

A COMPACT MOBILENETV3-UNET FOR RESOURCE-EFFICIENT BRAIN TUMOR SEGMENTATION

Syed Zahid Shah^{*1}, Atif Jan², Fawad Ahmad³, Syed Yousaf Shah⁴, Umar Farooq⁵

^{*1,2,3,4}Department of Electrical Engineering, University of Engineering and Technology Peshawar Pakistan

⁵World Web Tree, Riyadh, Kingdom of Saudi Arabia

¹syedzahidshah@uetpeshawar.edu.pk, ²atifjan@uetpeshawar.edu.pk, ³engrfawad64@gmail.com, ⁴24pwele6354@uetpeshawar.edu.pk, ⁵umarpak995@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22653400>

Keywords

Brain tumor segmentation, BraTS2021, lightweight deep learning, pseudo-labeling, Dice score, semi-supervised learning, U-Net, MobileNetV3.

Article History

Received: 14 July 2026

Accepted: 18 August 2026

Published: 31 August 2026

Copyright @Author

Corresponding Author: *

Syed Zahid Shah

Abstract

Introduction: Manual segmentation of brain tumors is labor-intensive and susceptible to inter-observer variability, yet accurate segmentation using magnetic resonance imaging (MRI) is essential for diagnosis, treatment planning, radiation therapy, and follow-up monitoring. Current approaches focus on complex 3D networks, transformers, or ensembles of multiple models, all requiring extensive GPU memory and training time, making them impractical for resource-constrained environments.

Objective: This paper presents a lightweight segmentation framework based on the MobileNetV3-UNet architecture.

Methodology: The model was trained and evaluated on BraTS2021, with initial validation on BraTS2020, segmenting the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) across four MRI modalities (FLAIR, T1, T1CE, T2). To reduce reliance on costly expert annotation, pseudo-label semi-supervised learning (SSL) was evaluated under limited-label budgets (20% and 40%).

Results: The fully supervised model achieved a Mean DSC of 0.9127 on BraTS2021 (WT: 0.9385, TC: 0.9114, ET: 0.8882) with only 10.1 M parameters and 2.14 GFLOPs, substantially reducing computational cost compared to heavy state-of-the-art architectures. Under limited labels, patient-wise 3D evaluation showed consistent improvement, with Mean DSC increasing from 0.8862 (supervised 20%) to 0.8919 (SSL 20%) and 0.8942 (SSL 40%).

Conclusion: These results demonstrate that a lightweight network can achieve competitive segmentation accuracy at a fraction of the computational cost of heavier models, and that unlabeled data can enhance performance when annotated data is scarce.

1. INTRODUCTION

Characterization of the brain tumor is difficult because the size and shape of the tumor and the tissue characteristics of the tumor are variable. Gliomas are often primary brain tumors that have irregular and infiltrative borders. MRI is the imaging modality of choice because of its high soft-tissue contrast that is free from the use of ionizing radiation. FLAIR, T1, T1CE and T2

sequences give complementary information to identify tumor regions and surrounding edema [1, 3]. Accurate segmentation can aid diagnosis, treatment planning, surgery, radiotherapy, and follow-up. Manual segmentation, however, can be time consuming and can differ among and between radiologists. These restrictions make this segmentation challenging to be scalable and reproducible, and therefore need automated

and efficient techniques for brain tumors segmentation [2, 19]. Today, deep learning has made a significant impact on this field. The number of handcrafted features and fragile, rule-based pipelines have been replaced by representations learnt directly from the data, and Convolutional Neural Networks (CNNs) subsequently came to the fore as the cutting edge method in biomedical segmentation [1]. This progress was driven by the BraTS benchmark, which offers multimodal data co-registered with expertly annotated and a standardized evaluation protocol with which methods can be compared. However, with repeated mismatches, accuracy has gradually increased and using an encoder-decoder with skip connections is a common approach [3, 4, 5].

1.1. Deep Learning and Computational Challenges

Two practical costs have been associated with this advancement, though, and this paper addresses the two following: The first is the computation based one. The state of the art segmenters or ensembles of several large models (based on U-Nets and transformers) have been shown to be the most effective, and are fully three-dimensional [6, 7]. These systems require a lot of energy, long training times, a large amount of GPU memory, and record rich, volumetric context, but are impractical to be implemented on commodity hardware. This is a major accessibility problem: more than many clinical and academic environments, where the only machine available to use would be a mid-range GPU could afford to install one of these machines. Such a cost-effective model which recovered most of the accuracy would significantly increase accessibility [13, 16].

1.2. Semi-Supervised Learning and Segmentation Challenges

The second cost is the cost of the annotations. Obtaining voxel-by-voxel labels for tumors requires trained experts and is slow, expensive, and throughout clinical archives is far less common than large amounts of unlabeled

scanning data. The idea of semi-supervised learning (SSL) takes advantage of this asymmetry, namely to learn an unlabeled set of data and a small labeled one [10]. A common SSL method is pseudo labeling, where the labels are generated from unlabeled images by a teacher model trained on labeled images, while consistency based methods such as Mean Teacher [9] have shown effectiveness in tasks like medical image analysis to solve the problem of the amount of labeled images required. These costs are compounded by some inherent problems. However, due to the excessive ratio of tumor to healthy images, and the distribution of the images that is usually very class-imbalanced, the learning problem needs to be formulated using some degree of overlap-tolerant design objectives. The tumor has an enhancing profile and is frequently highly variable amongst different patients, and small and irregular in size, which are the most difficult areas to segment. Finally, there can be a loss of model performance when the image source changes due to patient variations and/or the acquisition protocol and scanners. It should be difficult to defeat with these properties and low in computation cost [7].

The goal of this research is to find the computational efficient solution that can maintain strong segmentation performance in the limited-label scenario. It aims to explore the ability of a lightweight MobileNetV3 encoder which combines inverted residual blocks and Squeeze-and-Excitation attention with a U-Net decoder to offer an effective segmentation accuracy and computation cost trade off. The study also assesses the performance of pseudo-label semi-supervised learning with confidence-based filtering to see if unlabeled data can enhance segmentation with only a small fraction of labeled data. BraTS datasets are used to evaluate model performance on the basis of global Dice score and patient-wise three-dimensional Dice score to achieve a comprehensive evaluation of segmentation performance as shown in Figure 1.

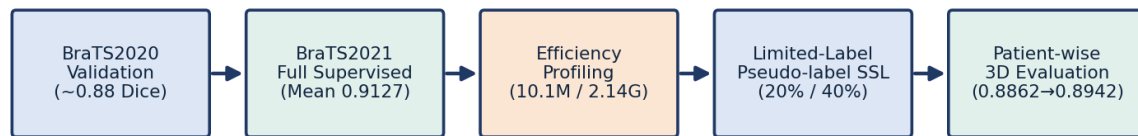


Fig. 1. Staged research and training pipeline, from BraTS2020 validation through full supervised BraTS2021 training and efficiency profiling to limited-label pseudo-label SSL with patient-wise 3D evaluation.

2. LITERATURE REVIEW

2.1. Encoder-Decoder Segmentation

The fully convolutional networks (FCN) replaced fully connected (FC) layers of classifiers with convolutional and up sampling layers and made end-to-end prediction of dense pixel-level semantic information trainable. This idea was adopted to biomedical data by the U-Net and now the U-Net is widely used to serve as the backbone for most of the biomedical applications [1]. It encodes finer and finer details of the image and gradually more and more abstract concepts and then decodes the image using skip connections which add back the fine detail of the image boundaries. This setting is extended with variants like attention gates to suppress the information coming from an irrelevant background, separable convolutions to enlarge the receptive field at a fixed cost [18] or by introducing residual or dense connectivity to help with optimization. In particular, the encoder is modular and can easily be replaced with an efficient backbone without changing the decoder as employed in this work [12].

2.2. Heavy 3D and Transformer Models

Whole-volume interpretation is achieved in fully three-dimensional networks, which include inter-slice context. On medical challenges, the self-configuring nuns-Net clearly assists as a great baseline, whilst the volumetric V-Net [6, 7] introduced the Dice objective for imbalanced segmentation that has been widely adopted. Recently the transformer-based hybrids, such as Transnet and Swin UNETR have performed best on the BraTS leaderboard for modelling long-range dependencies [13]. The models can

have high compute cost including the need for 3D convolutions (requiring significant memory), slow training and the computational costs of assembling, but lightweight alternatives are encouraged [16].

2.3. Lightweight Networks

The use of depth wise separable and inverted-residual convolutions, squeeze-and-excitation (SE) channel attention [8] and hardware-aware neural architecture search, like in MobileNetV3 are employed to reduce computation. The building blocks of all of these backbones are the same, but they are all tailored towards natural-image classification. This is thus a principle-based approach to a compact model based on a backbone such as segmentation encoding [15].

2.4. Semi-Supervised Learning

In semi-supervised learning, we need a very large amount of unlabeled data and a relatively small amount of labeled data, under the premise that both labeled and unlabeled data have some fruitful information about the distribution of the input. In pseudo labeling [10], a teacher is trained on the small subset of labeled data and the prediction made by the teacher with the confidence taken as the label on unlabeled test data. Alternatively, consistency methods give stable predictions when making changes: Using an exponential moving average of student weights, the Mean Teacher is a more consistent teacher; Fix Match [11] combines weak and strong augmentation along with using a confidence threshold; and dual-task consistency has been extended to medical segmentation. Typical solutions include some types of balance

between using more unlabeled data and preventing systematic errors (confirmation bias), which involve confidence thresholding, uncertainty estimation and weight averaging. Here the pseudo-labeling technique is followed due to its simplicity and low overhead for the sake of efficiency and confidence filtering is used to overcome its main drawback [17].

2.5. Research Gap

All three of these properties are rarely discussed together in this literature: the motivation for efficient architectures is typically to work under full supervision; SSL is typically implemented on heavier networks; and the evaluation of the architectures is not necessarily performed on a patient-wise or volumetric basis. Together with the compact model and limited-label pseudo-label SSL, patient-wise three-dimensional evaluation is little explored, and this paper seeks to fill this gap.

3. MATERIALS AND METHODS

3.1. Research Design

The design of this study is quantitative experimental research design, with a sequential three-stage design. Stage one is feasibility on BraTS2020. In Stage 2, full supervised training and testing are performed on BraTS2021 to assess peak segmentation accuracy and the computational efficiency. Stage three is dedicated to both evaluating the label efficiency of semi-supervised learning with limited label budgets (20% and 40%). A comparative approach places the proposed lightweight model

in comparison with its supervised baselines and compares it with heavy architectures found in the literature. The slice-wise global Dice metric and patient-wise Dice metric are both reported to provide clinical relevance.

3.2. Data Collection

The secondary analysis was performed on benchmark datasets publicly available at the Brain Tumor Segmentation (BraTS) challenge through the Cancer Imaging Archive (TCIA). All cases contain co-registered skull stripped MRI volumes and expert annotated voxel level tumor annotations. This study did not involve the collection of primary data. The BraTS data are fully de-identified before they are published and the original institutions keep ethical approval for the data.

3.3. Datasets

Two public BraTS datasets are used. The BraTS2020 training set consists of 369 cases with annotations, to be used initially as a validation set. The primary experiments are carried out on BraTS2021, consisting of 1,251 annotated cases. Four MRI modalities: FLAIR, T1, T1CE, and T2, are included in each case, as are annotations by experts that label three types of tissue: necrotic and non-enhancing tumor core, peritumoral edema, and enhancing tumor. These are native labels organized into three evaluation regions, nested: Whole Tumor (WT), Tumor core (TC), and Enhancing tumor (ET). As Table 1 summarizes the datasets.

Table 1: Dataset Summary

Dataset	Purpose	Cases	Modalities	Output Regions
BraTS2020	Initial validation	369	FLAIR, T1, T1CE, T2	Tumor Dice
BraTS2021	Main supervised + SSL	1,251	FLAIR, T1, T1CE, T2	WT, TC, ET Dice

3.4. MRI Modalities and Preprocessing

The four modalities are stacked into a 4-channel input so the network can learn cross-modal relationships directly. FLAIR highlights edema and abnormal hyperintense tissue; T1 provides anatomical structure and baseline contrast; T1CE identifies enhancing tumor after gadolinium injection; and T2 emphasizes fluid-

rich and abnormal regions. Each modality is z-score normalized independently over its non-zero brain voxels:

$$\hat{x} = (x - \mu) / \sigma$$

where μ and σ are the mean and standard deviation computed over non-zero voxels. This normalization ensures all modalities share similar intensity ranges. The dataset is split at the

patient level into 70% training, 15% validation, and 15% test. Adjacent slices from the same patient are highly correlated; therefore, slice-level mixing is avoided to prevent data leakage.

Standard augmentations including random horizontal and vertical flips, small rotations up to 10° , and mild intensity perturbations are applied during training.

Table 2: MRI Modalities and Their Role

Modality	Role in segmentation
FLAIR	Edema and abnormal hyperintense tissue
T1	Anatomical structure and baseline contrast
T1CE	Enhancing tumor after contrast
T2	Fluid-rich and abnormal regions

3.5. MobileNetV3-UNet Architecture

The proposed model combines a MobileNetV3-Small encoder with a standard U-Net decoder. The encoder comprises a stack of inverted-residual bottleneck (MBConv) blocks. Each block applies a 1×1 expansion convolution, a depthwise 3×3 convolution, a squeeze-and-excitation (SE) attention module, and a 1×1 projection convolution. A residual connection is added when input and output dimensions match. The SE module generates per-channel gates via global average pooling:

$$s = \sigma(W_2 \delta(W_1 z))$$

where z is the pooled descriptor, δ is ReLU, σ is the sigmoid, and s rescales the channels. The decoder progressively upsamples encoded features and fuses them with matching encoder features through skip connections to restore boundary detail. The final 1×1 convolution maps decoder output to class probabilities. The complete model has approximately 10.1 M parameters and requires 2.14 GFLOPs per forward pass, making it suitable for modest hardware.

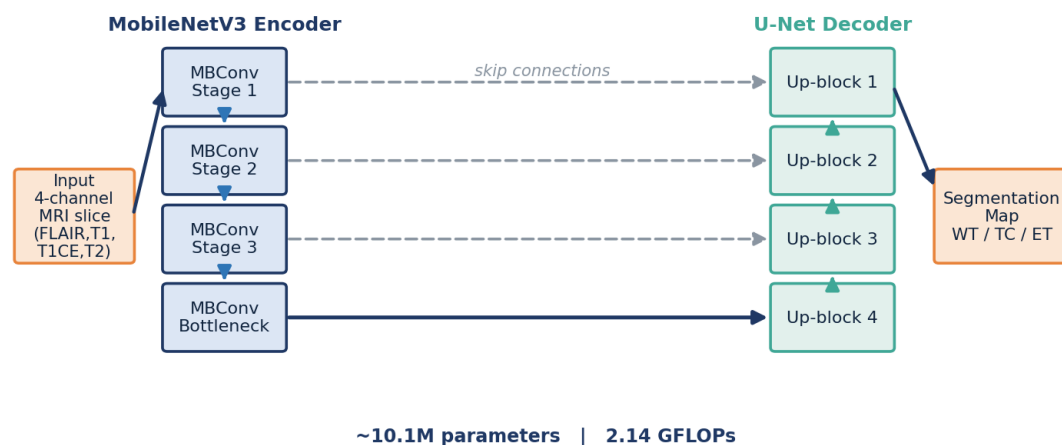


Figure 2: MobileNetV3-UNet architecture. A MobileNetV3 encoder of inverted-residual (MBConv) blocks feeds a U-Net decoder through skip connections; the model uses about 10.1 M parameters and 2.14 GFLOPs.

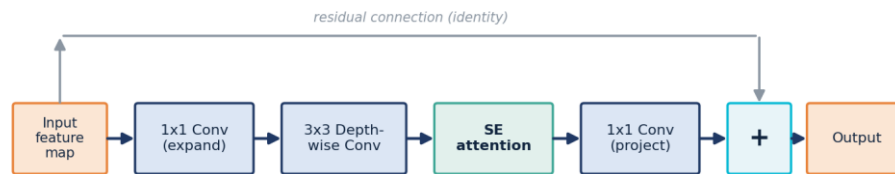


Figure 3: Inverted-residual (MBConv) block with squeeze-and-excitation attention used in the MobileNetV3 encoder.

3.6. Loss Function and Optimization

Because healthy voxels greatly outnumber tumor voxels, a composite loss combining soft Dice loss and weighted cross-entropy is used. The soft Dice loss for a class is:

$L_{\text{Dice}} = 1 - (2 \sum_i p_i g_i + \epsilon) / (\sum_i p_i + \sum_i g_i + \epsilon)$
 where p_i is the predicted probability, g_i is the ground-truth indicator at voxel i , and $\epsilon = 1 \times 10^{-6}$ ensures numerical stability. The weighted cross-entropy term is:

$$L_{\text{CE}} = - \sum_c g_c \log p_c$$

The total objective is $L = L_{\text{Dice}} + \lambda L_{\text{CE}}$ with $\lambda = 1.0$. The model is optimized with AdamW, which decouples weight decay from the gradient update. Training runs for 50 epochs with a batch size of 16, learning rate of 1×10^{-4} , and weight decay of 1×10^{-2} . The checkpoint with the best validation Mean DSC is retained for final evaluation. Algorithm 1 outlines the supervised training procedure.

Algorithm 1: Supervised Training

Input: labeled set D_l ; model f_θ ; epochs E

- 1: initialize θ
- 2: for $e = 1$ to E do
- 3: for each mini-batch (X, Y) in D_l do
- 4: $P \leftarrow f_\theta(X)$
- 5: compute $L = L_{\text{Dice}}(P, Y) + \lambda L_{\text{CE}}(P, Y)$
- 6: $\theta \leftarrow \text{AdamW update on } \nabla_\theta L$
- 7: evaluate Mean DSC on validation set
- 8: save θ if validation DSC improved
- 9: return best θ

3.7. Pseudo-Label Semi-Supervised Learning

The SSL pipeline first trains a teacher model on the labeled subset. The teacher predicts class probabilities for each unlabeled slice, and hard pseudo-labels are assigned pixel-wise by taking the argmax over classes. A slice is retained for student training only if its mean maximum class confidence exceeds a threshold $\tau = 0.90$:
 $\text{mean}(\max_c f_\theta(X_j)_c) \geq \tau$

This confidence filtering ensures only high-quality predictions contribute, shielding the student from unreliable teacher outputs and reducing confirmation bias. The retained pseudo-labeled slices are combined with the original labeled data to train a student model minimizing the same composite loss L over both sets. Algorithm 2 details the procedure.

Algorithm 2: Pseudo-Label SSL

Input: labeled D_l ; unlabeled D_u ; threshold τ

- 1: train teacher θ^T on D_l (Algorithm 1)
- 2: $P_s \leftarrow \emptyset$
- 3: for each X_j in D_u do

```

4:  $p \leftarrow f\theta^T(X_j); \hat{y} \leftarrow \arg \max p$ 
5: if  $\max(p) \geq \tau$  then add  $(X_j, \hat{y})$  to  $P_s$ 
6: train student  $\theta^s$  on  $D_l \cup P_s$  (Algorithm 1)
7: return  $\theta^s$ 

```

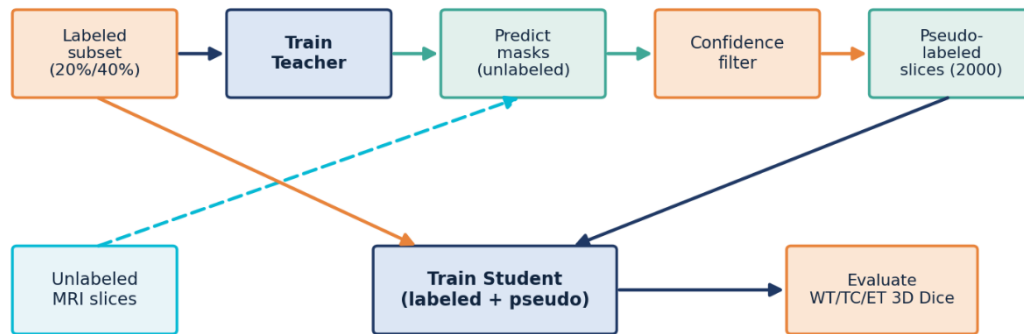


Figure 4: Pseudo-label semi-supervised workflow. A teacher trained on the labeled subset predicts masks for unlabeled slices; confidence-filtered pseudo-labels are combined with the labeled data to train the student, which is then evaluated with patient-wise 3D Dice.

3.8. Limited-Label Experimental Design

Two different limited label settings were tested – 20% and 40% of the whole training patients of the BraTS2021 cohort (about 250 and 500 patients, respectively). A supervised baseline was created for each setting based on the provided labels. A matched SSL variant was trained on the same labeled patients and a fixed number of 2000 pseudo-labeled slices that were selected from the remaining unlabeled slices. This paired design also allows the effect of unlabeled data to be determined while keeping other experimental factors constant.

3.9. Evaluation Metrics

Performance is measured using the Dice Similarity Coefficient (DSC), which quantifies

overlap between predicted region P and reference region G :

$$DSC = 2|P \cap G| / (|P| + |G|)$$

DSC is reported separately for WT, TC, and ET, with the Mean DSC summarizing the three. In addition, patient-wise three-dimensional Dice is computed by aggregating all slices of a patient into a volume before scoring:

$$DSC_{3D}(k) = 2 \sum_v p_v g_v / (\sum_v p_v + \sum_v g_v)$$

for patient k over all voxels v , then averaged across patients. This volumetric measure is closer to clinical practice, where the quantity of interest is the whole tumor volume rather than the slice average.

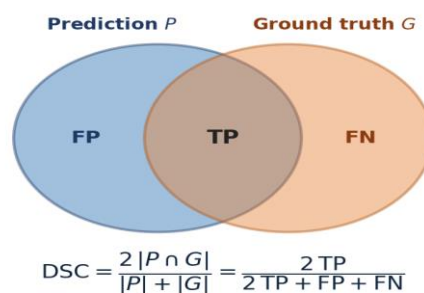


Figure 5: Dice similarity coefficient as the overlap between prediction P and ground truth G .

3.10. Implementation Details

All experiments were implemented in PyTorch and executed on a single NVIDIA Tesla T4 GPU via Google Colab, deliberately reflecting the

modest hardware targeted by the efficiency objective. Table 2 summarizes the training configuration.

Table 3: Training Configuration

Setting	Value
Framework / Hardware	PyTorch, single NVIDIA T4 GPU (Google Colab)
Encoder Backbone	MobileNetV3-Small
Optimizer	AdamW
Batch Size	16
Input Slice Size	128×128
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-2}
Epochs	50
Loss Weight λ	1.0
Confidence Threshold τ	0.90
Pseudo Slices (SSL)	2,000
Train / Val / Test Split	70% / 15% / 15% (patient-level)

4. Results and Discussion**4.1. BraTS2020 Validation**

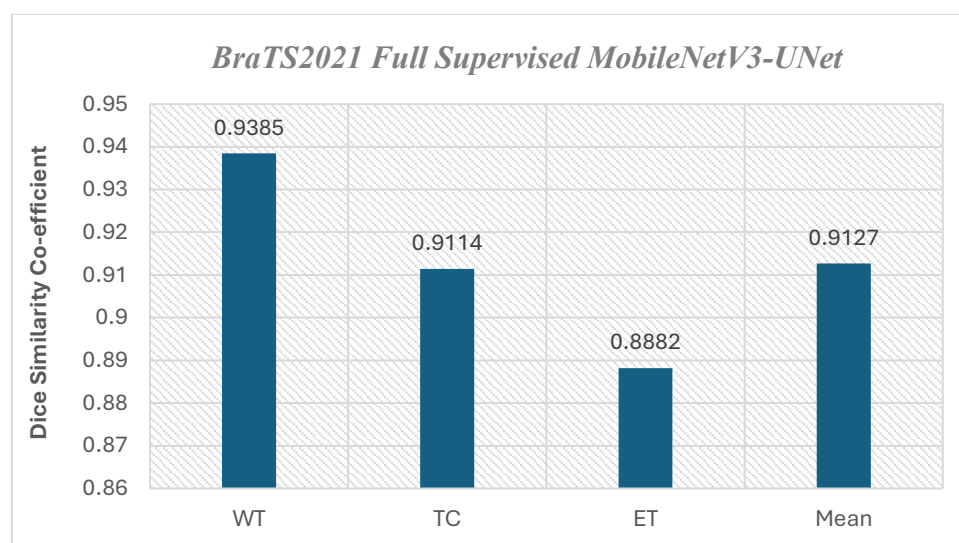
The first experiment that we ran on BraTS2020 showed proof of concept as the lightweight model MobileNetV3 could achieve a Dice of ~ 0.88 , competitive for an approach using such a small 2D model, encouraged us to test it on the larger dataset BraTS2021.

4.2. Full Supervised BraTS2021

The fully supervised MobileNetV3-UNet attained reasonable region-wise scores: WT 0.9385, TC 0.9114, ET 0.8882, and Mean DSC 0.9127 on BraTS2021. As might be expected, WT was segmented most accurately and ET least accurately, which was related to the size and regularity of the regions.

Table 4: BraTS2021 Full Supervised Result

WT	TC	ET	Mean	Params	GFLOPs
0.9385	0.9114	0.8882	0.9127	10.1 M	2.14

**Figure 6: Region-wise Dice of the full supervised model on BraTS2021**

4.3. Efficiency and Complexity Comparison

The Mean DSC score is 0.91, and the model has 10.1M parameters, 2.14 GFLOPs, the model is significantly smaller in size than the usual 3D networks and transformer hybrids. It can be trained and executed on a single small GPU, solving the issue encompassed by 'accessibility'.

Since there is a certain amount of variation in counts depending on configuration, the comparison values for other methods are only approximate figures available in the literature, and are intended to only indicate order of magnitude differences.

Table 5: Approximate Model Complexity Comparison

Method	Type	Params (approx.)	Relative cost
3D U-Net	3D CNN	≈ 19 M	High
nnU-Net [6]	3D CNN	configuration-dependent	High
TransUNet [13]	2D + Transformer	≈ 105 M	High
Swin UNETR [16]	3D + Transformer	≈ 62 M	High
Proposed (MobileNetV3-UNet)	2D CNN	≈ 10.1 M (2.14 GFLOPs)	Low

4.4. Limited-Label Global Dice

Set to the default settings for all countries. At the global slice level, pseudo-labeling's effect is small: Mean DSC is close to its supervised baseline (0.9072) with 20% SSL and slightly better (0.9081) with 40% SSL. Individual effects are

best seen in the Tumor Core. The weak global impact is due to the fact that the slice level scores are an average of a large number of slices which may contain little or no tumor and therefore improvements in a smaller number of slices are not reflected in the average score.

Table 6: Limited-Label Global Dice Results

Experiment	Labeled patients	Pseudo slices	WT	TC	ET	Mean DSC
Supervised 20%	60	0	0.9137	0.9350	0.8729	0.9072
Pseudo-label SSL 20%	60	2000	0.9115	0.9368	0.8736	0.9073
Supervised 40%	120	0	0.9137	0.9350	0.8729	0.9072
Pseudo-label SSL 40%	120	2000	0.9155	0.9365	0.8724	0.9081

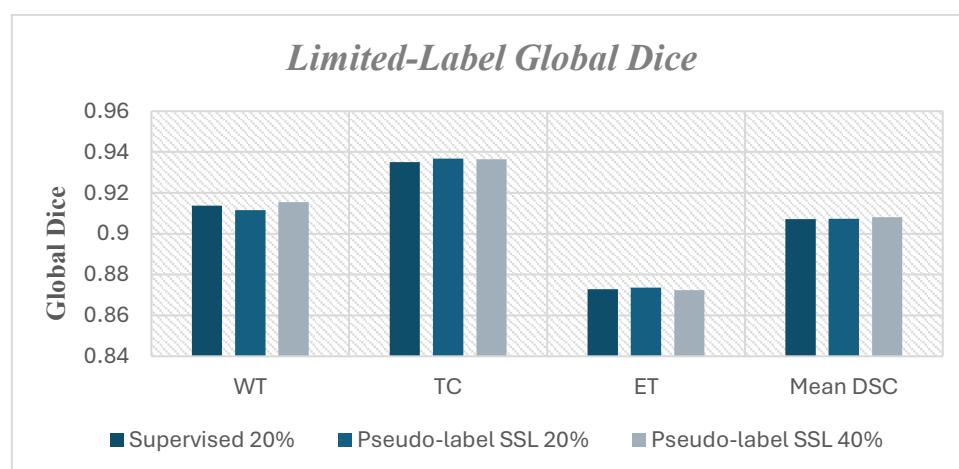


Figure 7: Limited-label global Dice by region: supervised 20% versus SSL 20% and SSL 40%.
Differences are small at the global slice level, with the most visible gains in Tumor Core

4.5. Patient-Wise 3D Dice

The advantage of unlabeled data is accentuated with patient-wise 3D evaluation. The Mean DSC score steadily increases from supervised baseline (0.8862) to SSL 20% (0.8919) and SSL 40% (0.8942) and this improvement is primarily driven by the Tumor Core and Enhancing

Tumor regions. The difference between global and patient-wise results is itself instructive, as verification of the former into pieces first and then into a score rewards spatial consistency that is not captured by slice-wise averaging, in a way that argues for the volumetric evaluation as the more clinically-relevant measure.

Table 7: Patient-Wise 3D Dice Results

Experiment	WT	TC	ET	Mean DSC
Supervised 20%	0.9028	0.9017	0.8541	0.8862
Pseudo-label SSL 20%	0.9007	0.9117	0.8632	0.8919
Pseudo-label SSL 40%	0.9044	0.9135	0.8645	0.8942

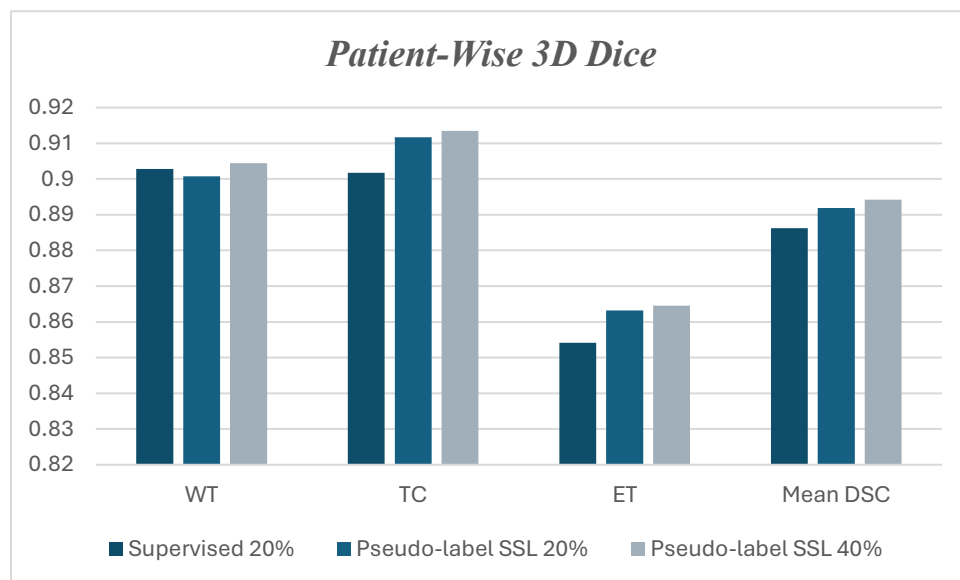


Figure 8: Patient-wise 3D Dice by region for the three settings. Gains concentrate in Tumor Core and Enhancing Tumor, the regions of greatest clinical interest

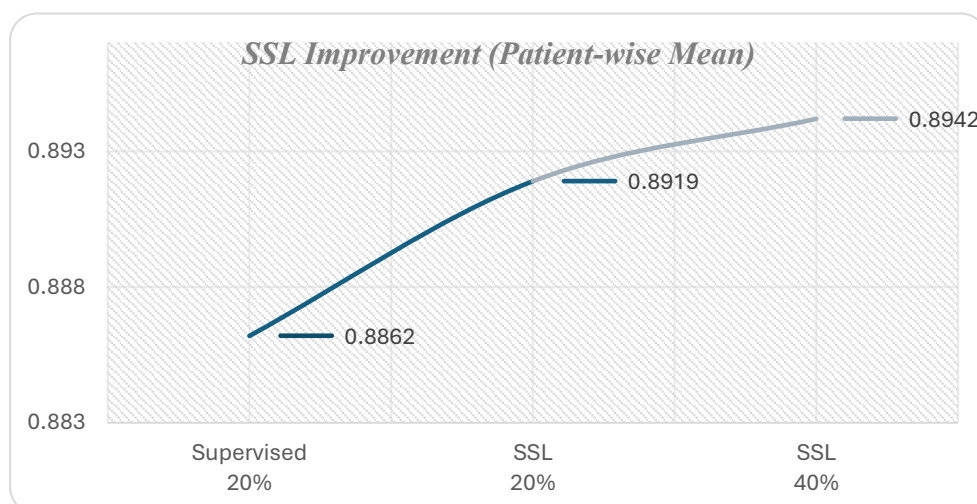


Figure 9: Mean patient-wise 3D Dice rises monotonically as SSL and labels are added

4.7. Region-Wise Analysis

The simple regions are always the best with a score above 0.90 in all settings as this encompasses the largest abnormal region with very distinct contrast on FLAIR and T2; not much room for SSL to improve on this. Patient-wise Dice scores increase from 0.9017 to 0.9135 for Tumor Core (TC) showing the most consistent SSL benefit, which indicates that the pseudo-labeled data provided a different internal structure of tumors that contributes the model learning. The difficulty of enhancing Tumor (ET) highlights its problem space, which can be small and irregular, as well as differ somewhat from patient to patient, but it does get better, from 0.8541 to 0.8645, with SSL, indicating a useful signal also in the unlabeled data.

4.8. How did the Two Evaluation Protocols differ?

In a global dice pool, all test dice are from easy slices, and many of them are not tumor-bearing or even tumor-free; so there is little value in having a model that is already good on easy slices and some improvements on hard slices are diluted. Patient-wise 3D Dice, on the other hand, scores the tumor volumes reconstructed for each of the patients, therefore the improvements in delineating the core and enhancing regions build up on a patient-by-patient basis and can be seen. This variation is not surprising and emphasized the importance of reporting volumetric data in the clinical studies.

4.9. Comparative statements across previous models.

The proposed model does not intend to outperform the leaderboard achievements against very complex 3D models and transformer hybrids [6], [13] and [16] but rather to restore most of the accuracy while keeping the cost relatively low. This measurement (per 10 M parameters, 2.14 GFLOPs) is very competitive against a Mean DSC above 0.91. In addition to adopting efficiency-oriented baselines, this work also utilizes a semi-supervised learning based approach and 3D reporting by patient, which classifies it from the simple supervised lightweight based approach.

4.10. The 2D-versus-3D Trade-off

Operates on two-dimensional slices, loss in inter-slice context with huge saving in memory and computation. The 3D structures are models of volumetric continuity, which tend to generate surfaces of smoother transitions along the slice axis, but have a much higher memory demand for the slice dimension and typically require training in chunks or patches. It is also found that, for the strong patient-wise 3D Dice, obtained here only using a 2D model, most of the volumetric coherence of a 3D model can be recovered with judicious per-modality normalization and patient-wise evaluation. An obvious compromise is to make a 3D formulation using neighboring slices as additional input channels, which is left as future work.

4.11. Computational and Deployment Considerations

Its footprint of 10.1 M parameters and 2.14 GFLOPs allows it to be easily deployed on a single entry-level GPU and makes it feasible to consider deploying it at the edge or on-premise, where data-governance rules may prevent sending scans to an external server. The depthwise separable factorization of Section IV-E results in a relatively low cost for inference as resolution increases, along with a potential further reduction in footprint with post-training quantization and/or pruning that is conventionally expected to have a limited effect on Dice. The setup allows clinics with lower IT power budgets to use this approach; a couple of leading 3D ensembles, such as Maxwell and Octavis, use the multi-GPU setup.

4.12. Statistical Considerations

The small differences reported at the global level (e.g. 0.9072 and 0.9073) are statistics from single training runs and are not meant to be taken as statistically significant, but as indicative. This looks like the improvements by patient are larger and monotone and therefore have more weight, but multiple seeds and a paired significance test by patient would be needed to make it a compelling claim. This is what is addressed by this work in its limitations (Section VI).

4.13. Discussion Summary

The results support the best claims, which are a full supervised Mean DSC of 0.9127 and an improvement in SSL of 0.8942 for each patient. When gathered, they demonstrate the ability of a compact model for accurate and label efficient results. The results also reveal, as stated here mainly, the limited global SSL benefit for low electricity rates and its maximum impact in a volumetric evaluation, and that ET is the natural and limiting region for the future research.

CONCLUSION

To tackle this challenge, this paper introduced a lightweight model with a backbone of MobileNetV3-UNet with a Mean DSC of 0.9127 on BraTS2021, using only 10.1 M parameters and 2.14 GFLOPs. Additionally, pseudo-label SRL was carried out, and the introduction of only 5-10 labels for each patient has increased the patient-wise Mean DSC of the simulated experiments for good segmentation from 0.8862 to 0.8942, demonstrating that even with the use of a compact model, good segmentation is obtainable when only limited labels are used. There are some caveats of this: It suffers from being performed at the 2D slice level, independently from the rest of the volume;

The global SSL gain is fairly low in the 2D slices; The experiment was performed within the context of the BraTS benchmark, and has not yet been external validated using clinical data; Pseudo-labels potentially carry noise after even applying confidence filters. In future work, we plan to expand the semi-supervised ablation, demonstrate the merits of increasing the size of the unlabeled medical pool, validate the model on external clinical test sets, quantify and simplify the model for use in mobile deployment and develop low cost 3D versions that provide more information about the space represented by the image planes.

LIMITATIONS

It is a purposive study with limited in scope. However, two experiments are conducted with two public BraTS data sets and the four common MRI modalities, the whole 3D volume is not used for training the model, the PNaNet only uses confidence-based pseudo-labeling, and the assessment process only tests the data sets in the BraTS benchmark and not with independent clinical data sets. All these limitations make it possible to do this study on one small GPU.

Table 8: Limitations and Future Directions

Limitation	Explanation	Future direction
2D context	Processes slices, not full 3D volumes	2.5D or lightweight 3D extensions
Pseudo-label noise	Wrong teacher masks can mislead the student	Uncertainty filtering or Mean Teacher
Dataset dependence	Based on BraTS benchmark data	Validate on external clinical datasets
Modest global SSL gain	Global Dice improvements are small	Larger SSL ablation with more unlabeled data

REFERENCES

- O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241.
- A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1314–1324.
- B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015, doi: 10.1109/TMI.2014.2377694.
- S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," Sci. Data, vol. 4, art. 170117, 2017.

- U. Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," arXiv:2107.02314, 2021.
- F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.
- J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
- A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NeurIPS*, 2017, pp. 1195–1204.
- D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop Challenges Represent. Learn.*, 2013.
- K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. NeurIPS*, 2020, pp. 596–608.
- O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," in *Proc. Med. Imag. Deep Learn. (MIDL)*, 2018.
- J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.
- I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018, pp. 4510–4520.
- A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. BrainLes, MICCAI Workshops*, 2021, pp. 272–284.
- X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI*, 2021, pp. 8801–8809.
- L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
- B. H. Menze et al., "A generative probabilistic model and discriminative extensions for brain lesion segmentation—with application to tumor and stroke," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 933–946, 2016, doi: 10.1109/TMI.2015.2502596.