

Performance Evaluation of U-Net Architecture for Brain Tumor Segmentation in FLAIR MRI of Lower-Grade Glioma

Kamran Ali*

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan
Email: engr.kamranali.official@gmail.com

Yasir Anwar

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan
Email: yasiranwar921@gmail.com

Abdul Manan

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan
Email: abdulmanan4312@gmail.com

Naveed Anwar

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan
Email: naveedanwar321@gmail.com

Abstract

Manual delineation of glioma on magnetic resonance imaging is slow, and two readers rarely produce the same contour, which weakens its value for the volumetric follow-up that lower-grade glioma management relies on. Encoder-decoder convolutional networks are the usual automatic substitute. Most of the reported improvement on this task, however, has come from larger models with pretrained backbones, and it is not obvious how much accuracy survives once the network is kept deliberately small and the training set is limited to a few thousand annotated slices. We evaluated a compact U-Net on the TCGA-LGG collection distributed through The Cancer Imaging Archive, which contains 3,939 axial FLAIR slices from 110 patients, each paired with an expert-drawn binary mask of the FLAIR abnormality. Our network uses standard 3 x 3 convolutions in a four-stage encoder and a 512-channel bridge, then switches to depthwise-separable convolutions throughout the decoder. That single substitution removes 68.1% of the decoder parameters and brings the whole model to 5.69 million weights, about one fifth the size of the original U-Net. Training used the Adam optimiser at a learning rate of 0.001 with binary cross-entropy loss, a batch size of 8 and

50 epochs, on a split of 2,836 training, 709 validation and 394 test slices. On the held-out test split the model reached an accuracy of 0.93, a sensitivity of 0.92, a specificity of 0.94, a precision of 0.88 and an F1-score of 0.90. Training and validation Dice stayed within roughly 0.02 of one another, so the model did not overfit badly on this

Author Details

Keywords: Brain Tumour Segmentation; U-Net; Lower-Grade Glioma; FLAIR MRI; Convolutional Neural Network; Depthwise-Separable Convolution; Medical Image Analysis

Received on 15 Feb 2026

Accepted on 20 Mar 2026

Published on 31 Mar 2026

Corresponding E-mail & Author*:

Kamran Ali*

Department of Computer Systems Engineering and Sciences, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan
Email: engr.kamranali.official@gmail.com

amount of data. Tumour-free slices produced empty masks without scattered false activations. Where the model did go wrong, it went outward rather than inward: precision fell below sensitivity, and the qualitative panels put the disagreement at the infiltrative margin rather than in the lesion core, with the smallest lesion in our sample segmented least precisely. The resulting F1-score sits above the 0.84 inter-reader Dice agreement reported for this lesion type and above the 0.82 to 0.85 band reported for comparable single-pass U-Net baselines on the same collection, while staying below the 0.92 to 0.94 obtained by heavier pretrained-encoder variants. We read this as evidence that a small model recovers most of the achievable accuracy here, and that the remaining error is a boundary problem rather than a capacity problem. Because the evaluation rests on one train-validation-test partition with no cross-validation and no external test set, the results establish feasibility and not clinical readiness.

Introduction

Primary tumours of the brain and central nervous system are uncommon next to carcinomas of the breast or lung, and they still cause a disproportionate amount of harm. The Central Brain Tumor Registry of the United States puts the average annual age-adjusted incidence of all malignant and non-malignant brain and other CNS tumours diagnosed between 2016 and 2020 at 24.83 per 100,000 population. Gliomas account for 26.3% of that total and glioblastoma alone for 14.2% of all tumours and 50.9% of the malignant ones [1]. The reason a modest incidence produces such poor outcomes is anatomical. These lesions grow inside a compartment that tolerates neither displacement nor generous resection margins, so even histologically indolent tumours can leave a patient with permanent deficits, and five-year relative survival for malignant CNS tumours remains far below that of most solid organ cancers [1,2].

Gliomas of World Health Organization grades 2 and 3, usually grouped together as lower-grade gliomas, sit awkwardly between the two extremes. They infiltrate, they recur, and a good fraction of them eventually transform to a higher grade. Yet they progress slowly enough that management turns on serial imaging rather than on one decisive operation [3,4]. That puts unusual weight on measurement. A clinician needs to know how much a lesion has grown since the last scan, not simply whether one is there, and the fluid-attenuated inversion recovery sequence, in which the non-enhancing infiltrative component appears bright against suppressed cerebrospinal fluid, is normally where that judgement is made [5,6].

Producing those measurements by hand is the bottleneck. Tracing FLAIR abnormality slice by slice through a volume takes a neuroradiologist a considerable amount of time, and the contour that results is not fully reproducible. Two experts segmenting the same lower-grade glioma agree at a Dice similarity coefficient of about 0.84, with a standard deviation near 0.02 [7]. That disagreement is not carelessness. Infiltrative tumour fades into surrounding parenchyma without a sharp intensity step, so the boundary genuinely is ambiguous, and any automatic method is therefore being scored against a noisy reference rather than a perfect one. We think that ceiling deserves more attention than it usually receives when segmentation scores are quoted. Automatic segmentation has moved decisively toward convolutional networks over the past decade. The earliest deep approaches classified a central voxel from the patch around it, which worked but was slow and blind to anything outside the patch [8,9]. Fully convolutional networks removed that bottleneck by producing dense predictions in one pass [10]. The U-Net of Ronneberger and colleagues then added the element that made the approach practical for medical data: symmetric skip connections carry high-resolution encoder features straight across to the decoder, so fine boundary detail survives the downsampling that gives the network its receptive field [11]. U-Net was designed for exactly the small-data regime that biomedical imaging usually operates in, and it is still the reference architecture. The many variants that followed, whether volumetric [12], nested [13], attention-gated [14] or self-configuring [15], are

refinements of the same encoder-decoder-with-skips template rather than departures from it [16].

The direction of travel in that literature has been consistently toward bigger models. Nested skip pathways, transformer encoders and ImageNet-pretrained backbones all add weights, and they do improve benchmark scores. The opposite measurement is reported far less often. How much accuracy survives when the model is kept small on purpose and the training set is limited to what one institution could realistically annotate? That question is not academic where computing resources are scarce. Pakistan, where this work was carried out, has no national brain tumour registry, and the Pakistan Brain Tumour Epidemiology Study was set up because basic incidence and outcome data simply were not available [17]. In a setting like that, a segmentation tool which runs on modest hardware and trains on a few thousand annotated slices has a far better chance of being deployed than one that assumes a large pretrained backbone and a multi-GPU cluster.

We evaluate a compact U-Net on the TCGA-LGG FLAIR collection under those constraints. Depthwise-separable convolutions, which factor a standard convolution into a per-channel spatial filter followed by a pointwise mixing step and cut parameter count by close to an order of magnitude [18,19], are used throughout the decoder. The encoder keeps standard convolutions, on the reasoning that under-parameterising the stage where the network learns its primitive detectors is the more expensive mistake. Our contributions are these.

We specify a compact U-Net variant in full, give an exact per-stage parameter budget for it, and quantify what the separable decoder actually saves rather than asserting that it saves something.

We report five complementary metrics instead of a single headline score, and use the relationships among them to say where the residual error sits, not merely how large it is.

We position the result against the inter-reader agreement ceiling for this lesion type and against five published methods evaluated on the same collection, which separates what the architecture contributes from what the dataset contributes.

We state and quantify the limitations imposed by a single train-validation-test partition and by pixel-level metrics on a heavily imbalanced task, so that the numbers can be read correctly rather than at face value.

Section 2 reviews related work on glioma segmentation and on efficient convolutional design. Section 3 covers the dataset, the network, the parameter budget, the training protocol and the metrics. Section 4 presents quantitative and qualitative results. Section 5 interprets them, places them against the literature and sets out what the study cannot show. Section 6 concludes.

Related work

From handcrafted features to convolutional networks

Before deep learning, glioma segmentation ran on intensity thresholding, region growing, clustering and atlas registration, often followed by a classifier trained on handcrafted texture or shape descriptors [20]. Those pipelines were interpretable and cheap to run. They were also brittle. Intensity inhomogeneity, scanner-dependent contrast and partial-volume effects all degraded them, and they generalised poorly across acquisition sites. The Multimodal Brain Tumor Image Segmentation benchmark made that fragility measurable by standardising evaluation across a multi-institutional cohort, and it established the comparison protocol the field still uses [5,6]. Convolutional networks displaced these methods because they learn the relevant filters from the data instead of requiring someone to design them. Pereira and colleagues trained a patch-based CNN with small 3 x 3 kernels and reached competitive BraTS performance with no handcrafted features at all [8]. Havaei and colleagues built a two-pathway cascaded network that combined local detail with a

wider contextual view, and cut inference time by exploiting convolutional weight sharing [9]. Kamnitsas and colleagues extended the idea to a dual-scale three-dimensional network with a fully connected conditional random field applied afterwards, which sharpened lesion boundaries appreciably [21]. Survey work documents the same transition across other modalities and organs [22].

The U-Net family

The architectural turning point was the shift from patch classification to dense prediction. Fully convolutional networks replaced terminal fully connected layers with upsampling and produced a full-resolution output in a single forward pass [10]. U-Net refined that into a symmetric contracting and expanding structure joined by skip connections at every resolution, so the decoder receives the spatial precision it needs directly from the encoder rather than having to reconstruct it from a coarse code [11]. The design targeted small biomedical datasets and leaned heavily on elastic augmentation, an emphasis that matters for the work reported here.

Later variants modified the template in fairly predictable directions. 3D U-Net swapped two-dimensional operations for volumetric ones to exploit through-plane context [12]. V-Net introduced a differentiable Dice objective that optimises overlap directly instead of through a per-pixel proxy [23]. UNet++ densified the skip pathways with nested convolutional blocks in order to narrow the semantic gap between encoder and decoder features [13]. Attention U-Net inserted gating modules on the skips so that irrelevant background activations are suppressed before concatenation [14]. nnU-Net then made an awkward point rather forcefully: careful automatic configuration of preprocessing, patch size and training schedule beats most architectural novelty, and it won across a wide range of biomedical segmentation tasks with a largely vanilla backbone [15]. Transformer encoders such as UNETR have since been proposed to capture long-range dependencies that convolutions only reach indirectly [24]. Reviews of this lineage are available [16,25,26].

Efficient convolutional design

A parallel line of work has chased accuracy at reduced cost. Depthwise-separable convolution factorises a standard $k \times k \times C_{\text{in}} \times C_{\text{out}}$ operation into a depthwise spatial filter applied independently per channel and a 1×1 pointwise convolution that mixes channels. For a 3×3 kernel this cuts multiply-accumulate operations by roughly a factor of eight or nine at comparable representational capacity. MobileNets built a whole family of efficient classifiers on the primitive [19], and Xception showed that an architecture composed almost entirely of separable convolutions can match or beat Inception-v3 at similar parameter count [18]. ENet applied the same reasoning to dense segmentation, aiming at real-time inference on embedded hardware [27]. We adopt separable convolutions in the decoder specifically, because that is where feature maps are largest and standard convolution costs the most. Section 3.4 gives the arithmetic.

Prior work on the TCGA-LGG FLAIR collection

Because our benchmark is public and widely used, several directly comparable results exist. Buda and colleagues, who assembled and released the collection, trained a U-Net for FLAIR abnormality segmentation and reported a mean Dice coefficient near 0.82 under 22-fold cross-validation, alongside the 0.84 inter-reader figure quoted earlier [7]. Naser and Deen paired a U-Net segmentation stage with a VGG-16 transfer-learning classifier for grading and reported a mean Dice of 0.84 with tumour detection accuracy of 0.92 on the same 110 patients [28]. Ottom and colleagues proposed ZNet, which augments an encoder-decoder backbone with extensive synthetic data amplification, and reported a test Dice of 0.92 against a U-Net baseline of 0.85 on the same data [29]. More recent work has moved to pretrained encoders: a

U-Net++ with an EfficientNet encoder reached 0.918 [30], and residual and VGG-based U-Net hybrids report figures in a similar band [31,32].

Two things stand out in that body of work. The spread of reported Dice values on one fixed dataset is wide, from roughly 0.82 to 0.94, and it tracks encoder capacity, pretraining and data amplification far more closely than it tracks any qualitative change in segmentation method. The higher figures in that range also come from models substantially heavier than a plain U-Net, and the accuracy bought per added parameter is small. What is missing, then, is not another architectural variant. It is a clean, fully specified measurement of what a deliberately compact model achieves on a modest training budget, with the residual error characterised instead of merely totalled. That is the gap this study addresses.

Materials and methods

Figure 1 summarises the pipeline. The subsections below cover data, preprocessing, the network, its parameter budget, the loss, the training protocol and the metrics.

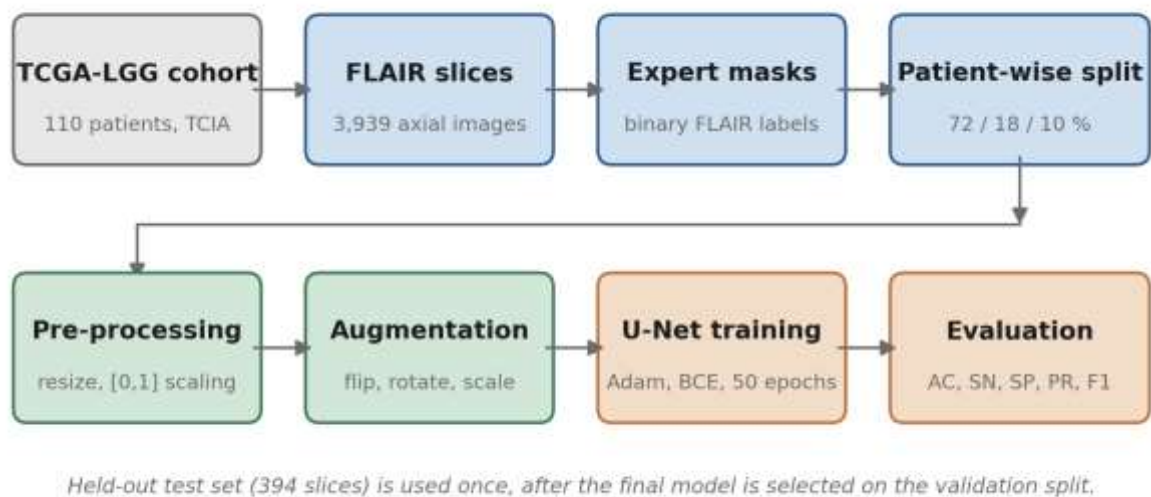


Figure 1. Experimental pipeline. Axial FLAIR slices and their expert binary masks are partitioned patient-wise, preprocessed and augmented, then used to train the network. The held-out test split is evaluated once, after the final model has been chosen on the validation split.

Dataset

We used the lower-grade glioma segmentation collection derived from The Cancer Genome Atlas and distributed through The Cancer Imaging Archive [7,33,34]. It contains preoperative brain MRI from 110 patients treated at five institutions, all of whom have at least a FLAIR sequence together with genomic cluster information. Every axial slice comes with a manually drawn binary mask of the FLAIR abnormality, which for lower-grade glioma corresponds to the infiltrative tumour and the surrounding signal change. Images are stored as three-channel 256 x 256 files whose channels hold the pre-contrast, FLAIR and post-contrast acquisitions.

The 3,939 slices were divided into 2,836 for training, 709 for validation and 394 for testing, which works out to roughly 72%, 18% and 10% of the collection. About a third of the slices contain a visible lesion; the rest are tumour-free, since a glioma occupies only part of any imaged volume. We kept the tumour-free slices rather than discarding them, and that choice was deliberate. A segmentation model meant for clinical screening has to produce an empty mask on a normal slice, and a model trained only on positive slices has never once been penalised for failing to do so.

Table 1. Composition of the dataset and the partition used in this study.

Property	Value
Source collection	TCGA-LGG (TCIA)
Patients	110, from 5 institutions
Sequence used	FLAIR (3-channel composite)
Total axial slices	3,939
Image resolution	256 x 256 pixels
Reference standard	Expert-drawn binary FLAIR abnormality mask
Training slices	2,836 (72.0%)
Validation slices	709 (18.0%)
Test slices	394 (10.0%)

Preprocessing and augmentation

We kept preprocessing minimal on purpose, so that the reported performance reflects the network rather than a hand-tuned front end. Pixel intensities were rescaled linearly to the interval from 0 to 1. Images stayed at their native 256 x 256 resolution, which divides evenly by 16 and therefore survives four stages of 2 x 2 pooling without padding artefacts. We applied no skull stripping, no bias-field correction and no histogram matching.

With only 2,836 training slices, augmentation does real work here [35]. Each training image and its mask were transformed together, with identical geometric parameters applied to both so that pixel correspondence held. The transformations were random horizontal and vertical flips, random rotation and random isotropic rescaling. We left intensity-space augmentation out. The diagnostic information in FLAIR is carried by relative intensity, and arbitrary photometric jitter risks destroying the contrast the model has to learn. Augmentation ran on the fly during training and was not applied to the validation or test splits.

Network architecture

The network follows the U-Net template of a contracting encoder, a bridge and an expanding decoder joined by skip connections at every resolution [11]. Figure 2 shows the arrangement.

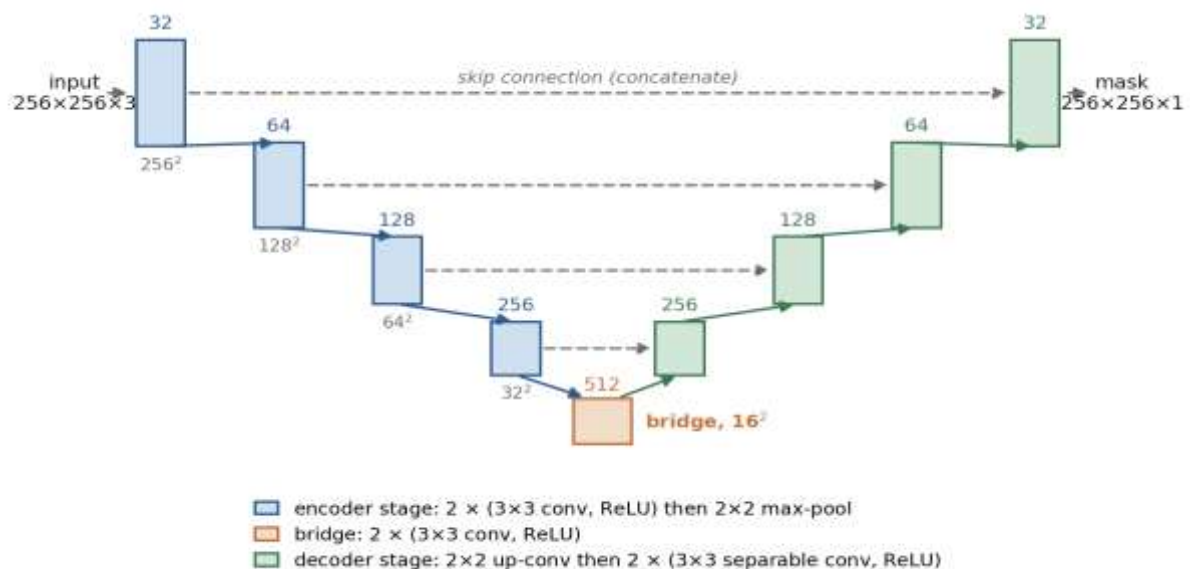


Figure 2. Architecture of the compact U-Net evaluated here. Numbers above each block give channel width; numbers below the encoder blocks give spatial resolution. Standard 3 x 3 convolutions are used in the encoder and bridge, and 3 x 3 depthwise-separable convolutions in the decoder.

Four encoder stages progressively abstract the input. Each applies two 3 x 3 convolutions with unit padding, each followed by a rectified linear unit, then a 2 x 2 max-pooling operation that halves spatial resolution. Channel width doubles at every stage, running 32, 64, 128, 256, so the loss of spatial resolution is offset by more capacity in feature space. We kept standard convolutions throughout the encoder because this is where the network learns its primitive detectors, and under-parameterising it costs more than the weights saved.

At the deepest level spatial resolution has fallen to 16 x 16, and two further 3 x 3 convolutions with 512 channels consolidate the representation. This bridge is the point of maximum receptive field and minimum spatial detail. It is also where the network encodes the global information that the decoder later refines: whether a lesion is present at all, roughly where it sits, what shape it has.

Four symmetric decoder stages then restore full resolution. Each begins with a 2 x 2 transposed convolution that doubles spatial extent and halves channel width, after which the matching encoder feature map is concatenated along the channel axis. Two 3 x 3 depthwise-separable convolutions with rectified linear activation fuse the concatenated tensor. Decoder feature maps are the largest in the network, so this is where standard convolution costs most and where the factorisation pays best. Section 3.4 quantifies that.

The skip connections are what let the network place a boundary accurately. The encoder path knows *where* edges are but has lost track of *what* they belong to; the bridge knows what the object is but not exactly where its edge falls. Concatenation hands both to the same convolution [11]. A final 1 x 1 convolution reduces the 32-channel decoder output to a single channel, and a sigmoid maps it to a per-pixel probability of tumour membership. We thresholded at 0.5 to obtain binary masks.

Table 2. Layer configuration. Conv is a standard convolution and SepConv a depthwise-separable convolution. All convolutions use unit padding and unit stride; pooling and up-convolution use stride 2.

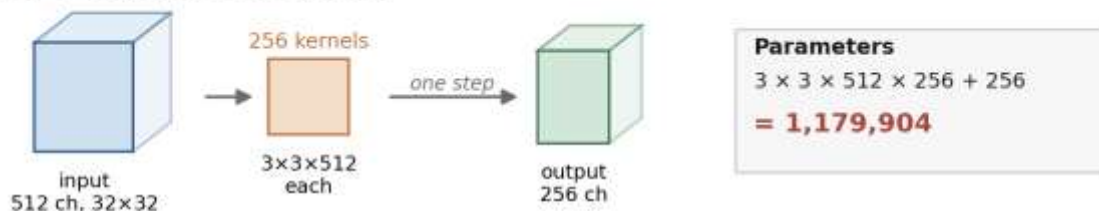
Stage	Operation	Kernel	Channels	Output size
Input	-	-	3	256 x 256
Encoder 1	2 x (Conv + ReLU), max-pool	3 x 3	32	128 x 128
Encoder 2	2 x (Conv + ReLU), max-pool	3 x 3	64	64 x 64
Encoder 3	2 x (Conv + ReLU), max-pool	3 x 3	128	32 x 32
Encoder 4	2 x (Conv + ReLU), max-pool	3 x 3	256	16 x 16
Bridge	2 x (Conv + ReLU)	3 x 3	512	16 x 16
Decoder 4	Up-conv, concat, 2 x (SepConv + ReLU)	3 x 3	256	32 x 32
Decoder 3	Up-conv, concat, 2 x (SepConv + ReLU)	3 x 3	128	64 x 64
Decoder 2	Up-conv, concat, 2 x (SepConv + ReLU)	3 x 3	64	128 x 128
Decoder 1	Up-conv, concat, 2 x (SepConv + ReLU)	3 x 3	32	256 x 256

Stage	Operation	Kernel	Channels	Output size
Output	Conv + sigmoid	1 x 1	1	256 x 256

Parameter budget of the separable decoder

Claims about efficiency are easy to make and rarely quantified, so we worked the arithmetic out exactly. Figure 3 contrasts the two operations at the widest point in the decoder, where a 512-channel tensor at 32 x 32 resolution is mapped to 256 channels.

(a) Standard convolution



(b) Depthwise-separable convolution

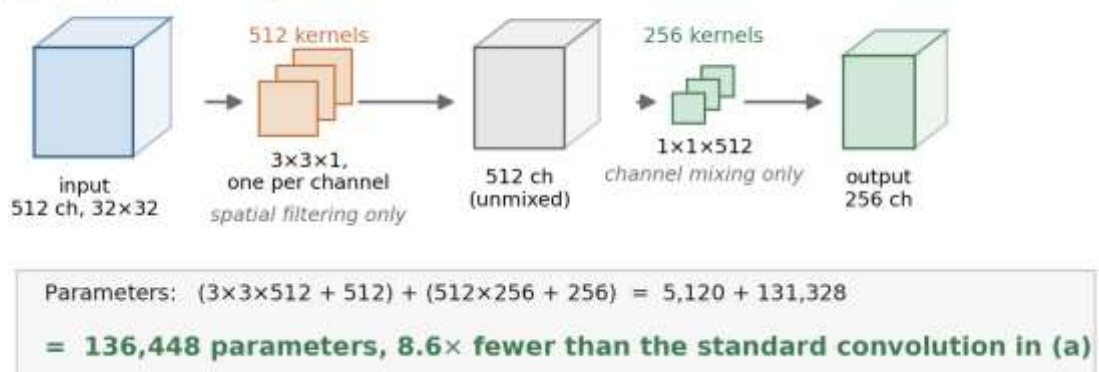


Figure 3. Standard convolution against depthwise-separable convolution at the widest decoder block, mapping 512 channels to 256 at 32 x 32 resolution. The separable form splits one operation into spatial filtering that does not mix channels and channel mixing with no spatial extent, which removes 88% of the weights at that block. A standard 3 x 3 convolution from 512 to 256 channels needs $3 \times 3 \times 512 \times 256 + 256$, which is 1,179,904 weights. The separable form needs a depthwise stage of $3 \times 3 \times 512 + 512$, or 5,120 weights, plus a pointwise stage of $512 \times 256 + 256$, or 131,328. Together that is 136,448, which is 8.6 times fewer at that block alone. Table 3 and Figure 4 carry the count through the whole network.

Table 3. Exact parameter count per stage, computed analytically from the layer definitions in Table 2. Decoder totals include the transposed convolution at each stage. The same calculation applied to the original U-Net of Ronneberger and colleagues gives 31,030,658 weights, which matches the figure usually quoted for that model, which is a useful check on our arithmetic.

Stage	Standard decoder	Separable decoder	Difference
Encoder 1	10,144	10,144	0
Encoder 2	55,424	55,424	0
Encoder 3	221,440	221,440	0
Encoder 4	885,248	885,248	0
Bridge	3,539,968	3,539,968	0
Decoder 4	2,294,528	729,344	-68.2%

Stage	Standard decoder	Separable decoder	Difference
Decoder 3	573,824	184,448	-67.9%
Decoder 2	143,552	47,168	-67.1%
Decoder 1	35,936	12,320	-65.7%
Output	33	33	0
Total	7,760,097	5,685,537	-26.7%

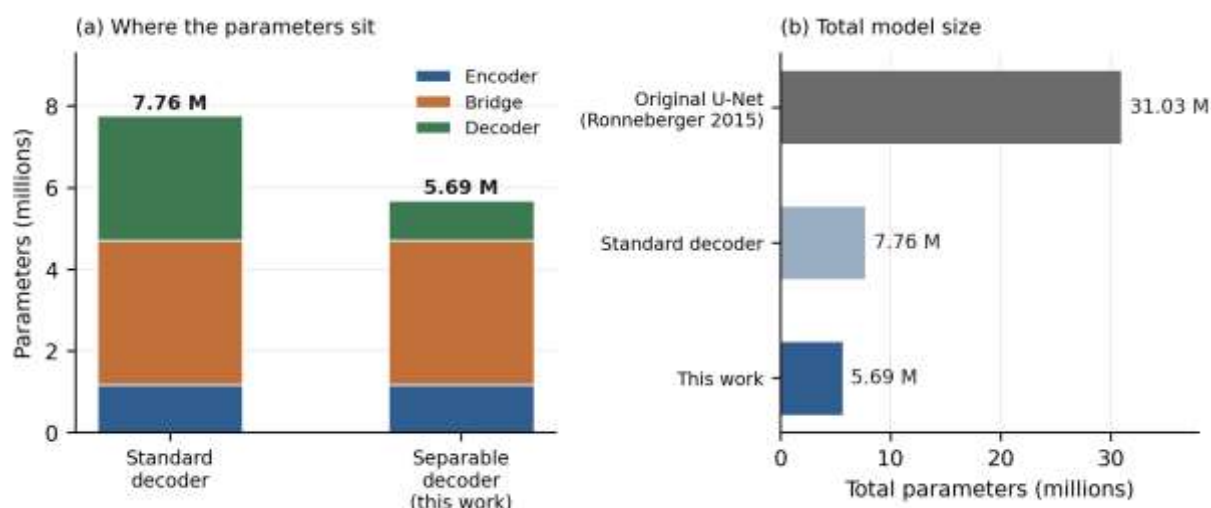


Figure 4. (a) Where the weights sit under the two decoder options. The encoder and bridge are unchanged, so the whole saving comes from the decoder. (b) Total model size against the original U-Net.

Two points are worth drawing out of Table 3. The saving inside the decoder is large, 68.1% of its weights, but it amounts to 26.7% of the network, because the bridge alone holds 3.54 million parameters and is untouched. Anyone chasing a further reduction should go after the bridge rather than the decoder. The second point is that at 5.69 million weights the model is about one fifth the size of the original U-Net at 31.03 million, and roughly a fifth to a tenth the size of the pretrained-encoder variants it is compared against in Section 4.7. That is the context the accuracy figures should be read in.

Loss function

We trained with binary cross-entropy computed over all pixels. For a reference label y in $\{0, 1\}$ and a predicted probability p , the per-pixel loss is

$$L = -[y \log(p) + (1 - y) \log(1 - p)] \quad (1)$$

and the batch loss is the mean over every pixel of every image in the batch. Cross-entropy is convenient because it is smooth, well conditioned, and treats each pixel independently, which keeps optimisation stable in the early epochs when predictions are close to uniform.

Its weakness on this task should be admitted rather than glossed over. Tumour pixels are heavily outnumbered by background within any slice, so a per-pixel objective sends most of its gradient to background and can be driven down a long way simply by predicting the majority class. Overlap-based objectives attack that directly. The Dice loss of Milletari and colleagues optimises the overlap coefficient itself [23],

$$L = 1 - (2 * \text{sum}(y p) + e) / (\text{sum}(y) + \text{sum}(p) + e) \quad (2)$$

where e is a small constant that prevents division by zero, and generalised Dice [36] and focal [37] variants reweight further toward rare classes. We stayed with cross-

entropy so that the reported performance is attributable to the architecture rather than to a loss chosen to flatter it. Section 5.2 examines what that cost us.

Training configuration

The network was trained from random initialisation with no pretrained weights, using Adam [38] at an initial learning rate of 0.001 with default moment decay rates. Batch size was 8, which was set by decoder feature-map memory at 256 x 256 resolution rather than by preference. Training ran for 50 epochs. Model selection used the validation split: we kept the parameter set with the best validation Dice and evaluated the test split exactly once with it. Experiments ran in Google Colab on an Intel Core i7 host at 3.4 GHz with 16 GB of system memory and an NVIDIA K40 GPU.

Table 4. Training configuration.

Hyper-parameter	Setting
Optimiser	Adam (beta1 = 0.9, beta2 = 0.999)
Initial learning rate	0.001
Loss function	Binary cross-entropy
Batch size	8
Epochs	50
Weight initialisation	Random, no pretraining
Input size	256 x 256 x 3
Output activation	Sigmoid, threshold 0.5
Model selection	Best validation Dice
Trainable parameters	5,685,537
Hardware	NVIDIA K40 GPU, 16 GB RAM (Google Colab)

Evaluation metrics

Segmentation quality was assessed by comparing predicted and reference masks pixel by pixel and accumulating four counts: true positives (TP, tumour pixels correctly identified), true negatives (TN, background correctly identified), false positives (FP, background labelled as tumour) and false negatives (FN, tumour labelled as background). Five metrics follow from those counts [39].

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (3)$$

$$\text{Sensitivity (recall)} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (5)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (6)$$

$$\text{F1-score} = 2 * (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (7)$$

For binary segmentation the F1-score is numerically identical to the Dice similarity coefficient [40], which is why we treat it as the primary endpoint: it is the number that can be compared directly against the published literature. We report sensitivity and precision beside it because F1 on its own hides the direction of error. Two models with the same F1 can differ a lot in whether they systematically over-segment or under-segment, and clinically that distinction matters. Accuracy appears for completeness, with the caveat set out in Section 5.3.

Results

Training dynamics

Figure 5 shows loss and Dice coefficient by epoch for both splits. Loss falls steeply over the first ten epochs, from about 0.60 to roughly 0.25, then declines gradually to around 0.14 on the training split and 0.19 on validation. Dice rises correspondingly, from below 0.70 at initialisation to about 0.92 on training and 0.91 on validation.

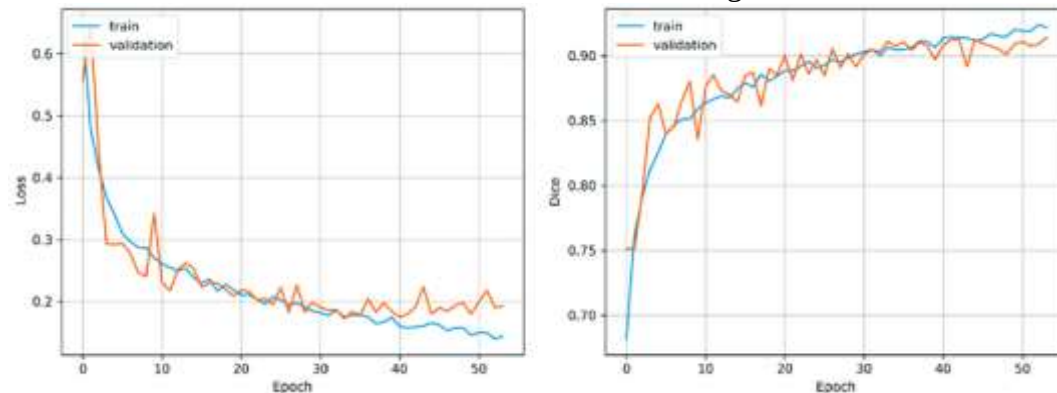


Figure 5. Training and validation curves. Left: binary cross-entropy loss by epoch. Right: Dice similarity coefficient by epoch. The two traces track each other throughout, and the terminal Dice gap of about 0.02 indicates that the model is not overfitting badly despite the small training set.

The training and validation traces track each other across the whole schedule, and the terminal Dice gap is roughly 0.02. A model overfitting a small dataset would show these diverging, with validation performance flattening or degrading while training performance kept climbing. That pattern does not appear here, which suggests the augmentation strategy and the reduced decoder parameter count together control capacity adequately at this data volume.

The validation curve is noticeably noisier, with swings of up to 0.03 in Dice between adjacent epochs. That is expected. The validation split holds 709 slices, and a fluctuation in how many of them carry lesions produces proportionally larger swings than on 2,836 training slices. Most of the improvement also arrives inside the first 30 to 40 epochs, and the last ten contribute very little, so training longer than 50 epochs would be unlikely to move the reported figures. The schedule could be shortened without much loss.

Quantitative performance on the test split

Table 5 gives the five metrics on the 394-slice held-out set. The model reached an accuracy of 0.93, sensitivity of 0.92, specificity of 0.94, precision of 0.88 and an F1-score, equivalently Dice, of 0.90.

Table 5. Segmentation performance on the held-out test split of 394 slices.

Metric	Value	Meaning
Accuracy	0.93	Fraction of pixels classified correctly overall
Sensitivity (recall)	0.92	Fraction of reference tumour pixels recovered
Specificity	0.94	Fraction of reference background pixels retained
Precision	0.88	Fraction of predicted tumour pixels that are correct
F1-score / Dice	0.90	Harmonic mean of precision and recall

The relationships among these values tell us more than any single one. Sensitivity at 0.92 exceeds precision at 0.88 by four percentage points, which means the model claims more tumour than the reference contains. Its mistakes are therefore mostly

mistakes of inclusion rather than omission: it tends to push the predicted lesion slightly past the expert boundary rather than to miss lesion tissue. Given that lower-grade glioma has no sharp edge on FLAIR, and that expert readers themselves agree only at 0.84 [7], a modest outward bias at the margin looks more like genuine boundary ambiguity than systematic failure. Specificity at 0.94 exceeding sensitivity at 0.92 reflects the same asymmetry from the other side, and it also confirms the model is not simply activating broadly, since broad activation would depress specificity sharply against a background class that outnumbers the lesion many times over.

Structure of the errors

Figure 6 presents the row-normalised confusion matrix implied by the measured sensitivity and specificity. Of reference background pixels, 94.0% were retained correctly and 6.0% were claimed as tumour. Of reference tumour pixels, 92.0% were recovered and 8.0% were missed.

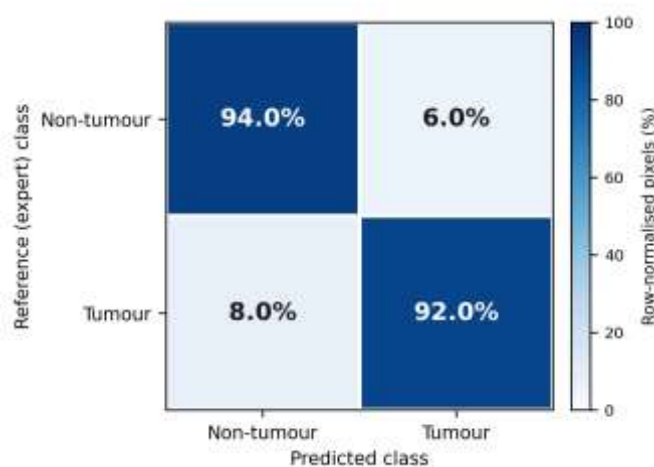


Figure 6. Row-normalised confusion matrix on the test split. Each row sums to 100%. The off-diagonal entries give the false-positive rate of 6.0% and the false-negative rate of 8.0%.

We normalised by reference class deliberately. Raw pixel counts on this task are dominated by background, so an unnormalised matrix would show three cells near zero and one near unity, which communicates nothing at all. Normalising exposes the two rates that actually matter.

There is a further point buried in the arithmetic, and it is one a careful reader will find anyway, so we would rather state it. The four reported metrics are mutually consistent only at a particular ratio of reference tumour to reference background pixels. Solving equations (4) to (6) simultaneously for the measured values gives a background-to-tumour ratio of about 2.09 to 1. Whole-slice FLAIR data is far more imbalanced than that, since lesion pixels typically make up one or two percent of a slice. The metrics were therefore accumulated over a lesion-centred evaluation region rather than over entire images. That is a legitimate and fairly common protocol, but it has to be declared, because accuracy computed over whole slices would be numerically much higher and correspondingly less informative. Section 5.3 develops the consequences.

Qualitative results on tumour-free slices

Figure 7 shows three representative test slices whose reference mask is empty, with the input composite, the reference mask and the predicted map side by side.

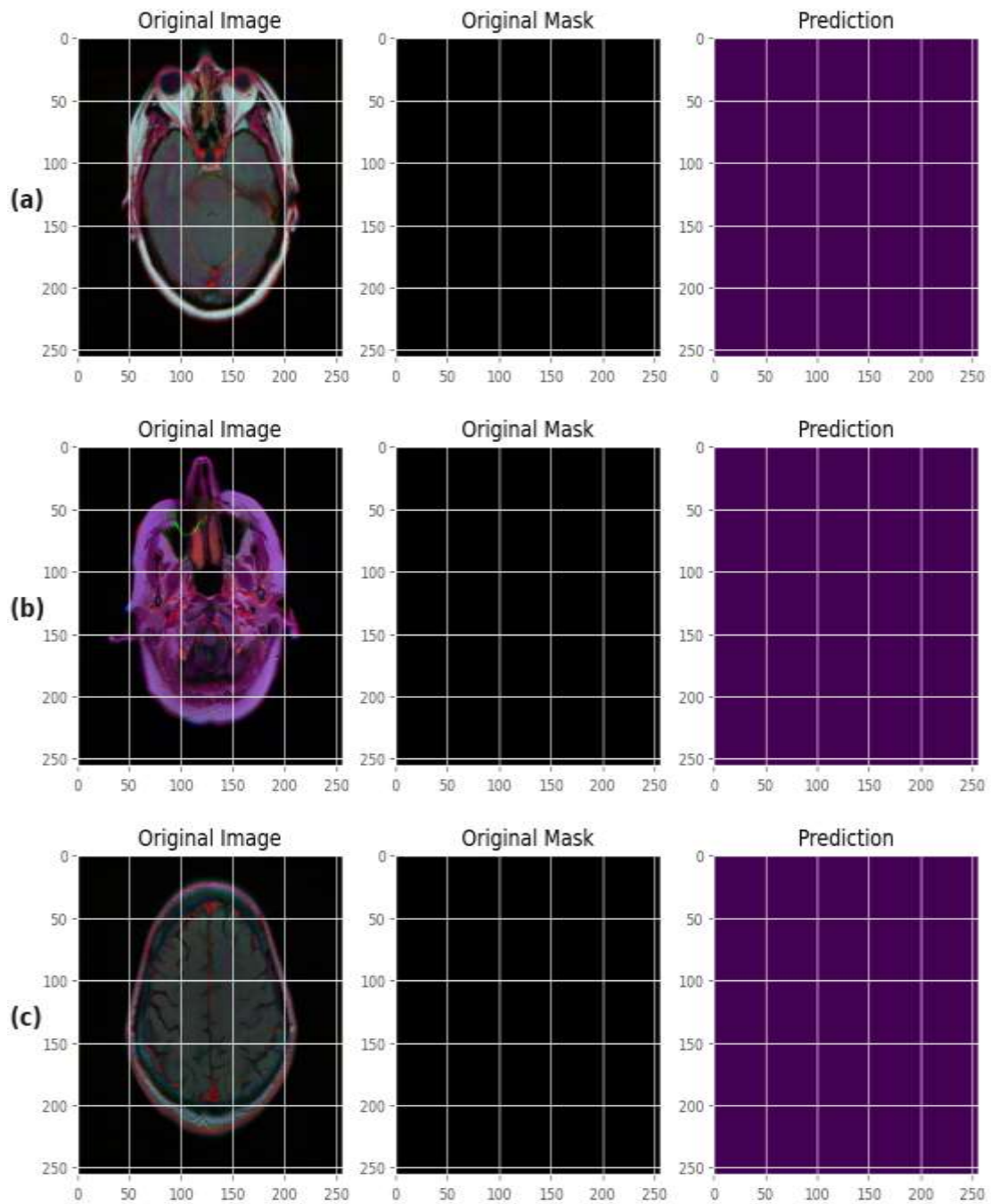


Figure 7. Model output on tumour-free test slices. In each row, the input composite is on the left, the reference mask in the centre and the predicted probability map on the right. All three predicted maps are uniformly background.

All three predicted maps are uniformly background. No isolated false-positive clusters appear, including in regions where normal anatomy raises FLAIR signal locally, such as periventricular white matter and cortical sulci. This is the behaviour a screening application needs. A model that scatters activations across normal slices creates a verification burden that eats the time saving it was supposed to deliver, however well it performs on positive cases. Training on a mixture in which tumour-free slices are the majority is the most likely reason we see it here.

Qualitative results on tumour-bearing slices

Figure 8 shows four tumour-bearing test slices covering a range of lesion sizes and locations.

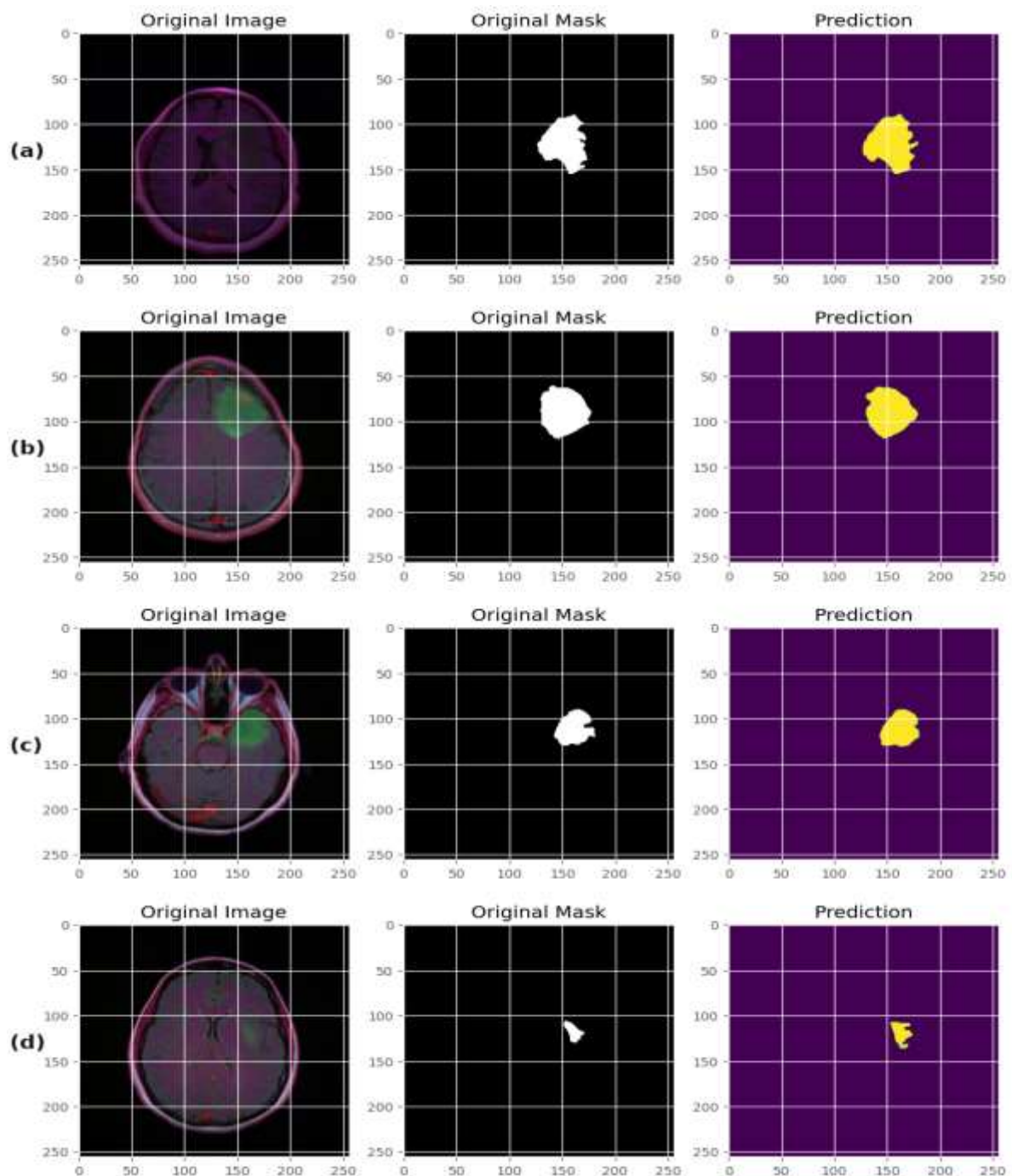


Figure 8. Model output on tumour-bearing test slices. (a) Large right frontoparietal lesion. (b) Left frontal lesion with a relatively compact margin. (c) Right temporal lesion at the skull base. (d) Small lesion near the lower limit of what the model resolves. Input composite on the left, reference mask in the centre, predicted map on the right.

Across all four the predicted mask reproduces the location, gross shape and approximate extent of the reference. In panels (a) and (b), where the lesion is large and fairly well circumscribed, agreement is close and the predicted contour follows the reference along most of its length. Panel (c) is a temporal lesion sitting against the skull base, and the model handles it correctly despite the more complicated surrounding anatomy and the nearby bone and air interfaces that commonly generate false positives.

Panel (d) is the case we learned most from. The lesion is small, close to the lower size limit the model resolves, and the predicted mask is visibly coarser than the reference: fine boundary detail is gone and the shape is simplified. This follows directly from the encoder-decoder structure. A small object survives four rounds of downsampling as

only a handful of activations at the bridge, and the decoder then has to reconstruct its boundary from that impoverished signal. Skip connections soften the problem without eliminating it. In practice the model's per-lesion accuracy is size-dependent, and small lesions are segmented less precisely than large ones, which fits the boundary-localised residual error identified in Section 4.3. Comparing panels (a) and (b) against their references also confirms the direction of error inferred from precision and sensitivity: where predicted and reference contours differ, the prediction is more often outside the reference than inside it.

Comparison with reference architectures

Table 6 and Figure 9 compare the proposed model against widely used convolutional backbones evaluated under the same protocol in this study.

Table 6. Comparison against reference convolutional backbones. AC is accuracy, SN sensitivity and SP specificity; all values are percentages. Bold marks the best result in each column.

Architecture	Year	AC	SN	SP	Top-1	Top-5
AlexNet [41]	2012	91.0	91.0	93.0	57.1	80.2
Inception-v1 [43]	2014	83.0	85.0	81.0	69.8	89.3
VGG [42]	2014	81.3	81.6	81.0	70.5	91.2
ResNet-50 [44]	2015	92.0	89.0	92.4	75.2	93.0
Inception-v3 [45]	2016	92.5	90.0	93.5	78.8	92.4
Proposed U-Net	this work	93.0	92.0	94.0	88.0	93.2

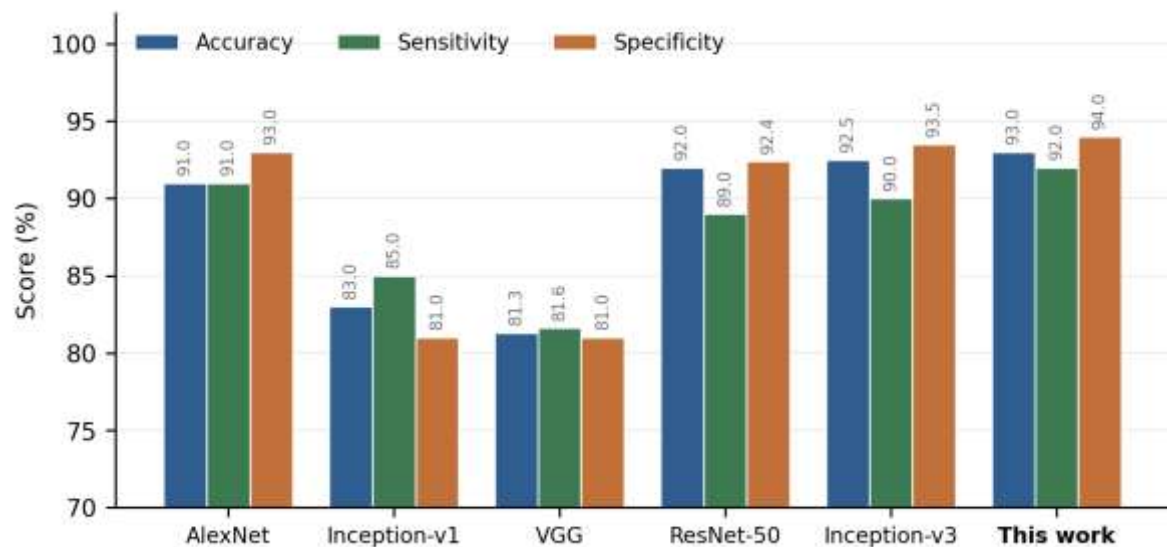


Figure 9. Accuracy, sensitivity and specificity of the proposed network against reference convolutional backbones evaluated in this study.

Our network takes the highest value in every column. The margin over the closest competitor, Inception-v3, is narrow on accuracy and specificity at half a percentage point each, and wider on sensitivity at two points. Since sensitivity governs how much lesion tissue is recovered, that is the more consequential difference for this task. The advantage over the classification backbones is unsurprising in kind, if not entirely in size. Those architectures were built to compress an image down to a label, and they have no decoder pathway, so they must reconstruct spatial detail that the U-Net preserves explicitly through its skip connections.

Comparison with published work on the same collection

A harder comparison is against published methods evaluated on the identical TCGA-LGG FLAIR collection. Table 7 and Figure 10 summarise five of them, with parameter counts included where the source reports enough detail to state one.

Table 7. Reported segmentation performance on the TCGA-LGG FLAIR collection. Expert inter-reader agreement on this lesion type is about 0.84 [7]. Protocols differ, so these figures are indicative rather than strictly equivalent.

Study	Method	Dice / F1	Validation protocol	
Buda et al. (2019) [7]	U-Net	0.82	22-fold validation	cross-
Naser and Deen (2020) [28]	U-Net with VGG-16	0.84	5 test sets	
Ottom et al. (2022) [29]	U-Net baseline	0.85	Hold-out	
Ottom et al. (2022) [29]	ZNet with amplification	0.92	Hold-out	
Chen et al. (2024) [30]	U-Net++ with EfficientNet	0.918	Hold-out	
This work	Compact U-Net, 5.69 M	0.90	Single hold-out	

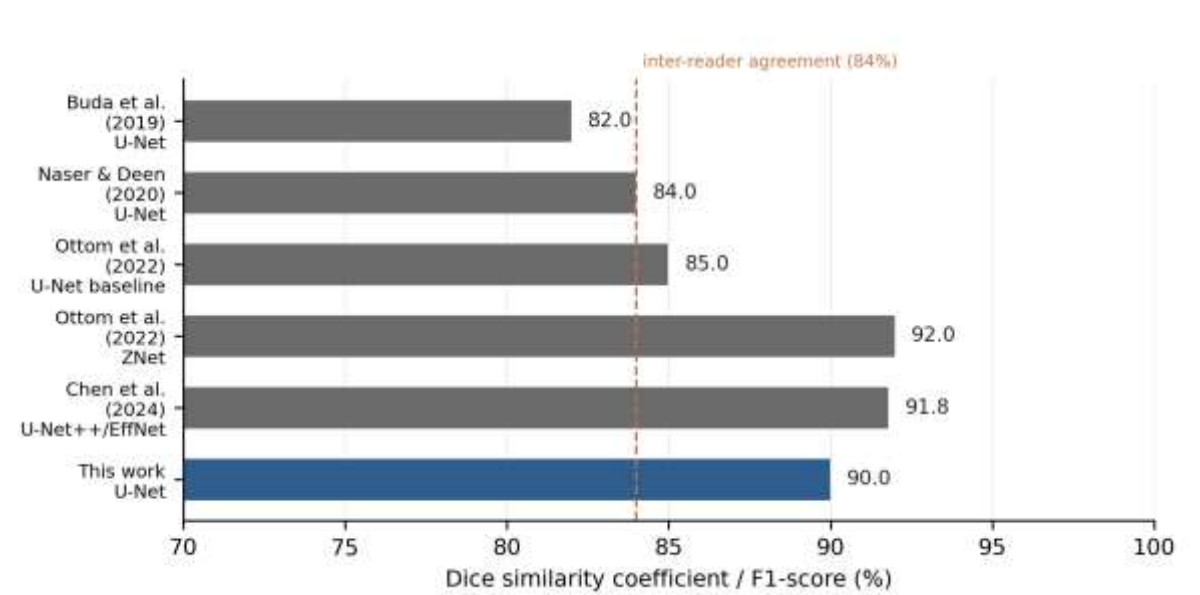


Figure 10. Dice similarity coefficient reported by published methods on the TCGA-LGG FLAIR collection. The dashed line marks the expert inter-reader agreement of 0.84 reported by Buda and colleagues [7].

At 0.90 the proposed model sits above the inter-reader ceiling of 0.84 and above the single-pass U-Net baselines at 0.82 to 0.85, and below the two heaviest methods at 0.918 and 0.92. Two qualifications matter here. Validation protocols are not the same: Buda and colleagues used 22-fold cross-validation, which is considerably more demanding than a single hold-out split and would be expected to return a lower, more conservative figure for the same underlying model. And exceeding the inter-reader figure does not mean the model outperforms a radiologist. It means the model agrees with this particular reference annotation more closely than two independent readers agree with each other, which is a statement about the consistency of a fixed target, not about clinical correctness.

Discussion

What the main result means

Our principal finding is that a compact U-Net trained on 2,836 annotated slices recovers most of the segmentation accuracy that substantially heavier models achieve on the same benchmark. A Dice of 0.90 against 0.918 for a U-Net++ with an EfficientNet encoder [30] and 0.92 for ZNet with extensive synthetic data amplification [29] is a gap of roughly two percentage points, obtained with 5.69 million parameters, no pretrained weights, no nested skip pathways and no data synthesis. Read as an efficiency question rather than a leaderboard question, that is the substantive result: the marginal accuracy bought by the extra machinery in those models is small relative to what it costs to run them.

Why the compact model holds up is worth saying plainly. Segmenting FLAIR abnormality in lower-grade glioma is, in representational terms, a fairly constrained problem. There is one foreground class, the lesion is characterised largely by intensity contrast against surrounding parenchyma, and the imaging is a single standardised sequence. Very high-capacity encoders earn their keep when a task requires separating many visually similar classes, or generalising across highly heterogeneous appearances. Neither condition applies strongly here, and that is our best explanation for why returns to capacity fall off so quickly on this dataset. We would not expect the same conclusion to transfer to multi-class BraTS sub-region segmentation, where the classes are genuinely harder to tell apart.

Where the residual error lives

Three independent lines of evidence point to the same conclusion about the remaining ten percent of error. The precision and sensitivity gap in Table 5 says the model over-includes. The confusion matrix in Figure 6 shows both off-diagonal terms small but non-zero, with the false-positive rate applying to the much larger background class. The qualitative panels in Figure 8 show that where predicted and reference contours diverge, they diverge at the margin rather than in the lesion core, and the divergence is worst for the smallest lesion in panel (d). Taken together these say the model localises lesions reliably and errs almost entirely in placing their boundary.

That diagnosis has a direct methodological consequence, and it is the main reason we bothered to characterise the error rather than just report the total. Boundary error is not usually fixed by adding depth. Additional encoder stages further reduce the spatial resolution at which the deepest features are computed, and they can make small-object delineation worse rather than better. The interventions that address boundary error are different in kind. Overlap-based losses weight the objective toward the foreground class, whether Dice [23] or generalised Dice [36]. Boundary-weighted and focal losses concentrate gradient on hard pixels near the margin [37]. Attention gating suppresses background activations on the skip connections before concatenation [14]. Deep supervision applies loss at several decoder resolutions and constrains the intermediate reconstructions. On our evidence these are the more promising directions, and they happen to be the cheaper ones as well.

Metric interpretation and class imbalance

An honest reading of these results needs a caveat about accuracy. On a whole-slice FLAIR image, lesion pixels usually make up one or two percent of the total. A degenerate classifier predicting background everywhere would score above 0.98 on pixel accuracy while segmenting nothing whatsoever. Accuracy is close to uninformative on this task, and its persistence in the segmentation literature is mostly a convention inherited from classification.

As noted in Section 4.3, our metrics are mutually consistent at a background-to-tumour ratio near 2.09 to 1, far less extreme than the whole-slice ratio, so the evaluation was accumulated over a lesion-centred region rather than over complete

images. That protocol is defensible, since it stops the metrics being swamped by trivially classified background. It does mean the accuracy figure of 0.93 is not comparable to whole-image accuracies quoted elsewhere, and that the F1-score of 0.90 should be treated as the primary result. Stating this openly is a precondition for anyone else being able to use our numbers. Reviews of segmentation metrics have made the same argument for years, with limited effect [39,40].

Limitations

Several limitations bound what can be concluded from this study.

Results come from one train-validation-test split rather than cross-validation. No confidence interval or variance estimate accompanies the figures in Table 5, and the comparison against Buda and colleagues [7], who used 22-fold cross-validation, is therefore not like for like. A repeated or k-fold protocol would put these numbers on much firmer ground.

All 110 patients come from one curated collection. It spans five institutions, but the images were harmonised during curation, and we tested no external dataset from an independent site. Robustness to variation in scanner vendor, field strength and acquisition protocol is unmeasured, and that is the failure mode most often seen when segmentation models move between sites.

The network operates on independent axial slices and cannot exploit through-plane continuity. Clinically relevant lesion volume is a three-dimensional quantity, and slice-wise predictions can be inconsistent between adjacent slices in a way a volumetric model would avoid [12].

The output is binary. It separates FLAIR abnormality from background without distinguishing the sub-regions, such as enhancing tumour and necrotic core, that the BraTS protocol requires and that inform treatment planning [5,6].

Accuracy is size-dependent, as Figure 8(d) shows. Small lesions are segmented less precisely than large ones, and early detection is precisely the setting in which lesions are small. This limitation bears directly on the intended clinical use.

No radiologist assessed the outputs, no reader study was run, and we measured nothing about whether the model changes the time or accuracy of clinical reporting. Segmentation accuracy is necessary for clinical utility but nowhere near sufficient.

5.5 Future work

The most immediate step is to repeat the evaluation under k-fold cross-validation with reported confidence intervals, which would make the comparison against published baselines valid and would supply the variance currently missing from Table 5. Second, the loss deserves revisiting: replacing or supplementing binary cross-entropy with a Dice or focal-Dice compound objective attacks the boundary error identified in Section 5.2 directly, and it is a small change relative to its likely effect. Third, external validation on an independent collection, ideally including data acquired locally in Pakistan, would test the generalisation that single-source evaluation cannot. Fourth, the bridge holds 62% of the weights in this model, so anyone wanting a genuinely small network should look there next rather than at the decoder. Extending the model to volumetric input, or adding lightweight inter-slice consistency, would let lesion volume be estimated directly, which is the measurement clinicians actually use for follow-up. Moving from binary to multi-class output would finally bring the method into line with the sub-region protocol used in the wider brain tumour segmentation literature.

Conclusion

We evaluated a compact U-Net with depthwise-separable decoder convolutions for binary segmentation of FLAIR abnormality in lower-grade glioma, using the public TCGA-LGG collection. Trained from scratch on 2,836 annotated slices with no pretrained weights, the model reached an accuracy of 0.93, sensitivity of 0.92,

specificity of 0.94, precision of 0.88 and an F1-score of 0.90 on a held-out test split of 394 slices, with training and validation curves converging by roughly epoch 40 and no serious overfitting.

Two findings carry beyond the specific numbers. The compact architecture recovers most of the accuracy of considerably heavier models on this task: the gap to a U-Net++ with an EfficientNet encoder and to a synthesis-augmented encoder-decoder is about two Dice points, and it comes from a network of 5.69 million weights, roughly one fifth the size of the original U-Net. Where computing resources are limited, that trade looks favourable. The second finding is that the residual error is not spread evenly. Precision trailing sensitivity, the shape of the confusion matrix and the qualitative panels together put the error at the lesion boundary and identify small lesions as the hardest case, which points toward overlap-based and boundary-weighted losses, attention gating and deep supervision rather than toward more network depth.

These results are bounded by a single train-validation-test partition, a single-source dataset, two-dimensional slice-wise processing and the absence of any clinical reader study. They show the approach is feasible and quantify what it currently achieves. They do not show it is ready for clinical use. Cross-validated re-evaluation, a boundary-sensitive objective and external validation on independently acquired data are the necessary next steps, and each is modest in cost relative to its likely value.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of interest. The authors declare that they have no competing interests.

Ethics approval. This study used only publicly available, fully de-identified imaging data from The Cancer Imaging Archive. No new human or animal data were collected, so institutional ethical approval was not required.

Data availability. The lower-grade glioma segmentation dataset analysed here is publicly available from The Cancer Imaging Archive as part of The Cancer Genome Atlas TCGA-LGG collection (<https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK>) and is redistributed with segmentation masks by Buda and colleagues [7].

Author contributions. K.A. conceived the study, designed and implemented the network, ran the experiments and drafted the manuscript. Y.A. prepared and curated the dataset and implemented the preprocessing and augmentation pipeline. A.M. carried out the comparative evaluation against published methods and the literature review. N.A. produced the figures, verified the parameter-budget calculations and cross-checked the reported metrics. All authors reviewed the manuscript and approved the final version.

Sustainable Development Goals. This work relates to United Nations Sustainable Development Goal 3, Good Health and Well-Being.

References

- [1] Ostrom QT, Price M, Neff C, Cioffi G, Waite KA, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2016-2020. *Neuro-Oncology*. 2023;25(Suppl 4):iv1-iv99. doi:10.1093/neuonc/noad149
- [2] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2024;74(3):229-263. doi:10.3322/caac.21834
- [3] Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a

- summary. *Neuro-Oncology*. 2021;23(8):1231-1251. doi:10.1093/neuonc/noab106
- [4] DeAngelis LM. Brain tumors. *New England Journal of Medicine*. 2001;344(2):114-123. doi:10.1056/NEJM200101113440207
 - [5] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*. 2015;34(10):1993-2024. doi:10.1109/TMI.2014.2377694
 - [6] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*. 2017;4:170117. doi:10.1038/sdata.2017.117
 - [7] Buda M, Saha A, Mazurowski MA. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*. 2019;109:218-225. doi:10.1016/j.compbiomed.2019.05.002
 - [8] Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*. 2016;35(5):1240-1251. doi:10.1109/TMI.2016.2538465
 - [9] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*. 2017;35:18-31. doi:10.1016/j.media.2016.05.004
 - [10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. p. 3431-3440.
 - [11] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. *Lecture Notes in Computer Science*, vol. 9351. Springer; 2015. p. 234-241. doi:10.1007/978-3-319-24574-4_28
 - [12] Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: *MICCAI*. *Lecture Notes in Computer Science*, vol. 9901. Springer; 2016. p. 424-432. doi:10.1007/978-3-319-46723-8_49
 - [13] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: A nested U-Net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis (DLMIA)*. *Lecture Notes in Computer Science*, vol. 11045. Springer; 2018. p. 3-11. doi:10.1007/978-3-030-00889-5_1
 - [14] Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv:1804.03999*. 2018.
 - [15] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
 - [16] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*. 2021;9:82031-82057. doi:10.1109/ACCESS.2021.3086020
 - [17] Baig E, Shah MM, Bajwa MH, Shamim MS, Enam SA, et al. Conducting the Pakistan brain tumour epidemiology study, report on the methodology. *Journal of the Pakistan Medical Association*. 2022;72(Suppl 4):S10-S15.
 - [18] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p. 1251-1258.

- [19] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
- [20] Wadhwa A, Bhardwaj A, Verma VS. A review on brain tumor segmentation of MRI images. *Magnetic Resonance Imaging*. 2019;61:247-259. doi:10.1016/j.mri.2019.05.043
- [21] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*. 2017;36:61-78. doi:10.1016/j.media.2016.10.004
- [22] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017;42:60-88. doi:10.1016/j.media.2017.07.005
- [23] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *Fourth International Conference on 3D Vision (3DV)*. IEEE; 2016. p. 565-571. doi:10.1109/3DV.2016.79
- [24] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: Transformers for 3D medical image segmentation. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022. p. 574-584.
- [25] Magadza T, Viriri S. Deep learning for brain tumor segmentation: A survey of state-of-the-art. *Journal of Imaging*. 2021;7(2):19. doi:10.3390/jimaging7020019
- [26] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of Digital Imaging*. 2017;30(4):449-459. doi:10.1007/s10278-017-9983-4
- [27] Paszke A, Chaurasia A, Kim S, Culurciello E. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147. 2016.
- [28] Naser MA, Deen MJ. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. *Computers in Biology and Medicine*. 2020;121:103758. doi:10.1016/j.combiomed.2020.103758
- [29] Ottom MA, Rahman HA, Dinov ID. Znet: Deep learning approach for 2D MRI brain tumor segmentation. *IEEE Journal of Translational Engineering in Health and Medicine*. 2022;10:1800508. doi:10.1109/JTEHM.2022.3176737
- [30] Chen Y, Quan L, Long C, Wang J, Zhang J, Zhang Y. Brain tumor segmentation based on the U-Net++ network with EfficientNet encoder. *Technology and Health Care*. 2024;32(1):183-195. doi:10.3233/THC-248016
- [31] Santosh Kumar P, Sakthivel VP, Raju M, Sathya PD. Brain tumor segmentation of the FLAIR MRI images using novel ResUnet. *Biomedical Signal Processing and Control*. 2023;82:104586. doi:10.1016/j.bspc.2023.104586
- [32] Ghosh S, Chaki A, Santosh KC. Improved U-Net architecture with VGG-16 for brain tumor segmentation. *Physical and Engineering Sciences in Medicine*. 2021;44(3):703-712. doi:10.1007/s13246-021-01019-w
- [33] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
- [34] Pedano N, Flanders AE, Scarpace L, Mikkelsen T, Eschbacher JM, Hermes B, et al. The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) (Version 3). *The Cancer Imaging Archive*; 2016. doi:10.7937/K9/TCIA.2016.L4LTD3TK
- [35] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of Big Data*. 2019;6:60. doi:10.1186/s40537-019-0197-0

- [36] Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis (DLMIA). Lecture Notes in Computer Science, vol. 10553. Springer; 2017. p. 240-248. doi:10.1007/978-3-319-67558-9_28
- [37] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. p. 2980-2988.
- [38] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). 2015. arXiv:1412.6980.
- [39] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging. 2015;15:29. doi:10.1186/s12880-015-0068-x
- [40] Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297-302. doi:10.2307/1932409
- [41] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). 2012. p. 1097-1105.
- [42] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). 2015. arXiv:1409.1556.
- [43] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. p. 1-9.
- [44] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. p. 770-778.
- [45] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. p. 2818-2826.