Towards CPU-Deployable nnU-Net: Training-Free Sensitivity-Guided Compression Across Clinical Domains

Willem P.E. Boonzaier[1,2] [0009-0007-7216-6543], Izak S. Pretorius[1,2] [0009-0003-0200-0602], Toufiq Musah[2,3] [0009-0001-0838-5190], Lourens J. Strauss[1] [0000-0002-2607-7941], Udunna C. Anazodo[2,4,5] [0000-0001-8864-035X]

[1] Department of Medical Physics, University of the Free State, Bloemfontein, South Africa
[2] SPRINT AI Training for African Medical Imaging Knowledge Translation (SPARK) Program, Montreal, Canada
[3] Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[4] Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada
[5] The Montreal Neurological Institute-Hospital, Montreal, QC, Canada

boonzaierwpe@ufs.ac.za

Abstract. Deep learning segmentation models have demonstrated strong performance across medical imaging tasks, yet their practical deployment in resource-constrained clinical environments — particularly in low- and middle-income countries — remains limited by high computational demands, large model sizes, and GPU dependence. We present a training-free method for dataset-adaptive architectural compression of the nnU-Net v2 ResEncM pipeline, implemented by applying the Jacobian sensitivity analysis of the XTinyU-Net framework within the nnU-Net v2 pipeline at initialisation time. By identifying the onset of representational saturation across both channel width and residual block depth configurations, the method selects the smallest stable architecture for each dataset without any additional training. The approach was validated across three datasets of increasing clinical complexity: the ACDC cardiac segmentation dataset, the BraTS-Africa glioma dataset, and the BraTS-METS brain metastasis dataset. Across all three, the lightweight models reduced model size by 94.9–96.9% (796.9 MB to 25.1–40.7 MB) and CPU inference time by 82.7–90.1%, reducing BraTS-METS inference from over 11 minutes to under 2 minutes on CPU-only hardware. Segmentation accuracy was preserved or improved: the lightweight model matched or marginally exceeded the baseline on BraTS-Africa, outperformed it on BraTS-METS across all sub-regions (ET DSC: 0.534→0.602, NETC: 0.547→0.624, SNFH: 0.497→0.561), and produced fewer gross boundary failures on ACDC despite marginally lower mean DSC. These results demonstrate that training-free sensitivity-guided compression is a principled and dataset-adaptive pathway to clinically deployable segmentation, with no trade-off between performance and deployability.

Keywords: Light-weight, nnUNet, resource constrained, cardiac segmentation, Brain tumour segmentation.

1 Introduction

The precise segmentation of brain tumors is a critical step towards accurate diagnosis and effective treatment[1]. Datasets like the BraTS Africa dataset [2, 3] has been crucial in research addressing the challenges of glioma segmentation in Sub-Saharan Africa, where limited MRI infrastructure and heterogeneous acquisition protocols can cause domain shift, highlighting the need for domain-adaptable models[4]. While top-performing solutions in brain tumor segmentation challenges have effectively addressed the clinical domain gap and achieved high accuracy, many of these models are large, single or multi-architecture state-of-the-art models that require high-end GPUs for inference[5–10]. The computational demands of these deep learning models can hinder their practical deployment and long-term maintainability in hospital environments, specifically in scarce resource environments[11].

The absence of GPU-based computing systems and reliance on CPU capability in many departments within Low and Middle-Income Countries (LMICs) presents significant challenges for the deployment of artificial intelligence (AI) models in clinical settings[12, 13]. Many solutions presented as LMIC solutions bypass this by providing web-based solutions. However the reliance on web-based solutions to outsource computational load for AI models in LMICs is significantly hampered by issues of network stability and speed[12, 14]. Lightweight models, characterized by smaller memory footprints and faster inference times on CPUs, are essential for practical deployment and long-term maintainability in these environments[15].

2 Previous work

Recently a XTinyU-Net concept was proposed as a stable method for limiting model size without hindering performance. Rather than relying on computationally exhaustive training cycles to determine optimal model size, this approach employs a training-free selection mechanism that identifies dataset-specific, ultra-lightweight U-Net configurations directly at initialization. By calculating the Jacobian sensitivity of discrete, width-scaled U-Net variants, the method detects the critical transition point—or "representational capacity collapse"—where performance degrades as model capacity is reduced. This sensitivity-based metric allows for the automated selection of the smallest stable architecture, providing a robust, data-adaptive solution for resource-constrained environments.[16]. This model provides a giant step in the right direction for lightweight model design but has yet to be tested on State-Of-The-Art (SOTA) architectures like nnU-Net[17].

This work presents a practical implementation of the XTinyU-Net sensitivity-guided compression framework within the nnU-Net v2 pipeline. Adapting the original Jacobian-based selection methodology to two nnU-Net-specific architectural parameters — encoder channel width and residual encoder block depth — we embed training-free architecture selection directly into the nnU-Net v2 workflow, democratising access to state-of-the-art segmentation performance in resource-constrained clinical environments.

3 Methods

3.1 Data Description

This study utilized three distinct open-source MRI datasets, each representing a progressively complex clinical scenario. The ACDC cardiac dataset[18], (Fig 1a), served as an initial baseline, characterized by a smaller sample size, high contrast, and predictable anatomical structures. To assess performance under increased heterogeneity, the BraTS Africa Glioma dataset[2] (Fig 1b), was subsequently employed, providing high-quality, real-world data reflective of an African patient population. Finally, the BraTS METS dataset[19] (Fig 1c), served as an extreme test case; its significantly larger population and the inherent unpredictability associated with the clinical presentation of brain metastases offered a rigorous environment for stress testing the model's robustness and generalizability. While the ACDC dataset consists exclusively of short-axis Steady-State Free Precession (SSFP) sequence images, the BraTS datasets provide a more diverse multi-parametric approach, incorporating T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and T2-FLAIR sequences. All three datasets present 3D segmentation challenges requiring the delineation of multiple anatomical and pathological regions. For the ACDC dataset, the focus is on the left ventricle, myocardium, and right ventricle. In contrast, the BraTS datasets require the segmentation of the necrotic core (NETC), surrounding non-enhancing FLAIR hypersensitivity (SNFH), and enhancing tumor (ET) for pre-operative cases, with the additional requirement of identifying the resection cavity (RC) in post-operative patients.

3.2 Deep Learning Frameworks

The nnU-Netv2 pipeline was used as the architectural backbone throughout this experiment. Preprocessing, training and inference was conducted with the standard nnU-Netv2 ResEncM 3D full resolution configuration with training running for 500 Epochs and the checkpointed best model being used for inference. All models were trained on a single NVIDIA RTX 5000 (16GB VRAM) GPU.

3.3 Model Development

The Control/Baseline model used the standard nnU-Netv2 Residual Encoder (ResEncM) configuration. Experimental models were obtained by systematically down-scaling network width and the number of residual encoder blocks per stage, following the XTinyU-Net design principles. After the standard nnU-Netv2 preprocessing stage (which produces the dataset fingerprint and the initial plans file), we inserted a Jacobian-sensitivity analysis implemented in a Python Jupyter notebook that interfaces directly with the nnU-Netv2 planning and network-instantiation code. For a candidate architecture the notebook:

1. Instantiates the corresponding residual U-Net (identical residual-block structure, batch-norm, and residual connections as the ResEncM baseline).

2. Draws a small batch of preprocessed training patches.
3. Computes the Jacobian matrix of the network output with respect to the input (or an intermediate feature map, as defined in the XTinyU-Net methodology) via automatic differentiation.
4. Summarises the Jacobian into a scalar sensitivity score.

In the first phase the notebook enumerated a discrete set of width multipliers while keeping the original number of residual blocks per stage. The configuration with the lowest Jacobian sensitivity was retained. In the second phase the selected width was fixed and the number of residual encoder blocks per convolutional stage was varied. Again, the variant with the lowest Jacobian sensitivity was chosen.

The final selected pair was written back into the nnU-Netv2 plans file by editing the architecture parameters (UNet_base_num_features, n_conv_per_stage_encoder, etc.). The source code for this section of the project may be found at https://github.com/Williamwhy/Tiny-nnU-Net-Via-XTiny-methodology.

Since the architecture search is performed once after nnU-Netv2 preprocessing and before training. It evaluates a small discrete set of candidate width and residual-block configurations (5 width settings × 4 depth settings in our experiments) using a light-weight proxy network. For each candidate we compute the Jacobian sensitivity on 10 training mini-batches via automatic differentiation. On a single modern GPU the entire selection procedure finishes in a few minutes and is negligible compared with a full nnU-Net training run. The selected parameters are then written into the plans file; all subsequent training uses the official nnU-Netv2 pipeline with no extra overhead making this approach computationally cheap, dataset adaptable and predictable.

3.4 Model evaluation:

Each experimentally compressed model was compared the nnUNet standard model trained on the same dataset. Model checkpoint size was used as a method to estimate model size compression. Performance mertics were Dice and 95th percentile Hausdorff distance (HD95) for the ACDC and BraTS-Africa dataset determined from the validation populations of the respective datasets, while for the BraTS METS dataset, the lesion-wise Dice and lesion-wise HD95 were gained from the official challenge platform (Synapse) validation. Finally, inference time on Kaggle`s cloud computing GPU (NVIDIA Tesla T4) and CPU (all 4 CPUs) system was measured as a measure of adaptability to resource constrained settings. Statistical testing of results was performed using Wilcoxon ranked t-test with statistically significant findings represented as those with p-values smaller than 0.05.

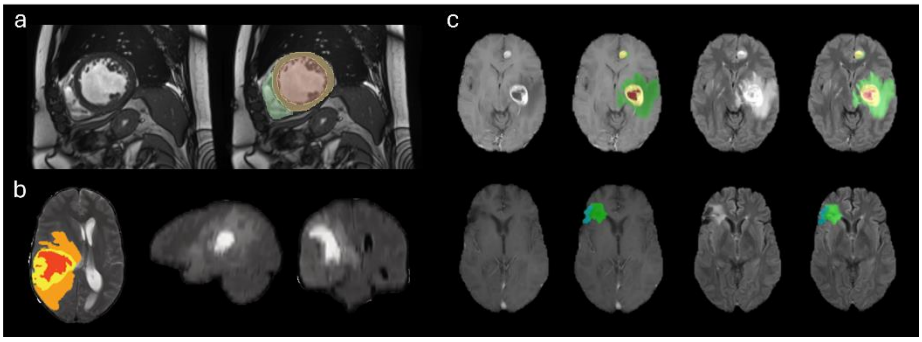

Fig. 1. Representative MRI examples from each dataset used in this study. (a) Short-axis Steady-State Free Precession cardiac MRI from the ACDC dataset showing a raw acquisition (left) and the corresponding segmentation overlay (right) with left ventricle (pink), myocardium (tan), and right ventricle (green) delineations. (b) Axial FLAIR MRI views from the BraTS-Africa glioma dataset showing a large pre-operative tumour with enhancing tumour (ET, yellow), necrotic core (NETC, red) and peritumoral edema (SNFH, orange) overlaid on the FLAIR as well as hyperintensity extent visible in the sagittal and coronal views. (c) Two representative BraTS-METS cases shown on T1 post-contrast and FLAIR sequences. Top row: a pre-operative case with NETC (red), ET (yellow), and SNFH (green) annotations. Bottom row: a post-operative case demonstrating the resection cavity (RC, cyan) as an additional sub-region unique to post-treatment presentations.

4 Results:

The Jacobian sensitivity analysis identified dataset-specific lightweight configurations for each of the three datasets, as shown in Fig. 2. For ACDC, the lowest sensitivity was achieved with channel widths of [4, 8, 16, 32, 64, 64] and [1, 1, 1, 1, 1, 1] residual encoder blocks per stage — the shallowest and narrowest configuration tested, reflecting the relatively low representational complexity of this dataset. For BraTS-Africa, the optimal configuration was [8, 16, 32, 64, 128, 128] channel widths with [1, 2, 2, 2, 2, 2] residual blocks per stage, requiring greater capacity to accommodate the multi-modality nature and the heterogeneity of glioma tumour anatomy of the dataset. For BraTS-METS, the selected configuration was [16, 32, 64, 128, 256, 256] channel widths with [1, 2, 2, 2, 2, 2] residual blocks per stage, reflecting the greater representational complexity and heterogeneity metastasis segmentation including the additional resection cavity sub-region in post-surgical cases. In all cases the selected depth was shallower than the nnU-Net ResEncM default of [1, 3, 4, 6, 6, 6] residual blocks per stage, and the selected channel widths were substantially narrower than the default [32, 64, 128, 256, 320, 320]. The baseline nnU-Net ResEncM checkpoint is consistent across all three datasets at 796.9 MB. The ACDC lightweight model achieved a checkpoint size of 25.1 MB (96.9% reduction), BraTS-Africa 39.9 MB (95.1% reduction), and BraTS-METS 40.7 MB (94.9% reduction).

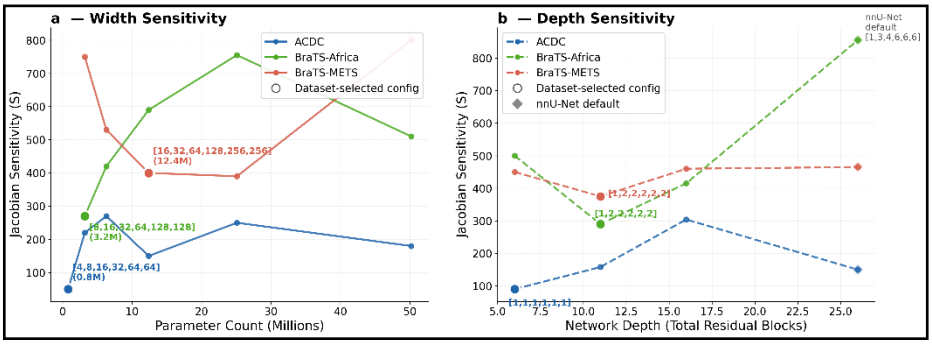

Fig. 2. Jacobian sensitivity analysis results for width and depth selection across all three datasets. (a) Sensitivity as a function of parameter count for five candidate channel width configurations. Each curve represents one dataset; the circled marker indicates the configuration selected at the onset of the low-sensitivity plateau. For ACDC the optimal width [4, 8, 16, 32, 64, 64] (0.8M parameters) was identified at far lower complexity than BraTS-Africa [8, 16, 32, 64, 128, 128] (3.2M) and BraTS-METS [16, 32, 64, 128, 256, 256] (12.4M). (b) Sensitivity as a function of total residual encoder blocks for four depth configurations, using the width selected in (a). ACDC converged to the shallowest configuration [1, 1, 1, 1, 1, 1] (6 blocks), while both BraTS datasets selected [1, 2, 2, 2, 2, 2] (11 blocks). All selected configurations are substantially shallower than the nnU-Net ResEncM default [1, 3, 4, 6, 6, 6] (26 blocks, diamond markers).

Regarding the segmentation accuracy, on the ACDC cardiac dataset, the lightweight model showed a marginal reduction in DSC across all three structures (RV: 0.909→0.898, Myocardium: 0.904→0.883, LV: 0.946→0.937). However, inspection of the HD95 distributions reveals a clinically important finding: while mean DSC was slightly lower, the lightweight model produced substantially fewer gross segmentation errors, with the extreme HD95 outliers present in the baseline model largely absent in the lightweight equivalent. This suggests that although the compressed model sacrifices a small degree of voxel-level overlap precision, it maintains a more reliable spatial understanding of cardiac anatomy and avoids the catastrophic boundary failures occasionally observed in the larger model (Fig 3a).

On the BraTS-Africa glioma dataset, the lightweight model performed comparably to or marginally above the baseline across ET, NETC, and SNFH in both DSC and HD95. This indicates that the Jacobian-guided compression preserved segmentation capacity fully for this dataset (Fig 3b).

On the BraTS-METS dataset, the lightweight model produced improvements in DSC and HD95 across ET, NETC, and SNFH relative to the baseline. Mean DSC increased from 0.534 to 0.602 for ET, 0.547 to 0.624 for NETC, and 0.497 to 0.561 for SNFH, suggesting that the compressed architecture generalises better on this heterogeneous multi-lesion dataset, likely due to the reduced tendency to overfit to dominant lesion phenotypes in the training set. For RC, the mean RC DSC increased from 0.293 to 0.338 in the lightweight model (Fig 3c).

Inference time reductions were consistent across all three datasets on both GPU and CPU hardware. On GPU, the lightweight models reduced inference time for ACDC by 83.1%, 71.8% for BraTS-Africa, and from 73.3% for BraTS-METS. CPU inference

time reductions were even more pronounced, with ACDC reduction of 90.1%, BraTS-Africa 82.7%, and BraTS-METS 88.7%. The CPU gains are particularly clinically significant — reducing BraTS-METS inference from over 11 minutes to under 2 minutes on CPU-only hardware (Table 2) directly addresses the deployment barrier in LMIC settings where GPU access is limited or absent. The consistency of these gains across datasets of markedly different complexity and modality confirms that the Jacobian-guided compression produces lightweight models that are not only smaller but substantially faster in both GPU-accelerated and resource-constrained CPU-only environments.

Table 2. Inference time and model size comparison between baseline nnU-Net and the proposed lightweight model across three datasets.

	Inference time (s)				Time reduction (%)	
Dataset	*GPU base*	*GPU light*	*CPU base*	*CPU light*	*GPU*	*CPU*
ACDC	0.7	0.1	690	69	83.1	90.1
BraTS-Africa	5.0	1.4	684	119	71.8	82.7
BraTS-METS	5.0	1.3	688	78	73.3	88.7

The qualitative performance derived from the ACDC dataset showed all differences to be statistically significant except for the HD95 of the right ventricle (p=0.572). For the BraTS-Africa dataset, Dice differences for all classes were statistically significant while none of the classes showed a statistically significant difference in HD95. Finally, all classes in the BraTS-MeTS dataset showed a statistically significant difference in HD95 and Dice, except for the resection cavity which had no statistically significant differences in Dice (p=0.420) or HD95 (p=0.528).

Discussion

The results of this study demonstrate that Jacobian sensitivity-guided compression within the nnU-Net framework produces lightweight models that match or exceed the segmentation accuracy of the standard ResEncM baseline across datasets of markedly different clinical complexity, modality, and structure — while reducing model size by approximately 95% and inference time by up to 90% on CPU hardware.

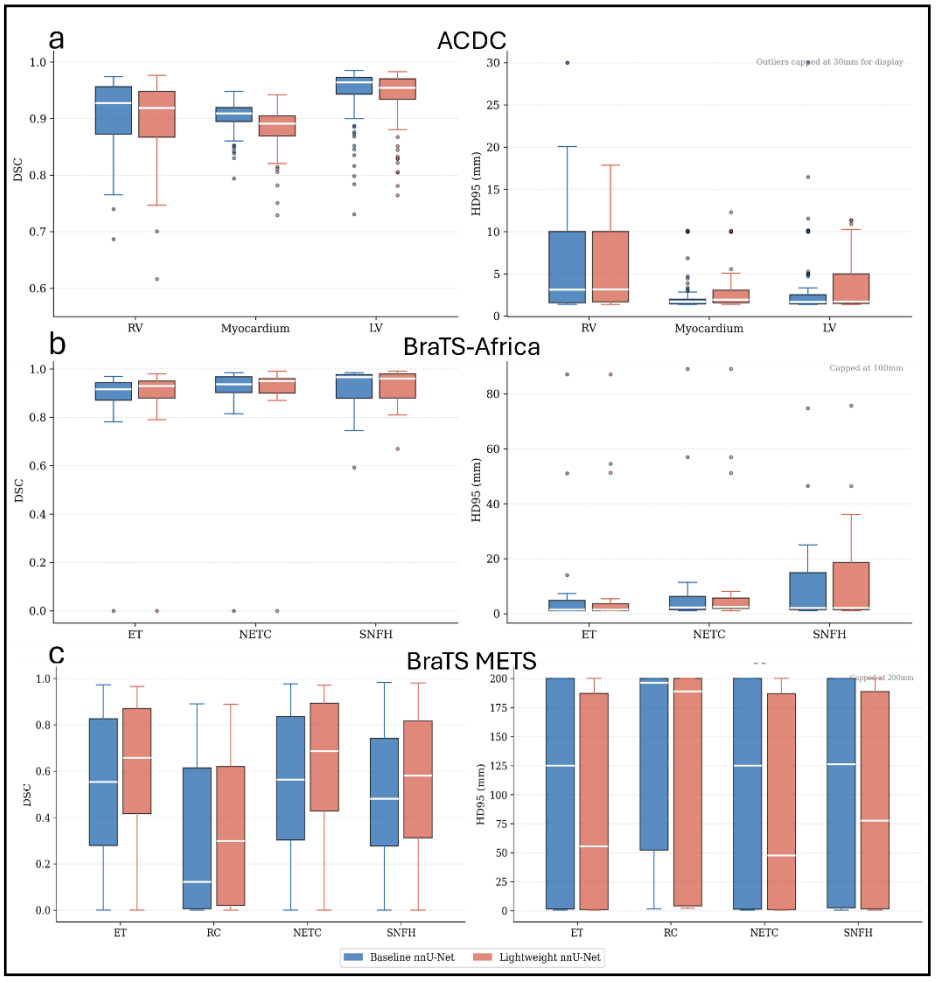

Fig. 3. Boxplots performance metrics for full nnUNet and lightweight nnUNet for all three datasets. a) DSC and HD95 for right ventricle (RV), Myocardium and left Ventricle (LV) for the ACDC dataset. b) DSC and HD95 for necrotic core (NETC), surrounding non-enhancing FLAIR hypersensitivity (SNFH), and enhancing tumor (ET) for the BraTS-Africa dataset. c) Lesion-wise DSC and HD95 for necrotic core (NETC), surrounding non-enhancing FLAIR hypersensitivity (SNFH), enhancing tumor (ET) and resection cavity (RC) for the BraTS METS dataset.

The finding that the lightweight model statistically significantly underperforms the baseline on ACDC DSC while simultaneously eliminating the gross boundary failures observed in the baseline HD95 distributions is an important nuance. Standard DSC is a voxel-level metric that rewards correct prediction of large structures disproportionately and is insensitive to spatial plausibility. The HD95 distribution tells a different story: the compressed model produces fewer catastrophic outliers, suggesting that reduced parameter count imposes an implicit regularisation effect that prevents the model from committing to implausible segmentation boundaries. In clinical cardiac imaging, a

model that is slightly less precise but consistently anatomically plausible is arguably more useful than one that achieves marginally higher mean DSC at the cost of occasional gross failures. This finding warrants further investigation into whether parameter reduction systematically reduces catastrophic failure modes across other segmentation tasks.

On BraTS-Africa, the lightweight and baseline models were indistinguishable in quality, confirming that the selected reduction fully captured the discriminative complexity of the glioma segmentation task. The fact that the BraTS-METS task required a wider configuration ([16, 32, 64, 128, 256, 256]) than the other datasets to achieve the same result, driven by the greater heterogeneity of multi-class metastasis presentations, further validates the dataset-adaptive nature of the approach.

The BraTS-METS improvements in ET, NETC, RC and SNFH DSC and HD95 represent clinically meaningful gains of 6–8 percentage points. This counter-intuitive result — that a substantially smaller model outperforms the larger baseline — is consistent with the overfitting hypothesis: the 101.5M parameter ResEncM baseline has excess capacity relative to the discriminative complexity of this dataset, and that excess capacity is exploited to memorise training-set idiosyncrasies rather than generalise to the validation distribution. These results together propose an interesting hypothesis, that if model compression is applied adaptively and in coherence with dataset and task complexity, the produced lightweight model might be a better generalizable and clinically useful model than an over parameterized model. Although statistical testing shows promise in many cases, it should be carefully considered as statistical gains may not equate to clinical gains as can be seen from comparing losses in Dice with gains in HD95 in the ACDC dataset.

5 Impact in Resource-Constrained Settings

The computational gains demonstrated in this study have direct and immediate implications for the deployment of AI-assisted segmentation in resource constrained clinical environments as all datasets used were constructed of multicentre data and of varying degrees of complexity and clinical environments. The primary barrier to AI adoption in these settings is not algorithmic performance but infrastructure: the absence of GPU-capable workstations, unreliable network connectivity that limits cloud-based solutions, and the maintenance burden of large model deployments in environments with limited technical support.

The lightweight models produced by the Jacobian-guided compression framework address each of these barriers directly. A reduction in CPU inference time of between 80 and 90% — transforms auto segmentation from a computationally prohibitive task requiring dedicated GPU hardware into one that can be executed on a standard clinical workstation CPU within routine reporting timeframes. These gains are achieved without web-based processing, eliminating the dependence on network stability and data transfer speeds that have been identified as critical barriers to cloud-based AI deployment in resource constrained settings.

The 94.9–96.9% reduction in model checkpoint size from 796.9 MB to 25.1–40.7 MB carries equally significant practical implications. Smaller model files reduce storage requirements, simplify distribution and version management, and lower the memory footprint during inference — all of which matter in environments where storage, RAM, and technical administration capacity are constrained. The ability to distribute a complete, functional segmentation model in under 41 MB means that updates and new model versions can be transferred over low-bandwidth connections, installed without specialist intervention, and run on hardware that would be unable to load the standard nnU-Net checkpoint into memory at all.

Taken together, these results support the argument that the XTiny U-Net paradigm, implemented here within the nnU-Net self-configuring pipeline, represents a viable and principled pathway toward clinically deployable AI segmentation in resource-constrained environments — one that does not require a trade-off between performance and deployability, but rather demonstrates that the two objectives are complementary when architectural design is guided by the discriminative complexity of the data rather than by the availability of high-end computational resources.

6 Limitations and future work

Although this work doesn't compare results with other lightweight solutions, it rather aims to compare the performance against state-of-the-art nnU-Net performance while demonstrating practical size and time reductions. As this method is a pruning only method for size reduction, it does not prohibit any down-the-line post training quantization techniques which might further reduce model size or computational cost. Future work will include testing this work beyond the use of MRI as input modality, exploring the performance on translation tasks in addition to segmentation tasks and including optimization for CPU based inference via model weight restructuration for CPU and quantization.

Acknowledgments. This work was part of the Sprint AI Training for African Medical Imaging Knowledge Translation (SPARK) Academy 2026 summer school on deep learning in medical imaging. The authors would like to thank the instructors of the summer for providing insightful background knowledge on brain tumours that informed the research presented here, most notably: Maruf Adewole, Mohannad Barakat, Craig Jones, Noha Magdy, Amal Saleh, Aondona Iorumbur, Bernes Atabonfack, Confidence Raymond, Craig Jones, Ilerioluwakiiye Abolade, Jeremiah Fadugba, Juampablo Heras Rivera, Kenneth Aguh, Lukman E. Ismaila, Maruf Adewole, Mohtady Barakat, Nourou Dine Bankole, Saikat Roy, Toufiq Musah, Willem Boonzaier.. The authors acknowledge the funding support provided to SPARK through the Lacuna Fund for Health and Equity (PI: Udunna Anazodo), the RSNA R & E Foundation (PI: Farouk Dako), the University of Washington Population Health Initiative Tier2 Grant (PI: Mehmet Kurt), McGill University Healthy Brain and Healthy Lives (HBHL, Anazodo), and the National Science and Engineering Research Council of Canada (NSERC) Discovery Launch Supplement (PI:UdunnaAnazodo, DGECR-2022-00136).

Disclosure of Interests. None.

Artificial Intelligence Declaration. Artificial intelligence tools were used for language editing, proofreading and graphical editing of the manuscript. All study methodology, experiments, data analysis, interpretation, and scientific conclusions were performed by the authors.

7 References

1. Zhang, W., Wu, Y., Yang, B., Hu, S., Wu, L., Dhelimd, S.: Overview of multi-modal brain tumor mr image segmentation. Healthc. 9, 1–20 (2021). https://doi.org/10.3390/healthcare9081051.
2. Adewole, M., Raymond, C., Omidiji, O., Ojo, N., Gbadamosi, Y., Cakmak, J., Fatade, A.: The BraTS-Africa Dataset : Expanding the Brain Tumor Segmentation Data to Capture African Populations. 1–6 (2025).
3. Adewole, M., Rudie, J.D., Gbdamosi, A., Raymond, C., Zhang, D., Omidiji, O.: The Brain Tumor Segmentation (BraTS) Challenge 2023 : Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). 1–10 (2023).
4. Abolade, I., Udo, A., Ojo, A., Oyetunji, A., Ajigbotosho, H., Iorumbur, A., Raymond, C., Adewole, M.: Domain-Adaptive Transformer for Data-Efficient Glioma Segmentation in Sub-Saharan MRI. 1–4 (2025).
5. Boonzaier, W.P.E., Moosa, F., Lebang, K., Jabaar, H., Iorumbur, A., Zhang, D., Raymond, C.: Domain Adaptation for Adult Glioma Segmentation in Sub-Saharan Africa: An Ensemble of nnU-Net v2 and MedNeXt. Lect. Notes Comput. Sci. 16376 LNCS, 248–261 (2026). https://doi.org/10.1007/978-3-032-16365-3_23.
6. Ankomah, C.T., Ayivor, L.E., Nyame, I., Wambo, L., Bonsu, P.Y., Iorumbur, A.M., Confidence, R., Musah, T.: How We Won BraTS-SSA 2025: Brain Tumor Segmentation in the Sub-Saharan African Population Using Segmentation-Aware Data Augmentation and Model Ensembling. (2025).
7. Barakat, M., Salah, O., Yasser, A., Ahmed, M., Arief, Z., Khan, W., Zhang, D., Iorumbur, A., Raymond, C., Barakat, M., Magdy, N.: Training Beyond Convergence: Grokking nnU-Net for Glioma Segmentation in Sub-Saharan MRI. Lect. Notes Comput. Sci. 16376 LNCS, 307–317 (2026). https://doi.org/10.1007/978-3-032-16365-3_28.
8. Parida, A., Capellán-martín, D., Jiang, Z., Tapp, A., Liu, X., Anwar, S.M., Dec, I. V: Adult Glioma Segmentation in Sub-Saharan Africa using Transfer Learning on Stratified Finetuning Data. 1–10.
9. Ferreira, A., Solak, N., Li, J., Dammann, P., Kleesiek, J., Alves, V., Egger, J.: How we won BraTS 2023 Adult Glioma challenge? Just faking it! Enhanced Synthetic Data Augmentation and Model Ensemble for brain tumour segmentation. (2024).
10. Amod, A.R., Smith, A., Joubert, P.: Bridging the Gap : Generalising State-of-the-Art U-Net Models to Sub-Saharan African Populations.
11. Lan, Q., Choi, A., Ma, J., Wang, B., Zhao, Z., Jiang, X., Hsu, Y., Mcwilliams, D.B., Informatics, B.: From Performance to Practice : Knowledge-Distilled Segmentator for On - Premises Clinical Workflows.
12. López, D.M., Rico-olarte, C., Blobel, B., Hullin, C.: Challenges and solutions for transforming health ecosystems in low- and middle-income countries through artificial intelligence. 1–22 (2022). https://doi.org/10.3389/fmed.2022.958097.
13. Ciecierski-holmes, T., Axt, M., Brenner, S.: Arti fi cial intelligence for strengthening

healthcare systems in low- and middle-income countries : a systematic scoping review. (2021). https://doi.org/10.1038/s41746-022-00700-y.

14. Zuhair, V., Babar, A., Ali, R., Oduoye, M.O., Noor, Z., Chris, K., Okon, I.I.: Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing. (2024). https://doi.org/10.1177/21501319241245847.
15. Vaze, S., Xie, W., Namburete, A.I.L.: Low-Memory CNNs Enabling Real-Time Mobile Deployment. IEEE J. Biomed. Heal. Informatics. 24, 1059–1069 (2020). https://doi.org/10.1109/JBHI.2019.2961264.
16. Kimbowa, A., Heidari, M., Liu, D., Hacihaliloglu, I.: XTinyU-Net: Training-Free U-Net Scaling via Initialization-Time Sensitivity. 1–11 (2026).
17. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods. 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z.
18. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohe, M.M., Pennec, X., Sermesant, M., Isensee, F., Jager, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Trans. Med. Imaging. 37, 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502.
19. Maleki, N., Amiruddin, R., Moawad, A.W., Yordanov, N., Gkampenis, A., Fehringer, P., Umeh, F., Chukwurah, C., Memon, F., Petrovic, B., Albalooshy, B., Cramer, J., Krycia, M., Shrickel, E.B., Ikuta, I., Thompson, G., Vidal, L., Kosovic, V., Goldman-Yassen, A.E., Hill, V., So, T.Y., Mhana, S., Alotaibi, A., Page, N., Bhatia, P., Guelen, M.S., Sharifi, Y., Jakovljevic, M., Abosabie, S.A.S., Abosabie, S.A., Ghonim, M., Ghonim, M., Manteghinejad, A., Baid, U., Janas, A., Krantchev, K., Ramakrishnan, D., Saluja, R., Jekel, L., Adewole, M., Albrecht, J., Anazodo, U., Aneja, S., Anwar, S.M., Bergquist, T., Chiang, V., Chung, V., Conte, G.M., Dako, F., Eddy, J., Ezhov, I., Khalili, N., Farahani, K., Iglesias, J.E., Jiang, Z., Johanson, E., Kazerooni, A.F., Kofler, F., Labella, D., Leemput, K. Van, Li, H.B., Linguraru, M.G., Liu, X., Meier, Z., Menze, B.H., Moy, H., Osenberg, K., Piraud, M., Reitman, Z., Shinohara, R.T., Wang, C., Wiestler, B., Wiggins, W., Shafique, U., Willms, K., Avesta, A., Bousabarah, K., Chakrabarty, S., Gennaro, N., Holler, W., Kaur, M., Lamontagne, P., Lin, M., Lost, J., Marcus, D.S., Maresca, R., Merkaj, S., Pedersen, G.C., Von Reppert, M., Sotiras, A., Teytelboym, O., Tillmans, N., Westerhoff, M., Youssef, A., Godfrey, D., Floyd, S., Rauschecker, A.M., Villanueva-Meyer, J., Pflüger, I., Cho, J., Bendszus, M., Brugnara, G., Guzman Perez-Carillo, G.J., Johnson, D.R., Kam, A., Yin, B., Kwan, M., Lai, L., Lall, N.U., Narayana Patro, S., Wu, L., Bansal, A., Barkhof, F., Besada, C., Chu, S., Druzgal, J., Dusoi, A., Farage, L., Feltrin, F., Fong, A., Fung, S.H., Gray, R.I., Iv, M., Postma, A.A., Mahajan, A., Joyner, D., Krumpelman, C., Letourneau-Guillon, L., Lincoln, C.M., Maros, E., Miller, E., Morón, F., Nimchinsky, E.A.: Analysis of the MICCAI Brain Tumor Segmentation - Metastases (BraTS-METS) 2025 Lighthouse Challenge: Brain Metastasis Segmentation on Pre- and Post-treatment MRI. 53, (2025).