0.001	0.564	0.323	0.278	0.169	-0.158	0.439
ViT3D	**0.507**	**0.465**	**0.407**	**0.241**	**0.651**	**0.462**	**0.407**	**0.308**	**0.141**	**0.634**

(b) Multimodal and FLAIR MRI

Model	T1w+T2w ($n = 75$)					FLAIR ($n = 80$)				
	Acc	BalAcc	F1	MCC	AUC	Acc	BalAcc	F1	MCC	AUC
BrainIAC [13]	0.400	0.322	0.190	-0.043	0.543	0.351	0.311	0.173	-0.090	0.533
Neuro-JEPA [5]	0.426	0.347	0.218	0.039	**0.676**	0.448	0.409	0.324	0.128	**0.684**
NeuroVFM [7]	0.413	0.333	0.195	0.000	0.556	0.376	0.333	0.182	0.000	0.598
Primus [14]	0.413	0.333	0.195	0.000	0.510	0.401	0.356	0.242	0.051	0.580
ViT3D	**0.533**	**0.482**	**0.434**	**0.272**	0.672	**0.466**	**0.436**	**0.411**	**0.175**	0.675

3.2 Per-class performance

Table 2 reports per-class F1 scores on the T1w+T2w configuration. The frozen backbones show a bias toward the Dementia class, with BrainIAC, NeuroVFM, and Primus achieving zero F1 for Parkinson's and very low F1 for Control (0.028, 0.021, and 0.117, respectively). Neuro-JEPA is a partial exception, achieving non-zero F1 for both Control (0.126) and Parkinson's (0.052). In contrast, ViT3D is the only model with a relatively non-trivial F1 across all three classes (Control 0.412, Dementia 0.629, Parkinson 0.128).

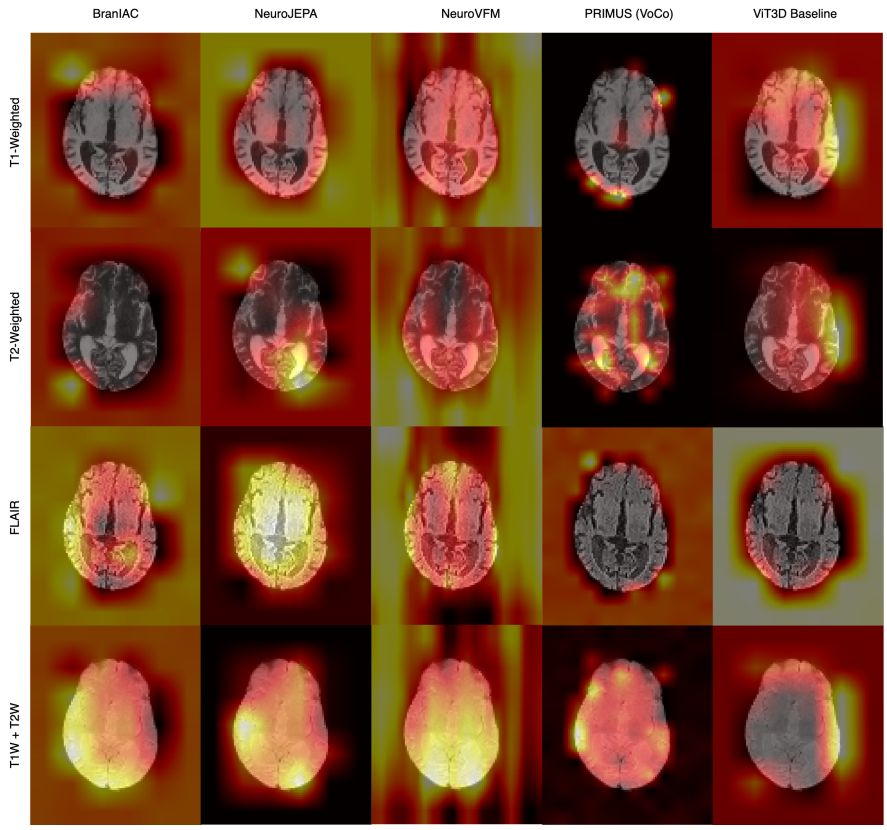

Fig. 1: Grad-CAM attention maps by modality (rows) and model (columns).

BrainIAC, NeuroVFM, and Primus predict almost exclusively the Dementia class, while ViT3D distributes correct predictions across all three diagonal cells, correctly classifying 11/22 Control, 26/31 Dementia, and 8/22 Parkinson's subjects, the only model to achieve non-zero recall on all three classes.

3.3 Explainability Analysis

To probe what the models are attending to, Figure 1 shows mean Grad-CAM attention maps per class. The lowest-performing frozen backbones (BrainIAC and NeuroVFM) produce diffuse, largely unstructured activation spanning the entire field of view rather than localizing to the brain. Primus attends to sparse, patchy regions concentrated at the brain periphery rather than deep gray-matter structures. Most notably, ViT3D exhibits a strong, spatially consistent hotspot at one lateral edge of the image across nearly every modality. A shortcut feature may support confident predictions without them being well grounded in pathology, tempering the otherwise favorable results of ViT3D in 1. Although, the attribution maps alone cannot establish causality.

Table 2: **Per-class F1 score**. Averaged over the four input tasks.

Model	Control	Dementia	Parkinson
BrainIAC	0.028	0.545	0.000
Neuro-JEPA	0.126	0.557	0.052
NeuroVFM	0.021	0.534	0.000
Primus	0.117	0.509	0.000
ViT3D	**0.412**	**0.629**	**0.128**

3.4 Statistical significance

Paired t-tests on 5-fold accuracy (ViT3D vs. each frozen backbone) reached significance only on the T1w task (BrainIAC $p = 0.044$, Neuro-JEPA $p = 0.041$, NeuroVFM $p = 0.021$); no other model/task comparison reached $p < 0.05$, consistent with the limited statistical power of a 5-fold protocol at this sample size ($n = 79$–80).

4 Discussion

Our findings suggest that reported generalization of brain MRI foundation models does not necessarily extend to small, heterogeneous and clinically distinct cohorts under a frozen-backbone protocol. This is consistent with prior work showing that foundation models do not necessarily yield superior fine-tuned performance compared with standard supervised models in low-data settings, and that direct benchmarking against simpler baselines remains essential [9]. In our benchmark, the foundation models did not exhibit uniformly transferable representations across downstream tasks, contrary to the broad transferability often reported in large-scale brain MRI foundation-model studies.

This pattern likely reflects a combined effect of both the model and the data. On the model side, using purely frozen features with a simple classification head may not allow sufficient adaptation to local disease phenotypes, scanner variability, or domain shift in this cohort. On the data side, the dataset is small, asymmetric, and non-representative, which increases fold-to-fold variability and weakens the reliability of fixed representations. These observations align with the broader literature, which shows that neuroimaging foundation models often benefit from large-scale training and carefully matched downstream adaptation. For example, BrainIAC was introduced as a self-supervised foundation model for brain MRI designed to support broad downstream use, but its success still depends on how well the downstream setting matches pretraining conditions [13]. Our results reinforce the view that frozen foundation models are not automatically sufficient in low-resource settings unless domain adaptation and dataset-specific considerations are taken into account.

Neuro-JEPA was the most interesting special case in our benchmark. It performed relatively well on FLAIR and showed improved behavior when T1 and

T2 were combined, suggesting that the learned representation may make more effective use of complementary sequence information in this cohort. This is especially plausible for FLAIR, which often highlights pathology-related differences more clearly than conventional T1 or T2 imaging. However, these gains were not uniform, indicating that even strong multimodal objectives do not fully overcome scanner shift, cohort heterogeneity, and limited sample size.

5 Conclusion

In this study, we evaluated four pretrained brain MRI foundation models under a frozen-backbone protocol on an African clinical cohort comprising dementia, Parkinson's disease, and control participants. Across T1w, T2w, FLAIR, and T1w+T2w inputs, the frozen foundation models generally failed to outperform a 3D ViT trained from scratch, with several models exhibiting uneven class-wise performance. Neuro-JEPA was a notable exception, achieving competitive AUC on multimodal and FLAIR inputs.

These findings challenges the prevailing assumptions that frozen foundation model backbones can universally adapt to low-data environments. The poor transferability observed is likely driven by severe domain shift: models pretrained predominantly on data from cohorts outside the target population struggle to interpret the heterogeneous acquisition protocols, lower-field scanner noise (0.3 T to 1.5 T), and distinct disease etiology profiles characteristic of African clinical settings. These findings underscore the need for neuroimaging foundation models to incorporate geographically and clinically diverse pretraining data.

6 Limitations

Our study has important limitations. The dataset is small and heterogeneous. The models were evaluated as frozen feature extractors with a simple classification head, which may underuse the pretrained representations. We also did not evaluate lightweight adaptation strategies such as LoRA fine-tuning on the target population, directly assess potential shortcut features contributing to confident predictions, or quantify the data and computational costs of calibrating these models for underrepresented populations. Future work should investigate parameter-efficient fine-tuning, shortcut-feature robustness, population-specific calibration, and richer multimodal fusion, particularly with T1+T2+FLAIR, to determine whether the observed performance limitations arise primarily from the data, the frozen setup, or both.

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

References

1. Adhikari, P.S., Sharma, A.: Dcgan-based data augmentation and gridsearch-optimized cnn for imbalanced brain tumor mri classification. In: International Con-

ference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026). pp. 159–175. Atlantis Press (2026)
2. Akinyemi, R.O.: Epidemiology of dementia in africa. Alzheimer's & Dementia **21**(Suppl 6), e105503 (2025). https://doi.org/10.1002/alz70860_105503
3. Caligiore, D., Giocondo, F., Silvetti, M.: The neurodegenerative elderly syndrome (nes) hypothesis: Alzheimer and parkinson are two faces of the same disease. ibro neurosci reports 13: 330–343 (2022)
4. Chandru, M., Abinesh, M., Ananth, M.S., Vashist, A., Manickam, P.: Artificial intelligence in neurodegenerative disease diagnosis: Advancing alzheimer's and parkinson's diseases. Current Opinion in Biomedical Engineering p. 100638 (2025)
5. Huang, H., Chen, L., Chen, J., Hyun, J., Loftus, J.R., Melmed, K., Orringer, D., Frontera, J., Dehkharghani, S., Masurkar, A., et al.: Learning sparse latent predictive foundation model for multimodal neuroimaging. arXiv preprint arXiv:2606.14957 (2026)
6. 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)
7. Kondepudi, A., Rao, A., Zhao, C., Lyu, Y., Harake, S., Banerjee, S., Joshi, R., Meissner, A.K., Hou, R., Jiang, C., Chowdury, A., Srinivasan, A., Athey, B., Gulani, V., Pandey, A., Lee, H., Hollon, T.: Health system learning achieves generalist neuroimaging models. arXiv preprint arXiv:2511.18640 (2025)
8. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
9. Nair, K., Razavian, N., Lui, Y.W.: Impact of dataset size on fine-tuning foundation models for neuroanatomic segmentation: Testing the foundation model hypothesis. Medical Physics **53**(2), e70337 (2026)
10. Ogbole, G.I., Adeyomoye, A.O., Badu-Peprah, A., Mensah, Y., Nzeh, D.A.: Survey of magnetic resonance imaging availability in west africa. The Pan African Medical Journal **30**, 240 (2018)
11. Oyeniran, O., Donald, K., Orunmuyi, A., Karikari, T.K., Evans, T.E., Cakmak, J., Draper, E., Ucheagwu, V.A., Hicks, J., Rosa-Neto, P., et al.: Imaging dementia in african populations: Closing the health equity gap. Alzheimer's & Dementia **21**(Suppl 2), e101439 (2025)
12. Ruffini, F., Mulero Ayllón, E., Shen, L., Soda, P., Guarrasi, V.: Benchmarking foundation models and parameter-efficient fine-tuning for prognosis prediction in medical imaging. Computer Methods and Programs in Biomedicine **275**, 109196 (2026). https://doi.org/10.1016/j.cmpb.2025.109196
13. Tak, D., Garomsa, B.A., Chaunzwa, T.L., Zapaishchykova, A., Climent Pardo, J.C., Ye, Z., Zielke, J., Ravipati, Y., Vajapeyam, S., Mahootiha, M., Smith, C., Familiar, A.M., Liu, K.X., Prabhu, S., Bandopadhayay, P., Nabavizadeh, A., Mueller, S., Aerts, H.J.W.L., Huang, R.Y., Poussaint, T.Y., Kann, B.H.: A generalizable foundation model for analysis of human brain mri. Nature Neuroscience (2026). https://doi.org/10.1038/s41593-026-02202-6
14. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., K"ohler, G., Maier-Hein, K.: Primus: Enforcing attention usage for 3d medical image segmentation. Transactions on Machine Learning Research (2026), arXiv:2503.01835

15. Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Kohler, G., Maier-Hein, K.: An openmind for 3d medical vision self-supervised learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23839–23879 (2025)
16. 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)
17. Wogu, E., Ogoh, G., Filima, P., Nsaanee, B., Caron, B., Pestilli, F., Eke, D.: Fair african brain data: challenges and opportunities. Frontiers in neuroinformatics **19**, 1530445 (2025)
18. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22873–22882. IEEE (2024)