RESEARCH PAPER

Lightweight CNN–Mamba 3D U-Net for Brain Tumor Segmentation and Quantitative Volume Assessment on an Edge Platform

D. Ramesh Reddy[1], Vijaya Rao Yamini Chandrika[2], MP. Sricharan.Reddy[3]
[1]Assistant Professor, Dept. of ECE, VNR VJIET, Hyderabad, India. Email: Rameshreddy_d@vnrvjiet.in
[2]Dept. of ECE, VNR VJIET, Hyderabad, India. Email: vychandrika99@gmail.com
[3]Dept. of ECE, VNR VJIET, Hyderabad, India. Email: mp.sricharan.reddy@gmail.com
***Corresponding author: Vijaya Rao Yamini Chandrika, Dept. of ECE, VNR VJIET, Hyderabad, India.**

Abstract— Accurate 3D brain tumor segmentation is essential for treatment planning and quantitative assessment, while their high computational and memory requirements limit practical deployment on resource-constrained edge platforms. This study proposes a hardware-aware lightweight CNN–Mamba 3D U-Net for multimodal MRI segmentation on the NVIDIA Jetson Orin Nano. The architecture combines depthwise-separable 3D convolutions for local feature extraction with low-resolution state-space modeling for long-range contextual representation, while maintaining ONNX and TensorRT compatibility. Evaluated on 32 held-out BraTS 2019 test cases, the proposed model achieved Dice scores of 0.8621, 0.8335, and 0.6695 for whole tumor, tumor core, and enhancing tumor, respectively, using only 2.24 million parameters compared with 22.58 million for the 3D U-Net baseline. INT8 TensorRT inference on the Jetson Orin Nano achieved 75.83 ms patch-level latency and 13.17 inferences/s at approximately 15W power consumption. Tumor-volume estimation showed strong agreement with ground truth, with Pearson coefficients of 0.9699, 0.9510, and 0.9195 for whole tumor, tumor core, and enhancing tumor, respectively, demonstrating that accurate quantitative brain tumor analysis is achievable on a 15W embedded GPU without server infrastructure.

Keywords— Brain tumor segmentation, CNN–Mamba, 3D U-Net, BraTS 2019, state-space model, NVIDIA Jetson Orin Nano, TensorRT, edge AI.

How to cite this article: *D. Ramesh Reddy1 Vijaya Rao Yamini Chandrika2 MP. Sricharan.Reddy3. Lightweight CNN–Mamba 3D U-Net for Brain Tumor Segmentation and Quantitative Volume Assessment on an Edge Platform. Int J Drug Deliv Technol. 2026;16(78s):1234-1236. DOI: 10.25258/ijddt.16.78s.133.*

I. INTRODUCTION

Brain tumor represents a significant clinical challenge due to their heterogeneous appearance, irregular margins, and varying location which can hamper diagnosis and treatment planning. Because of MRI's versatility, it has become the most used imaging modality for brain tumours, which in most protocols include T1, T1ce, T2 and FLAIR sequences, among many others, that yield complementary information.

The delineation of clinically relevant regions like whole tumor (WT), tumor core (TC), and enhancing tumor (ET) can prove helpful in treatment planning, disease monitoring, and quantitative assessment. Nonetheless, manual segmentation of volumetric MRI is a time-consuming, observer-dependent, and difficult-to-scale process. Consequently, there is a growing interest in automated deep-learning-based segmentation systems.

The hierarchical spatial feature extraction capability of convolutional neural networks (CNNs) has allowed them to obtain good performance in medical image segmentation. However, the local receptive field of convolution may limit the modelling of long-range dependencies in complex 3D tumor structures. Thus, volumetric medical image segmentation has been explored using transformers. The CKD-TransBTS architecture is a combination of convolutional feature extraction, transformer-based global modeling and modality-correlated cross-attention for multimodal brain tumor segmentation [1]. In a similar vein, 3D Transformer is used in nnFormer for capturing contextual relationships of object. MedNeXt shows that specially designed convolutional architectures can produce competitive performance in medical segmentation while enhancing architectural efficiency [2]. Due to these developments, Transformer networks can incur high computational and memory cost for 3D volumes of high resolution. State-space models have recently surfaced as an efficient alternative for long-range sequence modelling. Mamba's state-space modeling is selective with computational complexity scaling linearly with sequence length [4]. According to [5], Vision Mamba then adaptively modified state-space modeling to visual representation learning.

Owing to above advancements, there has been great interest in segmentation of medical images using Mamba based architectures. Mamba was applied in SegMamba for long-range sequential modeling in 3D medical image segmentation [6], while Swin-UMamba incorporated Mamba into a U-Net-style biomedical segmentation framework [7]. Zhou et al. demonstrated the infusion processes of convolutional local-feature extraction and state-space-based contextual modeling through the design of a Mamba-structured U-Net with residual multiscale convolution for brain tumor segmentation [8].

Despite Mamba-based segmentation methods showing promising results, deploying 3D medical imaging networks on resource-limited edge hardware remains a challenge in practice. It requires a lot of memory bandwidth, computation and activation storage. More specifically, a deployment platform like the NVIDIA Jetson Orin Nano comes with even more limitations related to unified memory, supported operators, inference precision, TensorRT support, etc. For edge deployment, a model's architecture must take into account hardware constraints rather than simply applying compression post-training. To tackle this issue, this research presents a hardware-aware lightweight CNN – Mamba 3D U-Net intended for the multimodal brain tumor segmentation and quantitative

assessment on NVIDIA Jetson Orin Nano. The depthwise-separable 3D convolution model efficiently extracts local features while the low-resolution Mamba-inspired state-space branch captures long-range context efficiently and at low computational cost. The proposed network achieves competitive segmentation performance on BraTS 2019 while using approximately 2.24 million parameters compared to approximately 22.58 million for the conventional 3D U-Net baseline.

The main contributions of this work are as follows:
(1) A hardware-aware lightweight CNN–Mamba architecture designed around the memory and computation constraints of the Jetson Orin Nano.
(2) An export-safe Mamba-inspired state-space formulation compatible with ONNX and TensorRT deployment without custom selective-scan plugins.
(3) Comprehensive edge evaluation using FP32, FP16, and INT8 TensorRT inference with latency, throughput, memory, GPU utilization, temperature, and energy measurements.
(4) Quantitative tumor-volume assessment for WT, TC, and ET using predicted segmentation masks, extending the framework beyond segmentation accuracy toward quantitative brain tumor analysis.

II. LITERATURE SURVEY

The study of computational efficiency and accuracy of 3D Medical Image Segmentation is significant in recent years. Dang et al. proposed LoG-VMamba that allows local feature extraction integrated with modeling global state-space for medical image segmentation [9]. Luo et al. designed a Multi-perspective Feature Extraction and Dense Attention Network (MPEDA-Net) for brain tumor segmentation. Zhang and colleagues suggested the ETUNet model, which combines efficient Transformer-based contextual modeling with U-Net for 3D brain tumor segmentation [10]. Li et al. proposed mResU-Net, a multimodal MRI-based brain tumor segmentation based on multiscale residual learning. Duman et al. proposed a clinically flexible and computation efficient method for tumor segmentation called RFS+
The efficient convolutional and Mamba architectures are the focus of recent techniques. Rahman et al. suggested the multi-receptive dilated convolutions-based EfficientMedNeXt. Xing et al. introduced SegMamba-V2 to address the long-range sequential modeling task for general 3d medical image segmentation [15]. The viability of using Mamba-based segmentation techniques for on-device medical image was proven by SwiM-UNet. Similarly, [17] proposed CIHM, a hybrid network structure of Mamba, which uses contextual and local feature modeling to achieve efficient segmentation.
Altogether, the studies have demonstrated the feasibility of lightweight CNN, Transformer and Mamba models. However, only limited research has been conducted that concurrently tackles the problem of 3D brain tumour segmentation that is hardware-aware, efficient deployment on resource-limited edge devices compatible with TensorRT, and quantitative assessment of tumour volumes. This divide is fueling the light-weight CNN – Mamba framework for NVIDIA's Jetson Orin Nano combining contextual and local feature modeling for efficient segmentation [17].

Table 1. Summary of Recent Related Work

Ref.	Method / Year	Application	Main Contribution	Limitation / Research Gap
[9]	LoG-VMamba, 2024	Medical image segmentation	Combines local feature extraction with global Mamba-based contextual modeling	Does not target Jetson-based hardware deployment
[10]	MPEDA-Net, 2024	Brain tumor segmentation	Lightweight network using multi-perspective feature extraction and dense attention	No Mamba-based long-range modeling or edge deployment
[11]	ETUNet, 2024	3D brain tumor segmentation	Efficient U-Net enhanced with Transformer-based global feature modeling	Transformer processing remains computationally demanding for edge devices
[12]	mResU-Net, 2024	Multimodal MRI brain tumor segmentation	Uses multiscale residual learning for effective tumor feature extraction	Hardware-aware deployment not investigated

[13]	RFS+, 2023	Brain tumor segmentation	Focuses on clinically adaptable and computationally efficient segmentation	Does not address Mamba or TensorRT-based embedded deployment
[14]	EfficientMedNeXt, 2025	Medical image segmentation	Uses multi-receptive dilated convolutions for efficient multiscale feature extraction	Does not specifically address brain-tumor edge deployment
[15]	SegMamba-V2, 2026	3D medical image segmentation	Employs Mamba-based long-range sequential modeling for volumetric segmentation	Embedded GPU and TensorRT deployment not the primary focus
[16]	SwiM-UNet, 2026	On-device medical image segmentation	Lightweight architecture designed for efficient on-device segmentation	Does not specifically evaluate Jetson Orin Nano-based brain tumor segmentation
[17]	CIHM, 2026	Medical image segmentation	Hybrid Mamba framework combining local and contextual information efficiently	Hardware-specific TensorRT validation is not the primary contribution
Proposed	**Lightweight CNN–Mamba 3D U-Net**	**3D brain tumor segmentation and volume assessment**	**Combines DSConv and low-resolution state-space modeling with hardware-aware TensorRT-compatible design**	**Designed and experimentally evaluated for NVIDIA Jetson Orin Nano deployment**

III. METHODOLOGY

A. Dataset Description

This study used the BraTS 2019 training dataset, which includes 335 multimodal MRI cases. Among them, 259 cases are high grade glioma and 76 cases are low grade glioma. Each patient record contains four MRI modalities, namely T1, T1ce, T2, and FLAIR. The images provided in the dataset are already skull stripped, aligned to the same anatomical space, and resampled to an isotropic resolution of 1 × 1 × 1 mm. The segmentation masks contain three tumor regions: necrotic and non-enhancing tumor core with label 1, edema with label 2, and enhancing tumor with label

4. For evaluation, these labels were grouped into whole tumor, tumor core, and enhancing tumor regions. The dataset was divided at the patient level into 269 training cases, 34 validation cases, and 32 test cases while maintaining the proportion of high grade and low-grade glioma cases. A fixed random seed of 42 was used, and the same split was used for all model experiments.

B. Preprocessing

Each modality volume is normalized independently using z-score normalization computed within the brain mask, defined as the volume mask, the only voxels that have intensity greater than zero. Background voxels remain at zero after normalization. All volumes are trimmed from 240x240x155 to 192x192x144. Remove the padding around brain area. This crop removes about 35% of background voxels from each forward pass and represents our first hardware-aware design decision that decreases activation memory and inference latency on the Jetson Orin Nano. The four modalities after normalization and cropping are stacked to get four channel input tensors of shape (4,192,192,144) Saved as a float32 NumPy array, the pre-processed volumes are accessible to the Data Loader.

C. Proposed Architecture

The suggested Lightweight CNN-Mamba 3D U-Net has a generic encoder-bottleneck-decoder with skip connections. The total parameter count is 2,244,868, which is 10.1 times less than the 3D U-Net baseline.

The overall architecture of the proposed hardware-aware lightweight CNN–Mamba 3D U-Net is illustrated in **Fig. 1**.

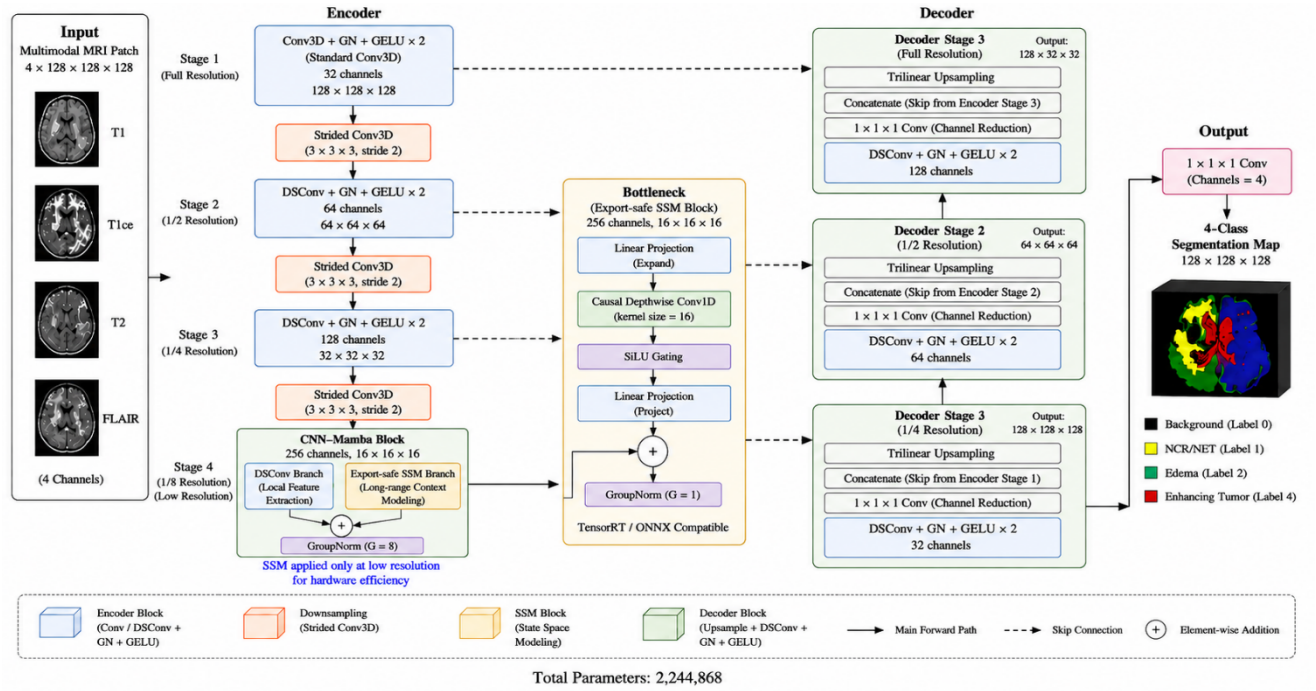

Fig.1.Architecture of the proposed hardware-aware lightweight CNN–Mamba 3D U-Net.

Encoder The encoder consists of four stages, with sequential down sampling and with channel widths of 32,64,128 and 256. Importantly, stage 1 operates on full resolution (128x128x128 patches) using regular 3D convolutions to maintain fine spatial details at entry. The second and third stages uses depth wise-separable convolutions (DSConv) such that each standard 3D convolution in factored into a depth wise spatial filter and pointwise channel mixer. Because of this, the number of parameters per layer is reduced by around 8 to 19 times compared to that of the standard convolutions.

The 8 GB unified memory limitation of the Jetson Orin Nano is the direct motivation for this reduction. Stage 4 works at a spatial resolution of 16x16x16. It applies a CNN-Mamba block, composed of a parallel DSConv branch for local feature extraction and export-safe SSM branch for long-range contextual modelling. The SSM is employed at this resolution only, as the sequence length 4,096 is manageable through computation. In contrast, applying it at a higher resolution (in this case 64x64x64) would yield a

sequence length of 262,144 well above the memory budget of the target device. Between stages, we perform down sampling using strided 3x3x3 convolutions with a stride of 2. These convolutions are learnable and keep more spatial information than max pooling.

The original Mamba selective scan is implemented as a custom CUDA kernel, which is not traceable by ONNX exporter and does not have a corresponding TensorRT plugin from version 8.5. The selective scan is switched out for a causal depthwise Conv1d with kernel size 16 to allow for full deployment, this preserves the causal temporal structure of the original formulation while making use of only standard ops that are ONNX opset 16 and TensorRT 8.5.2 compliant. The block is structured upon mamba expand-gate-project. The input is projected to a size double of the inner dimension. It is split into a content branch processed by the causal convolution and a gate branch

processed by SiLU activation. The two branches are multiplied together for input-dependent gating. Finally, the output is projected back to the model dimension with a residual connection. To substitute LayerNorm, which is not supported natively in TensorRT 8.5, we use GroupNorm with one group.

Bottleneck: An additional SSM block is applied at the 16x16x16 bottleneck to further capture global context before the decoder begins up sampling.

Decoder: The decoder has three up sampling stages which correspond to three encoder stages numbered 3, 2 and 1. Each up-sampling layer comes with a trilinear interpolation and a 1x1x1 convolution to decrease the number of channels. The corresponding skip connection is concatenated with the output of every stage. The concatenated output then goes to two blocks named DSConv for the spatial refinement of the final output. We don't use transposed conv because it causes checkerboard artifacts in 3D, and is hard to export to ONNX and TensorRT.We don't use transposed conv because it causes checkerboard artifacts in 3D, and it is hard to export to ONNX and TensorRT.

Normalization and activation: The network consistently applies GroupNorm, utilizing 8 groups. We don't use BatchNorm, because the batch size during 3D inference is typically 1, and its running stats is unreliable. The convolutional blocks contain GELU activation and the SSM branches SiLU activation in adherence to the original Mamba.

Output: A 1x1x1 convolutional layer maps the final decoder feature map to a four-class logit volume, which is passed through softmax during inference to produce per-voxel class probabilities.

D. Training Configuration

All models get trained using a GPU (24 GB VRAM) of the NVIDIA RTX 4500 Ada generation. The equal weights soft Dice loss and cross-entropy loss computed on all four outputs is used as the combined loss function. In the cross-entropy term, the class weights of background, NCR/NET, edema, and ET have values of 0.1, 1.0, 1.0, and 2.0 respectively. We employ the AdamW optimizer with a learning rate of 1e-4 which decays via cosine annealing to 1e-6 over 40 epochs. Validation whole-tumor Dice mitigates overfitting early stopping with patience of 10 epochs. The training patches of size 128x128x128 are randomly extracted on pre-processed volumes with the batch size as 2. This includes random flipping along each axis independently with probability 0.5, as well as random intensity scaling per channel in the range 0.9, 1.1.

E. TensorRT Deployment on Jetson Orin Nano

The model is exported to ONNX opset 16 format after transferring weights to the architecture compatible with TensorRT. The Jetson Orin Nano 8 GB (JetPack R35.5.0, TensorRT 8.5.2.2 and CUDA 11.4) compiled three TensorRT engines: FP32, FP16 and INT8. The Entropy calibration is used by INT8. In short, all benchmarks were run at 15 W power with the clocks locked using Jetson clocks for reproducible measurements. All configurations of precision are tested with a duration of 10 warmup runs followed by a sustained window of 30 seconds with trtexec.

The developed edge inference interface deployed on the NVIDIA Jetson Orin Nano is shown in Fig. 2, including multimodal MRI input, TensorRT inference, segmentation visualization, and quantitative tumor-volume estimation.

(a) Multimodal MRI input and inference interface

(b) Segmentation and tumor-volume output

Fig. 2. Jetson Orin Nano deployment interface for multimodal MRI segmentation and tumor-volume estimation.

F. Evaluation Metrics

The performance of the segmentation is evaluated on the 32-case test set by the Dice similarity coefficient (DSC), Hausdorff distance at 95th percentile (HD95) in mm and intersection over union (IoU) for each of the 3 tumor regions. If the ground truth ET is empty, and predicted ET is empty then DSC is assigned 1.0 (BraTS challenge protocol). Various metrics are used when evaluating hardware performance and these include the GPU inference latency, the end-to-end latency, peak RAM use, GPU utilization, GPU temperature, power usage in watts, and energy per proton and MRI volume in joules (which is the product of mean power and the end-to-end latency). The Bland-Altman analysis uses the predicted versus ground truth volumes to validate the tumor volume estimation. This is repeated for the 32 test cases and the following results are reported: the mean bias, standard deviation, 95% limits of agreement, and Pearson correlation coefficient. Because all of the volumes are acquired at 1 mm isotropic resolution, every predicted voxel corresponds to exactly 1 mm3. So, volume computation is precise and there is no need of a resampling correction.

IV. RESULTS AND DISCUSSION

A. Training Convergence Analysis

Training loss, validation loss, and regional validation Dice scores were used to investigate the convergence behaviour of the proposed CNN–Mamba model. The loss values were not changing significantly in the later epochs, and the Dice score (WT, TC and ET) for validation also had a slight variation, as shown in Fig. 3, which means that the proposed network was converging steadily.

B. Segmentation Performance

The CNN–Mamba model was tested to 32 held-out BraTS 2019 test cases, along with the conventional 3D U-Net under the same experimental protocol. (WT), tumor core (TC), and enhancing tumor (ET). The absolute gain overall in comparison to 3D U-Net was 0.0017 for WT, 0.0661 for TC, and 0.0298 for ET. TC constituting about 8.61% relative increase in Dice score was the largest improvement observed.

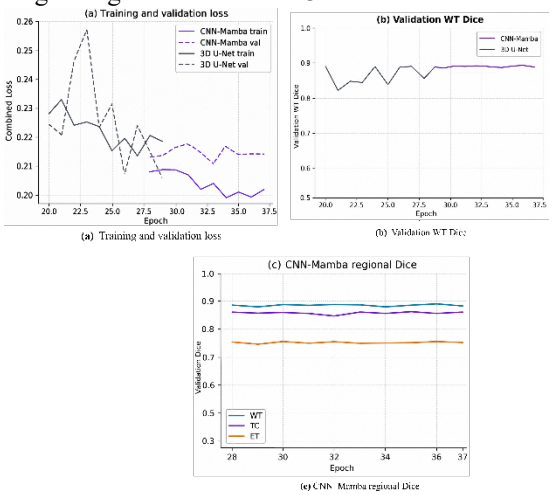

Fig. 3. Training convergence analysis: (a) training and validation loss, (b) validation WT Dice, and (c) regional validation Dice of the proposed CNN–Mamba model.

As seen in Table 1, the proposed model attained Dice scores of 0.8621, 0.8335, and 0.6695 for respective whole tumor. The proposed model enhanced the IoU of all three tumor regions. Most importantly, substantial gains in boundary localization were also observed. The value of WT HD95 reduced from 8.1238 mm to 4.0525 mm, a reduction of 50.12% while the TC and ET HD95 dropped by 22.51% and 17.00% respectively. As per the results, the proposed framework enhances spatial consistency with reference segmentation and WT and TC benefits the most.

Table 2. Segmentation performance on the BraTS 2019 test set

Metric	3D U-Net WT	CNN – Mamba WT	3D U-Net TC	CNN – Mamba TC	3D U-Net ET	CNN – Mamba ET
Dice	0.8604	0.8621	0.7674	0.8335	0.6397	0.6695
IoU	0.7681	0.7702	0.6714	0.7309	0.5428	0.5707
HD95 (mm) ↓	8.12	4.052	4.68	3.630	43.40	36.02
Sensitivity	0.8957	0.9239	0.7649	0.8738	0.6361	0.6361
Precision	0.8529	0.8223	0.8391	0.8154	0.6359	0.6473

CNN–Mamba increased WT sensitivity from 0.8957 to 0.9239 and TC sensitivity from 0.7649 to 0.8738, indicating improved detection of tumor voxels. ET sensitivity remained essentially unchanged.

As shown in Fig. 4, CNN–Mamba improves Dice performance while reducing HD95 across WT, TC, and ET

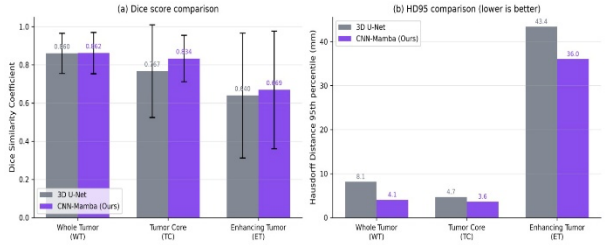

Fig. 4. Dice and HD95 comparison of 3D U-Net and CNN–Mamba.

It was observed that WT, TC accuracies had reduced moderately, which predicted a higher sensitivity, but also some false-positive predictions. The precision of ET improved from 0.6359 to 0.6473.

The representative segmentation examples are displayed in Fig. 5, which shows that the tumor regions predicted and the ground-truth tumor regions agree well in space. As such, the proposed approach introduces a desirable overall overlap and boundary-localization improvement, as opposed to uniformly improving all individual metrics.

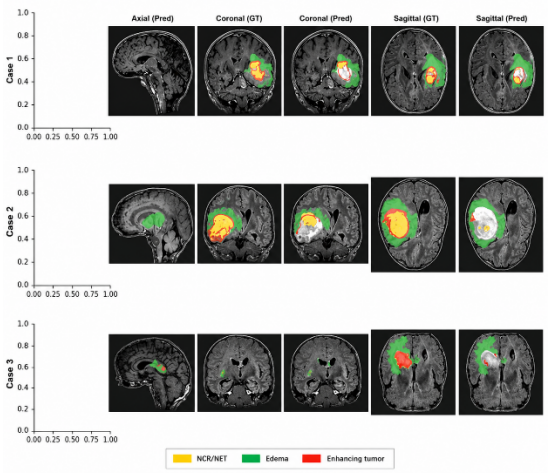

Fig. 5. Qualitative CNN–Mamba segmentation results on representative BraTS 2019 test cases.

The variability of TC Dice was also reduced substantially, with the standard deviation decreasing from 0.2431 for 3D U-Net to 0.1215 for CNN–Mamba. This reduction, together with the higher mean TC Dice, indicates more consistent tumor-core segmentation across the test cases.

C. Model Efficiency and Edge Deployment

The proposed framework significantly reduces the total number of parameters in the model. CNN–Mamba has 2,244,868 parameters, while the regular 3D U-Net has 22,580,964 parameters. The parameters dropped by 90.06%

indicating that the proposed model is about 10.06 times smaller while having better Dice and HD95 performances. Its efficiency arises from the use of depthwise-separable convolutions and limiting state-space modeling to low-resolution feature maps. As long-range processing is computationally expensive, SSM doesn't apply it throughout the network but at a spatial resolution of 16 × 16 × 16 where global contextual information can be modelled with a shorter sequence length. The NVIDIA Jetson Orin Nano implementation of TensorRT showed that the architecture can be used for inference in constrained resources. The average patch-level GPU latency for the FP32 engine is 95.25 ms.

On the other hand, FP16 was able to reduce latency to 83.99 ms. Furthermore, INT8 achieved the lowest latency of 75.83 ms. Subsequently, it achieved the highest throughput of 13.17 inferences/s, giving rise to a speedup of 1.26× over FP32.

Table 3. TensorRT performance on NVIDIA Jetson Orin Nano

Precision	Latency (ms)	Throughput (qps)	Speedup
FP32	95.25	10.48	1.00
FP16	83.99	11.88	1.13
INT8	**75.83**	**13.17**	**1.26**

During FP16 inference, GPU utilization remained between 93–99%, with a GPU temperature of approximately 62–63°C and power consumption of approximately 15 W.

Table 4. Bland–Altman tumor-volume analysis

Region	Bias (mm³)	95% Limits of Agreement (mm³)	Pearson r
WT	+18,151.2	−26,037.1 to +62,339.4	**0.9699**
TC	+4,740.0	−18,289.9 to +27,769.9	**0.9510**
ET	+3,775.0	−10,171.6 to +17,721.6	**0.9195**

The Bland–Altman plots in Fig. 7 illustrate the agreement between predicted and ground-truth tumor volume This observation is consistent with the increased WT and TC sensitivity and the corresponding reduction in precision. Although the strong Pearson correlations demonstrate that predicted volumes follow the variation in ground-truth tumor burden closely, the observed biases emphasize that correlation alone should not be interpreted as complete volumetric agreement.

The effect of TensorRT precision optimization on latency and throughput is shown in Fig. 6.

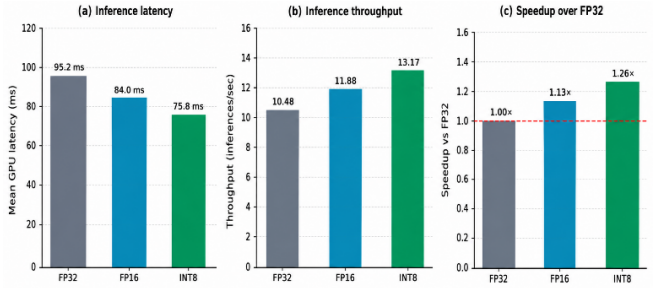

Fig. 6. TensorRT inference performance on NVIDIA Jetson Orin Nano.

The estimated energy consumption was approximately 1.26 J per inference patch. These results demonstrate that the proposed network is not only lightweight in terms of parameter count but is also practically executable using reduced-precision TensorRT inference on the target embedded GPU.

D. Quantitative Tumor Volume Assessment

Quantitative assessment was performed by comparing tumor volumes derived from predicted segmentation masks with ground-truth volumes. Strong correlations were observed for all three tumor regions, with Pearson correlation coefficients of 0.9699 for WT, 0.9510 for TC, and 0.9195 for ET. Bland–Altman analysis revealed positive biases for WT, TC, and ET, indicating a systematic tendency toward volume overestimation.

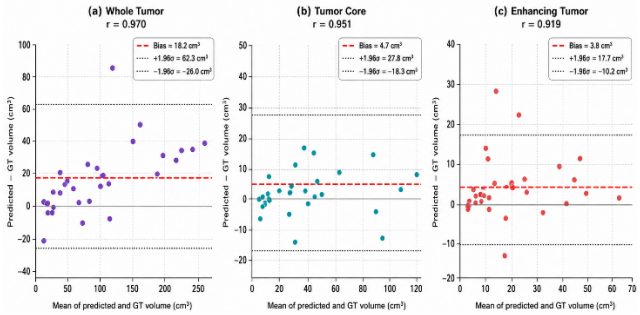

Fig. 7. Bland–Altman analysis of WT, TC, and ET volume estimation.

E. Overall Discussion

The results demonstrate that a good compromise of segmentation accuracy and efficiency is obtained by using efficient 3D convolution processing and a reduced resolution state-space model. CNN–Mamba achieved the highest improvement in tumor-core segmentation, which showed a Dice score of 0.7674 to 0.8335 and less inter-case variability, and a sensitivity score of 0.7649 to 0.8738. In the meantime, there are significant decreases in the amount of boundary localization indicated by the HD95 values. In this regard, the primary advantage of the proposed method is that it not only improves the segmentation accuracy, but does so with approximately 10% of the parameters of a

conventional 3D U-Net, and can run on TensorRT on the NVIDIA Jetson Orin Nano. In the field of 3D medical imaging, it is crucial to strike the right balance between accuracy and efficiency. Comparing the dice score of ET with WT and TC with respect to the HD95, it is observed that ET remains the most challenging region with dice score of 0.6695 as compared to the dice score of WT and TC, which are 0.3616 and 0.3654 respectively. Further, the current evaluation is carried out with respect to the held-out test set of 32 cases and BraTS 2019. The latency measured is for patch-level and not end-to-end network inference. Therefore, future directions ought to involve validation on external datasets, statistical analysis using multiple seeds, full volume latency characterization and further optimization of enhancing tumor segmentation.

V. CONCLUSION

A hardware-aware lightweight CNN–Mamba 3D U-Net functional design study on multimodal brain tumor segmentation and quantitative assessment on the NVIDIA Jetson Orin Nano. The suggested framework utilizes 3D convolutions which are depthwise-separable, to extract the local features efficiently. Moreover, long-range contextual representation is afforded via the low-resolution state-space modeling. On the 32-case BraTS 2019 test set, it achieved Dice scores of 0.8621, 0.8335 and 0.6695 for whole tumor, tumor core and enhancing tumor, respectively. The parameter amount of the model proposed is 2.24 million which is a reduction from 22.58 million in the 3D U-Net (90.06% reduction). Also, the Dice and HD95 performance improved in all three tumor areas. The Jetson Orin Nano deployment of TensorRT enabled real-world edge inference, with INT8 executing with mean patch-level latency of 75.83 ms and throughput of 13.17 inferences/s. Quantitative tumor-volume estimation further demonstrated intense linear correlations with reference volumes, with Pearson coefficient r values of 0.9699, 0.9510, and 0.9195 for WT, TC, and ET, respectively. The findings suggest that CNN–Mamba architectures are useful for resource-efficient 3D medical image analysis. In the future, we will verify externally, optimize full-volume latencies, implement statistical multi-run analysis, and refine enhancing-tumor segmentations.

REFERENCES

[1] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, B. Huang, C. Liang, G. Han, Z. Liu, and C. Han, "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 8, pp. 2451–2461, Aug. 2023, doi: 10.1109/TMI.2023.3250474.

[2] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. Maier-Hein, "MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation," *arXiv preprint arXiv:2303.09975*, 2024.

[3] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnFormer: Volumetric medical image segmentation via a 3D Transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023, doi: 10.1109/TIP.2023.3293771.

[4] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2024.

[5] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[6] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*, Lecture Notes in Computer Science, vol. 15009, pp. 578–588, 2024, doi: 10.1007/978-3-031-72111-3_54.

[7] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang, H. Zheng, and S. Wang, "Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*, pp. 615–625, 2024, doi: 10.1007/978-3-031-72114-4_59.

[8] R. Zhou, J. Wang, G. Xia, J. Xing, H. Shen, and X. Shen, "Cascade residual multiscale convolution and Mamba-structured UNet for advanced brain tumor image segmentation," *Entropy*, vol. 26, no. 5, Art. no. 385, Apr. 2024, doi: 10.3390/e26050385.

[9] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. Shahbaz Khan, "UNETR++: Delving into efficient and accurate 3D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 9, pp. 3377–3390, Sep. 2024, doi: 10.1109/TMI.2024.3398728.

[10] J. Megha and G. Thimmaiah, "A lightweight CNN framework for real time liver lesion classification from CT images with YOLOv8 guided localization," *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.*, vol. 18, pp. 616–628, Jul. 2026, doi: 10.70917/ijcisim-2026-3776.

[11] E. Alwadee, X. Sun, Y. Qin, and F. Langbein, "LATUP-Net: A lightweight 3D attention U-Net with parallel convolutions for brain tumor segmentation," *Comput. Biol. Med.*, vol. 184, Art. no. 109353, 2024, doi: 10.1016/j.compbiomed.2024.109353.

[12] X. Sun, W. He, J. Ruan, Z. Yuan, Z. Sun, and J. Zhang, "An edge enhanced 3D Mamba U-Net for pediatric brain tumor segmentation with transfer learning," *Med. Phys.*, vol. 52, no. 10, Art. no. e70002, Oct. 2025, doi: 10.1002/mp.70002.

[13] M. M. Rahman, M. Munir, and R. Marculescu, "EfficientMedNeXt: Multi-receptive dilated convolutions for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*, Lecture Notes in Computer Science, vol. 15963, pp. 196–206, 2025, doi: 10.1007/978-3-032-04965-0_19.

[14] Z. Xing, T. Ye, Y. Yang, D. Cai, B. Gai, X.-J. Wu, F. Gao, and L. Zhu, "SegMamba-V2: Long-range sequential modeling Mamba for general 3-D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 45, no. 1, pp. 4–15, 2026, doi: 10.1109/TMI.2025.3589797.

[15] SwiM-UNet with multi-dimensional adaptor for

efficient on-device medical image segmentation," *Sci. Rep.*, vol. 16, Art. no. 5807, 2026, doi: 10.1038/s41598-026-35771-4.

[16] Y. Ma, Y. Liu, J. Cheng, H. Zhan, G. Qi, X. Zhang, and Z. Zhu, "CIHM: Context-insight hybrid Mamba for efficient medical image segmentation," *Vis. Intell.*, vol. 4, Art. no. 12, 2026, doi: 10.1007/s44267-026-00113-5.

[17] T. D. Q. Dang, H. H. Nguyen, and A. Tiulpin, "LoG-VMamba: Local-Global Vision Mamba for Medical Image Segmentation," in *Computer Vision – ACCV 2024*, Lecture Notes in Computer Science, vol. 15473, pp. 222–240, Springer, 2024.