Efficient Latent Space Operators for 3D Brain Tumor Segmentation

Cleveland Crossling, Hima Vadapalli[0000−0001−9040−3601], and Dustin van der Haar[0000−0002−5632−1220]

Academy of Computer Science and Software Engineering, University of Johannesburg, Cnr University Road and Kingsway Avenue, Auckland Park, Johannesburg, Gauteng, 2092
221096021@student.uj.ac.za, himav@uj.ac.za, dvanderhaar@uj.ac.za

Abstract. Automated brain tumor segmentation is highly useful for clinical diagnosis, yet state-of-the-art deep learning models are often too computationally demanding for deployment in resource-constrained settings, such as those in Sub-Saharan Africa. This paper explores the trade-off between segmentation accuracy and computational efficiency by integrating various latent space operators into a 3D V-Net architecture. A standard Transformer is compared to its efficient variants (Reformer, Linformer) and a recurrent model (LSTM) using the BraTS Sub-Saharan Africa 2024 dataset. The results demonstrate that efficient transformers can outperform their more complex counterparts on specific, clinically relevant metrics. The Linformer achieved the highest volumetric overlap with a Dice Score of 0.7916, while the Reformer yielded the most precise boundary delineation with a 95th percentile Hausdorff Distance of 18.6832.

Brain Tumor Segmentation, Medical Image Analysis, Deep Learning, Vision Transformer (ViT), CNN-Transformer, Latent Space

1 Introduction

Gliomas account for approximately 80% of primary intracranial tumours and remain the most malignant form of brain cancer [1]. While innovations in the Global North have reduced glioma mortality by 10–30%, Sub-Saharan Africa and other Low-Middle Income Countries (LMICs) face rising mortality rates of about 25% due to inadequate healthcare infrastructure, delayed diagnoses, and socioeconomic challenges [1, 2]. The accuracy of tumor segmentations directly impacts clinical decision-making, and therefore the quality of diagnosis. Automated systems, often driven by machine learning, have proven to be a viable approach to medical segmentation, and therefore must effectively segment key areas of interest [3].

Convolutional neural networks (CNNs) are highly effective for tasks that require spatial context awareness [4]. As a result, they have been widely adopted

to medical image segmentation, with architectures such as U-Net [5] and V-Net [6]. The introduction of situationally context aware paradigms, such as the transformer and other attention based systems, has led to the development of hybrid and multi-modal architectures in order to improve on purely convolutional models [7], often achieving state-of-the-art results [8].

While transformer-based models have significantly improved segmentation accuracy in medical imaging, they often come at the cost of computational efficiency. Transformers are inherently data-intensive and substantially increase model complexity due to their quadratic scaling in attention mechanisms and larger parameter counts [9]. This results in higher memory consumption and longer training times, factors that undermine deployment feasibility in resource-constrained settings, particularly in LMICs.

The primary contributions of this work are as follows: First, a systematic comparison of multiple sequence-processing paradigms is conducted, including a standard Transformer, efficient variants (Reformer and Linformer), and a recurrent model (LSTM). These act as operators within the latent space of a 3D convolutional V-Net. Secondly, these hybrid architectures are evaluated on the BraTS Sub-Saharan Africa dataset, providing insights relevant to resource-constrained clinical environments. Finally, a clear performance-efficiency trade-off is demonstrated, showing that computationally efficient models can achieve segmentation accuracy comparable, or even superior to, more complex architectures on specific, clinically relevant metrics.

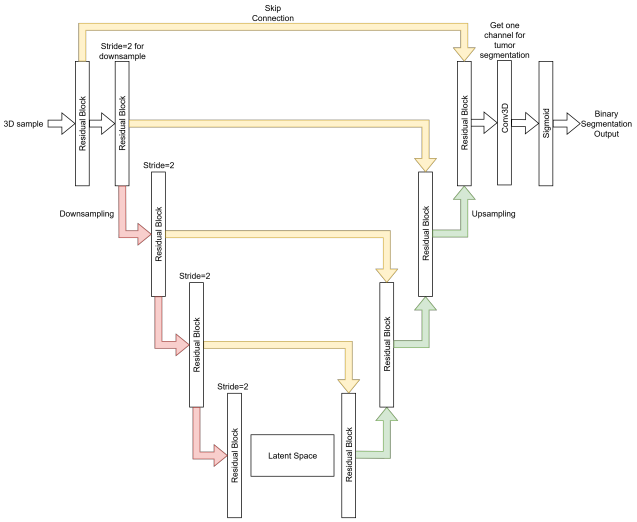
The remainder of this paper is organized as follows: Section 2 details the V-Net architecture and the specific latent space operators investigated. Section 3 describes the experimental setup, including the dataset, data preparation techniques, and evaluation metrics. Section 4 presents the quantitative and qualitative results of the experiments. Finally, Section 5 discusses the implications of the findings, and Section 6 concludes the paper.

2 Methodology

The technical setup for the experiment is outlined in this section in several parts: The baseline model, traditional transformer, efficient transformer comparison, and implementation details.

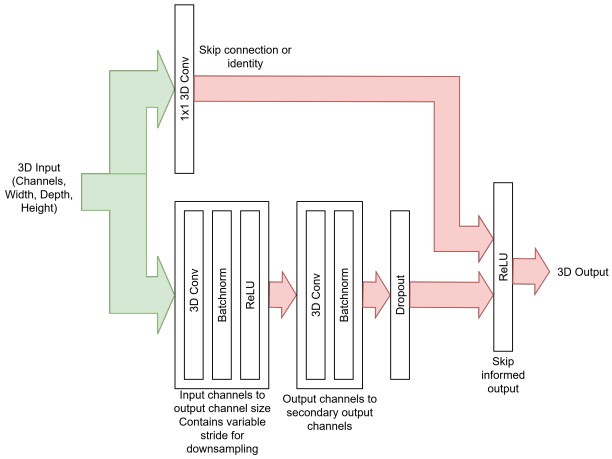
2.1 Base Model

The model utilised as a baseline for experimentation consistency is a V-Net, specifically selected for the 3 dimensional nature of the MRI modalities provided for brain tumor segmentation. The model consists of an encoder, responsible for downsampling the input, and a decoder for upsampling to output segmentation. Connections between the encoder and decoder at respective levels of compression are included as skip connections for informing upsampling, which conforms with a typical V-Net structure [6].



[H]

Fig. 1. The V-Net to be utilised as the base model. Red indicates downsampling, green indicates upsampling, and yellow illustrates the skip connections between the encoder and decoder.



[H]

Fig. 2. Diagram of the flow of the utilised Residual Block within the base V-Net architecture.

Figure 1 clearly illustrates the encoder and decoder structure of the V-Net, as well as the resulting latent space.

As illustrated in Figure 2, the network employs residual blocks for downsampling within the encoder. To achieve this, the first 3D convolutional layer within the block is configured with a stride of 2, effectively halving the spatial

dimensions of the input feature map. In parallel, a skip connection processes the original input through a 1x1 convolution, also with a stride of 2, to match the new dimensions and channel depth of the main path. The output of the main convolutional path, which includes a dropout layer for regularization, is then combined with the skip connection's output via element-wise summation. This final sum is passed through a ReLU activation layer.

In the decoder, the same residual block architecture is adapted for feature refinement during upsampling. This block processes an input that has already been enlarged via trilinear upsampling and concatenated with the corresponding feature map from the encoder's skip connection. For this role, the block's convolutional layers are set with a stride of 1 to preserve the input's dimensions. Its purpose is not to upscale, but to merge the features from the lower layer and the skip connection. The internal residual connection allows the block to refine the concatenated features before passing them to the next stage of the decoder.

2.2 Latent Space Operators

The base model receives a latent operator as a parameter that utilises the encoded values in the latent space as input. It then performs its operation and provides its nuanced output to the decoder. This is where the latent space transformers and algorithms are provided to the model before training.

To prepare the 3D feature maps for the attention mechanism, the spatial dimensions (depth, height, and width) are flattened into a single sequence of vectors. Each vector in this sequence retains the original channel depth as its feature dimension. This sequence is then passed to the latent operator, which processes the entire latent representation before its output is restructured back to its original 3D volumetric shape for the decoder [10].

The latent space operators utilised are as follows:

Baseline Transformer The canonical Transformer architecture, with its multi-head self-attention mechanism, is implemented as the performance benchmark in the experiments. This model is expected to represent the upper bound of segmentation efficacy, providing a reference against which all other latent space operators are evaluated [9].

Reformer To address the computational demands of the standard Transformer, the Reformer model is evaluated. This architecture introduces two key optimizations: locality-sensitive hashing (LSH) to approximate full attention, and reversible residual layers to reduce memory consumption. It is selected to test the hypothesis that a more efficient attention mechanism can maintain high performance with significantly lower computational overhead [11].

Linformer The Linformer is included as an example of a model that achieves linear complexity. It projects the self-attention mechanism's key and value matrices into a lower-dimensional space, drastically reducing computational load.

While this linear projection is highly efficient, its performance can be sensitive to the input sequence length [12, 13].

LSTM As a counterpoint to attention-based models, a Long Short-Term Memory (LSTM) network is included. LSTMs are a type of recurrent neural network (RNN) that processes sequence data iteratively using a series of gates to manage its internal state. This model is included to evaluate how a classic recurrent approach, which relies on sequential processing rather than global attention, performs on this volumetric segmentation task.

None To establish a performance floor, the base encoder-decoder architecture is used without any specialized operator in the latent space. The features from the encoder bottleneck are passed directly to the decoder. This baseline serves as the lower-bound benchmark, allowing for a direct measurement of the performance gain contributed by each of the investigated latent space models.

2.3 Operator Complexity

This section is intended to highlight the complexity of the selected latent space operators.

Table 1. Computational Complexity of Latent Space Operators

Latent Space Operator	**Time Complexity**
Traditional Transformer	$O(N^2 \cdot d)$
Reformer	$O(N \log N \cdot d)$
Linformer	$O(N \cdot d)$
LSTM	$O(N \cdot d^2)$
None	N/A

Table 1 indicates the computational complexity of each latent space operator, where N represents the sequence length of the flattened latent space and d is the feature dimension of each vector in the sequence. As can be seen, each model presents a different computational profile, creating a distinct trade-off between efficiency and effectiveness.

2.4 Configuration

This section describes the various hyperparameters and hardware used to train the models during experimentation.

Loss Function Dice loss is used, as it is the typical loss function for V-Net segmentation models [6]. Preliminary testing included both Tversky loss and a combination of Binary Cross Entropy and Dice losses as considerations, however, they lack the stability of stand-alone Dice loss in this context.

Optimizer The choice of optimizer was determined through preliminary experiments comparing Adam and AdamW [14]. The AdamW optimizer, which implements decoupled weight decay, demonstrated superior performance and stability during initial training runs. This is consistent with training transformer based models which benefit from a more nuanced form of regularisation than L2 [15]. Consequently, AdamW was selected for all subsequent experiments, configured with a learning rate of 1×10^{-4} and a weight decay value of 1×10^{-5}.

Hardware In order to emphasise training simplicity, each model was trained on a RTX3060 with 12GB of VRAM.

Other A batch size of 2 over 100 epochs was consistently used for all experiments.

3 Experiment Setup

The dataset, data preparation techniques, and metrics are discussed in this section.

3.1 Dataset

Experiments were conducted using the publicly available BraTS Sub-Saharan Africa 2024 dataset [16]. The official training set contains 60 subjects. Each subject includes four MRI modalities: T1-Weighted, Post-contrast T1-weighted, T2-weighted, and T2-weighted Fluid-Attenuated Inversion Recovery (FLAIR). Ground truth data is provided as a multi-class semantic segmentation map identifying distinct tumor sub-regions. The official validation dataset, which is provided without ground truth labels for blind evaluation in the challenge, was not used in the experiments.

Data Preparation To facilitate model training and validation, the official training dataset of 60 subjects was partitioned into a new training set (48 subjects, 80%) and a validation set (12 subjects, 20%) using a random split.

To reduce computational requirements, all four MRI volumes for each subject were downsampled by a factor of 2 in each dimension using a 2×2×2 max-pooling operation. These downsampled volumes were then stacked along the channel dimension to form a single 4-channel input tensor. The corresponding ground truth segmentation maps were similarly downsampled. For this study, the original multi-class labels were merged into a single binary mask to define the whole tumor region.

3.2 Metrics

To evaluate the performance of each model, two standard metrics are employed: the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95).

1. **Dice Score** measures the volumetric overlap between the predicted segmentation and the ground truth, and is considered a standard for medical image segmentation tasks [6].
2. **Hausdorff Distance (95th percentile)** measures the distance between the boundaries of the predicted and ground truth segmentations. It is sensitive to segmentation accuracy and is effective at penalizing predictions that deviate significantly in shape from the ground truth. By using the 95th percentile, we ensure the metric is not affected by a small number of extreme outliers, providing a stable assessment of large boundary errors [17].

Hausdorff distance is computationally demanding to run during training, thus Dice on validation is selected for model checkpointing.

4 Results

In this section, the results of each experiment, as well as their respective metrics, are provided.

4.1 Baseline Model

The results for the selection of the baseline model are as follows:

Table 2. Dice Scores for Various Ablation Settings

Latent Space Size	Loss Function	Best Dice Score
32	Dice	0.4842
64	Dice	0.4298
128	Dice	0.6802
256	Tversky	0.7335
256	Dice	0.7651
512	Dice	0.7460

In Table 2, latent space size indicates how many feature channels are at the bottleneck of the V-Net. A larger number of channels indicates a more memory intensive model.

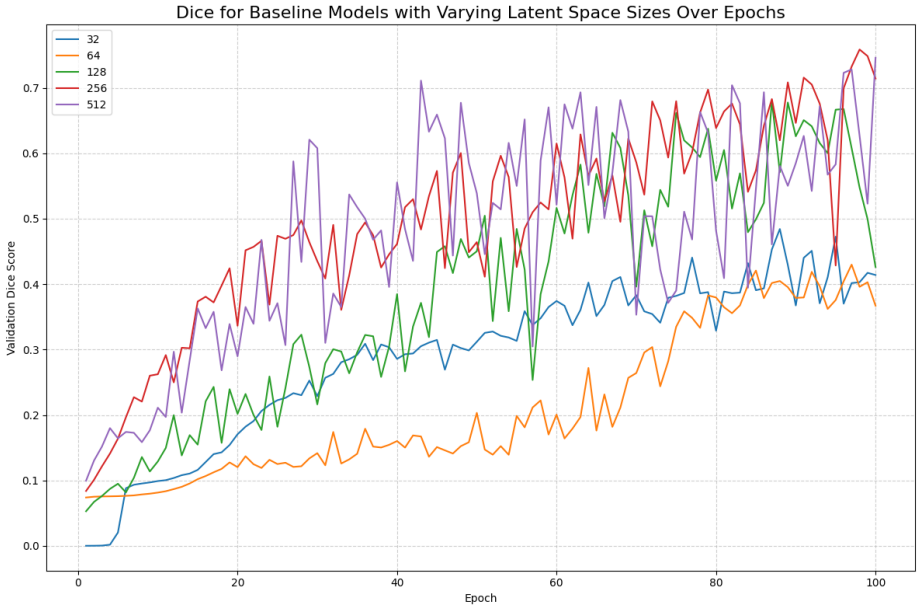


Fig. 3. The baseline model validation Dice Scores over epochs for different latent space sizes.

Figure 3 shows each latent space size during training and its Dice Score performance as a result.

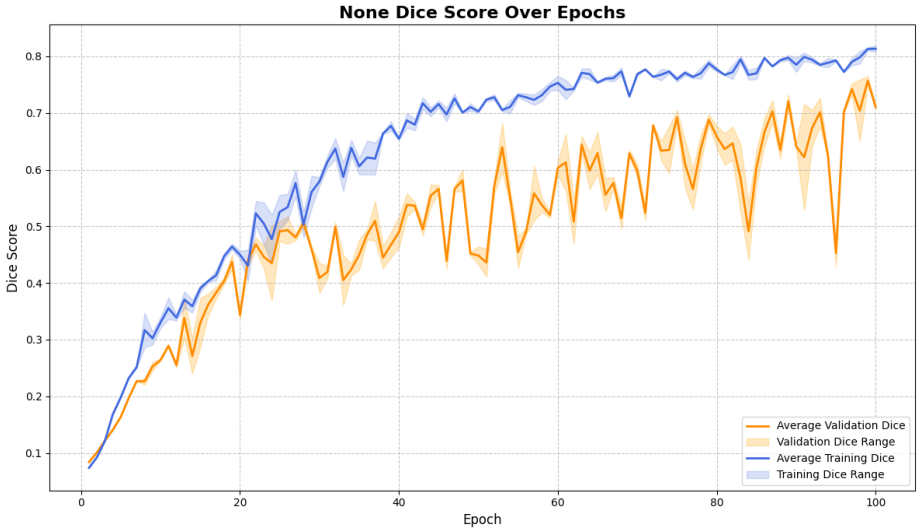


Fig. 4. Selected baseline model validation Dice Scores over epochs. The range indicates the variance between different training instances.

These results prompted 256 channels with Dice loss at the latent space to be selected for the baseline model.

4.2 Dice Metrics

The Dice results of the experiments with latent space operators.

Table 3. Dice Score Ranges for Different Latent Space Operators

Latent Space Operator	Dice Range
Transformer	0.7576 – 0.7578
Reformer	0.7304 – 0.7461
Linformer	0.7773 – 0.7916
LSTM	0.6743 – 0.6856
None	0.7586 – 0.7651

4.3 Hausdorff Distance

The Hausdorff distance at the 95th percentile for each experiment.

Table 4. Hausdorff Distance 95 (HD95) for Different Latent Space Operators

Latent Space Operator	HD95
Transformer	20.0814
Linformer	22.0027
Reformer	18.6832
LSTM	22.6156
None	25.2894

5 Discussion

The experimental results present a clear trade-off between model complexity and segmentation performance. The primary finding of this study is that efficient Transformer variants, specifically Linformer and Reformer, not only match, but can selectively outperform the standard Transformer in a 3D medical segmentation task. This challenges the notion that maximum computational complexity is a prerequisite for maximum performance.

A key observation is the divergence in performance between the Dice Score and Hausdorff Distance metrics. The Linformer model achieved the highest Dice Score (0.7916). This suggests its linear attention mechanism is effective at capturing the core structure of the tumor region. However, its corresponding Hausdorff Distance was average (22.0027), indicating a lower accuracy at the tumor boundaries.

The Reformer achieved the best Hausdorff Distance (18.6832), demonstrating a more precise segmentation of the tumor's edge. This suggests its locality-sensitive hashing (LSH) attention is useful for refining boundary details.

The standard V-Net with no latent operator achieved a Dice Score (0.7651) nearly identical to the full Transformer (0.7578). This suggests that for this architecture, the quadratic complexity of the vanilla attention mechanism offers no immediate benefit for volume overlap. However, the baseline model produced the worst Hausdorff distance by a significant margin (25.2894), confirming that the inclusion of a latent space operator is crucial for refining segmentation boundaries.

Finally, the LSTM model's performance was the lowest across both metrics. This aligns with the expectation that its sequential processing is less suited for capturing the complex, non-local spatial relationships in 3D volumetric data.

5.1 Limitations

A potential limitation of the study is the use of max-pooling for downsampling the input volumes and segmentation masks. While this method is computationally efficient and emphasizes intense features, it may result in the loss of finer textural details and can produce block-like artifacts at segmentation boundaries compared to interpolation-based methods. Future work should explore the impact of using trilinear and nearest-neighbour interpolation for the volumes and

masks respectively, as this could potentially lead to more precise boundary delineation and improved overall performance.

Furthermore, the use of a final post processing operator, often seen in state-of-the-art segmentation models, could be highly effective in reducing the high Hausdorff metrics shown in Table 4.

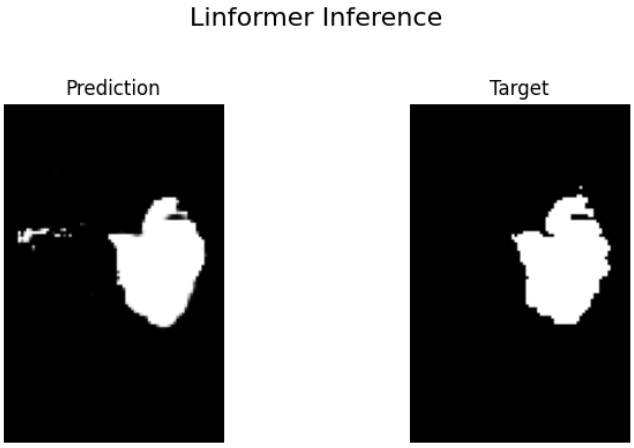


Fig. 5. Visualisation of a slice of inference output from the validation set for the best Linformer model compared to the target.

The Figure 5 further reinforces this point. There are several examples of similar small inference errors such as this, particularly for the Linformer and Reformer based models.

6 Conclusion

This study evaluated the impact of different latent space operators on the performance and efficiency of a V-Net for 3D brain tumor segmentation. The findings reveal that the most computationally complex model, the standard Transformer, does not guarantee the best performance. Instead, = a clear and practical trade-off is demonstrated: the Linformer performs well at volumetric overlap, achieving the highest Dice Score, while the Reformer provides superior boundary delineation, evidenced by the lowest Hausdorff Distance. Both of these efficient models outperformed the standard Transformer on their respective metrics. Furthermore, the results confirm that a purely convolutional baseline model struggles with boundary precision and a sequential LSTM-based approach is not suited for capturing the complex spatial context of this task. Ultimately, this work establishes that efficient Transformer variants are an effective choice for developing automated segmentation systems.

References

1. G. . Brain and O. C. C. Collaborators, "Global, regional, and national burden of brain and other cns cancer, 1990-2016: a systematic analysis for the global burden of disease study 2016," *Lancet Neurol*, vol. 18, no. 4, pp. 376–393, 2019.
2. S. Z. Y. Ooi, R. de Koning, A. Egiz, D. U. Dalle, M. Denou, M. R. D. Tsopmene, M. Khan, R. Takoukam, J. Kotecha, D. Sichimba, Y. C. H. Dokponou, U. S. Kanmounye, and N. D. A. Bankole, "Management and outcomes of low-grade gliomas in africa: A scoping review," *Annals of Medicine & Surgery*, vol. 74, February 2022.
3. Y. Gao *et al.*, "Medical image segmentation: A comprehensive review of deep learning-based methods," *Tomography (Ann Arbor, Mich.)*, vol. 11, no. 5, p. 52, Apr. 2025.
4. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, 2015, pp. 234–241.
6. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.
7. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, and A. L. Yuille, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
8. S. Wu, Z. Chen, Y. Xia, L. Yang, D. Wang, A. L. Yuille, and Y. Luo, "Mednext: A convnext-inspired framework for efficient 3d medical image analysis," *arXiv preprint arXiv:2411.15872*, 2024.
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762
10. A. Hatamizadeh, V. Nath, Y. Tang, J. Kautz, and D. Yang, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.
11. N. Kitaev, Łukasz Kaiser, and A. Levskaya, "Reformer: The efficient transformer," 2020. [Online]. Available: https://arxiv.org/abs/2001.04451
12. Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *arXiv preprint arXiv:2009.06732*, 2020.
13. S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020. [Online]. Available: https://arxiv.org/abs/2006.04768
14. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.
15. Z. Li, T. Cai, Y.-G. Kim, Y. Li, S. Zhou, and J. D. Lee, "Why transformers need adam: A hessian perspective," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://arxiv.org/abs/2405.14856
16. "Brats 2024," 2024, accessed: 2024/08/28. [Online]. Available: https://www.synapse.org/Synapse:syn53708249/wiki/627501
17. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.