ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

Comparative Analysis of U-Net and Segment Anything Model for Brain MRI Tumour Segmentation

[1] Vandan Jain, [2] Shreyans Jain H, [3] Pronov Mazumdar

[1] [2] [3] Department of Data Science and Business Systems, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

*Corresponding Author Email: [1] vj3786@srmist.edu.in, [2] sj0261@srmist.edu.in, [3] pm2235@srmist.edu.in

Abstract*— Accurate segmentation of brain tumours from magnetic resonance imaging (MRI) scans is critical for diagnosis, treatment planning, and monitoring of neurological diseases. This paper presents a comparative study between two fundamentally different segmentation paradigms: (1) a task-specific U-Net model trained from scratch on brain MRI data, and (2) Meta's Segment Anything Model (SAM), a large-scale foundation model applied in a zero-shot setting with bounding box prompts. Using the LGG (Lower-Grade Glioma) MRI Segmentation dataset, we evaluate both approaches on an identical validation set of 786 slices (266 tumour-containing) using five metrics: Dice coefficient, IoU, 95th- percentile Hausdorff distance (HD95), sensitivity, and precision. SAM achieves superior overall segmentation quality (Dice: 0.877 vs. 0.745, $p < 0.001$) and significantly better boundary precision (HD95: 5.37 vs. 15.53 pixels) compared to U-Net. On tumour-containing slices, both models achieve comparable Dice scores (0.637 vs. 0.638) but exhibit fundamentally different error profiles: SAM favours high sensitivity (0.987) at the cost of precision (0.494), while U-Net favours precision (0.750) over sensitivity (0.661). These findings demonstrate that foundation models hold significant promise for medical image segmentation in interactive clinical workflows, while task-specific models remain the practical choice for fully automated pipelines.

Keywords: Brain tumour segmentation, U-Net, Segment Anything Model, medical image segmentation, deep learning, MRI, foundation models.

I. INTRODUCTION

Brain tumours are a big health problem around the world, and gliomas make up about 80% of all malignant brain tumours. For planning surgery, radiation therapy, and long-term monitoring of how the disease is getting worse, it is important to be able to accurately see the edges of tumours on MRI scans. Radiologists' manual segmentation is the best way to do it, but it takes a long time, is subjective, and can vary from person to person. Deep learning has changed the way medical images are segmented in the last ten years. The **U-Net architecture** [1], which came out in 2015, became the standard for biomedical image segmentation because its elegant encoder-decoder structure with skip connections keeps spatial information across scales. U-Net and other task-specific models are trained from start to finish on data that is specific to their domain. When they are used for inference, they automatically segment the data. Recently, foundation models that have been trained on huge datasets have become a strong new way of doing things. Meta AI's **Segment Anything Model (SAM)** [2] was trained on more than 11 million images and 1.1 billion masks. It can segment images in many different visual domains without any training. SAM works through a prompt-based interface that takes points, bounding boxes, or text as input to help with segmentation. This poses a significant inquiry for medical imaging: is it possible for a general-purpose foundation model, devoid of any domain-specific fine-tuning, to achieve performance that matches or surpasses that of a task-specific model trained on medical data? This paper examines this question via a controlled comparative study on brain MRI tumour segmentation. We make a comparison:

1. A U-Net model trained from the ground up on the LGG MRI dataset using modern training methods (data augmentation, batch normalisation, dropout, combined loss), which is the task-specific paradigm.
2. SAM (ViT-B variant) with bounding box prompts based on ground truth masks, showing what the foundation model paradigm looks like when prompting is done correctly.

Here are the things we've done:

- A methodical comparison of task-specific and foundational model methodologies for brain tumour segmentation, employing five complementary evaluation metrics alongside statistical significance testing.
- An examination of the divergent error profiles between the two paradigms, demonstrating that similar Dice scores can obscure fundamentally distinct segmentation behaviours.
- Talk about the real-world trade-offs between fully automatic and prompt-guided segmentation paradigms for use in the clinic.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

II. RELATED WORK

A. U-Net and its Variants for Medical Images Segmentation

The U-Net architecture [1] was first made for biomedical image segmentation, but it is now one of the most popular architectures in medical imaging. Because it has a symmetric encoder-decoder structure with skip connections, the model can pick up on both high-level semantic features and fine-grained spatial details. Since then, a lot of different versions have been proposed, such as attention mechanisms, nested skip connections, and self-configuring pipelines. Each of these attempts to fix specific issues, like fusing features at different scales and choosing the right architecture [12, 18]. More recently, hybrid architectures like **TransUNet** [10] have added Vision Transformer encoders to the U-Net framework. These encoders use self-attention to find long-range dependencies that models that only use convolutional layers don't see.

There are a lot of public benchmarks for brain tumour segmentation that have been used to test U-Net-based methods. These methods usually get Dice scores between 0.75 and 0.90, depending on the tumour sub-region, the complexity of the dataset, and how many architectural changes were made. Some of the main trends that have helped improve performance recently are attention gates that help focus on areas that are important for tumours, residual connections that help the gradient flow better, and advanced loss functions like Dice loss and its variants [11], [12]. It is now even easier to choose an architecture thanks to self-configuring frameworks like **nnU-Net** [15]. You can get good results on a lot of different segmentation tasks without having to do a lot of manual tuning.

B. Foundation Models for Medical Imaging

Foundation models have changed the way we think about computer vision. SAM [2], which was trained on the SA-1B dataset with 11 million images, showed that it could transfer knowledge very well without any prior training. The next version, **SAM 2** [9], added unified image and video segmentation to the architecture, making it more efficient and accurate. Multiple studies have examined SAM's relevance to medical imaging tasks. Mazurowski et al. [4] assessed SAM across various medical imaging modalities and observed inconsistent performance, with SAM typically lagging specialised models lacking domain-specific fine-tuning. He et al. [5] specifically evaluated SAM on brain MRI and observed that prompt quality substantially affects segmentation accuracy, with bounding box prompts typically surpassing point prompts.

Specialised adaptations have come about because people are trying to close the gap between natural and medical images in terms of domain and capacity. **GoodSAM** [13] suggested prompting strategies that consider distortion to deal with distribution shifts in deployment settings. The Medical SAM Adapter [14] introduced lightweight adapter modules that make it easy to fine-tune SAM's pretrained weights for medical segmentation without having to retrain the whole model. **SAM 2**[8], which was trained on more than 1.5 million pairs of medical images and masks from 10 different imaging modalities, showed that adapting foundation models to the medical field can produce results that are as good as those of specialist models. This suggests that the difference between general-purpose and domain-specific models is getting smaller quickly.

C. Comparative Studies

Several studies have been conducted comparing conventional deep learning models with foundation models in the context of medical segmentation. All such studies have established the superiority of task-specific models over foundation models in terms of performance, despite the flexibility and lower requirement of training data offered by the latter [6]. The swift shift in focus from the original SAM model to the newer SAM 2 model and then the domain-specific variants such as **MedSAM** [8] and again SAM [9] emphasizes the importance of regular benchmarking in the interest of understanding the dynamic balance between the two. Our work adds to the existing pool of such comparisons with a detailed analysis of the LGG brain MRI dataset.

III. LITERATURE SURVEY

A lot of work has been conducted in recent years to enhance task-specific segmentation models for brain MRI scans and to make foundation models work better for medical imaging.

Attention mechanisms are currently used more frequently than any other technique to develop something for a specific task. Abrar et al. [11] showed that using attention-based convolutional U-Net architectures in brain tumour segmentation pipelines can help in identifying the edges of tumours, especially if they have abnormal shapes. However, this work is only applicable to 2D slice-by-slice processing using a single dataset. Thus, the efficacy of this attention-based design for other types of tumours remains unverified. Another residual U-Net with a Swin transformer encoder has been proposed by Angona and Mondal [17]. This has shown improved segmentation performance in brain MRI images using the local inductive bias of convolutional layers and the global context information of the transformer encoder. This has a major drawback in that it demands a lot of computational power. This hybrid model demands more GPU memory and training time than models using only convolutional layers. Moreover, the test is conducted on a small number of patients and hence cannot be generalized. Three-dimensional models of these frameworks have been explored by scientists. Puldajoov et al. [16] tested various 3D deep learning models like U-Net, V-Net, Attention U-Net, ResNet-based U-Net, and transformer-based models for brain

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

tumor segmentation. They found that the attention mechanism and residual connection always improve the accuracy of the segmentation of the brain tumor. However, the comparison is only made on a single dataset and does not include the inference time and the feasibility of the models in real-world scenarios.

Jin and Ibrahim [12] and Jiangtao et al. [18] conducted a comprehensive analysis of the U-Net and other U-Net architectures. They found that the attention mechanism and depth of the models improve the accuracy of the segmentation but require more computational power. However, the problem with the above two surveys is that they are more descriptive in nature. This makes it difficult to compare the results of the two surveys because neither of the surveys provides a comprehensive re-evaluation of the methods. The problem of choosing an architecture is solved by a self-configuring approach that automatically configures preprocessing, network structure, and post-processing for any segmentation dataset. It is a strong and reproducible approach for running tests for medical segmentation. However, the biggest problem that is associated with the use of nnU-Net is that it requires a lot of computational power to be able to self-configure, even though it is a good solution in many instances. It is not designed for a scenario that is resource-constrained or for a scenario that is considered to be at the edge.

The fast development of SAM and its derivatives has sparked the interest of the research community in the foundation model domain. Mazurowski et al. [4] conducted a study that tested the zero-shot SAM in a variety of medical imaging modalities, showing that the results were inconsistent and that SAM performed worse than other models in the domain. However, the major drawback of the study was that it did not investigate the domain adaptation aspect of the SAM at all, leaving the reader wondering how well the model would perform with a little bit of supervision for the tasks at hand. He et al. [5] also tested the SAM in 12 different medical imaging datasets, showing that the quality of the prompt has a major impact on the segmentation accuracy of the images, with bounding box prompts working much better than point prompts. However, the paper was a preprint, and it only touched upon the different modalities that the SAM was tested in without going into the details of the failures that the SAM may experience in a single task, such as the segmentation of brain tumors.

These adaptations have resulted from the quest to narrow the gap between the disparity in the images formed by nature and medicine in domain and capacity. GoodSAM [13] has developed a distortion-aware prompting technique to address the domain shift in images. However, the technique is only applicable in breaking down images that are panoramic in nature and has not been tested in MRI or volumetric images in medicine. The Medical SAM Adapter [14] has also been used to develop lightweight adapter modules that facilitate the modification of the pre-trained weights of SAM in medical segmentation images. Although the technique has been effective in modifying the parameters, it still requires a set of labelled medical images to fine-tune the technique, which may be hard to find in a medical environment that is poorly funded. In the survey conducted by Huang et al. [6], there is an overview of the application of SAM with various forms of medical images. It is always mentioned that gliomas and other low-contrast images are difficult for zero-shot SAM to process. However, the problem with this survey is that there is no information about how the problems mentioned can be solved and tested. It is more of a diagnostic tool. In the MedSAM [8], where the model is fine-tuned with over 1.5 million image-mask pairs with 10 imaging modalities, it is shown that the results can be similar to the results obtained with special models. However, the problem is the large number of data needed for fine-tuning the model, and the difficulty in working with low-contrast images with unclear boundaries, such as lower-grade gliomas.

The body of work indicates that there is a set of unresolved tensions between task-specific models that demand labelled data and have limited flexibility, and foundation models that call for precise prompting or expensive domain adaptation to work effectively with medical images. In this work, we conduct a direct comparison between the two paradigms using the LGG brain MRI dataset.

IV. METHODOLOGY

A. Dataset

We use the LGG (Lower-Grade Glioma) MRI Segmentation dataset [3], which is free to download from Kaggle. The dataset includes brain MRI scans from 110 patients from The Cancer Genome Atlas (TCGA), along with manual FLAIR abnormality segmentation masks for each scan. The images are given to you as TIFF files at their full resolution. There are 3,929 image-mask pairs in total, with each patient folder containing several 2D slices. The dataset has both tumour and non-tumor (healthy) slices, which is a realistic example of a class-imbalanced situation: only 33.9% of the slices have tumours.

B. U-Net Architecture

We use a U-Net structure and the best ways to train:

Encoder: Three blocks are used to encode. Each block has two 3x3 convolutional layers, then batch normalisation and ReLU activation. Then, a 2x2 max pooling is done. There are 32, 64, and 128 filters. After every block, the dropout rate is set to 0.1, 0.1, and 0.2.

Bottleneck: Two 3x3 convolutional layers with 256 filters, batch normalisation, ReLU activation, and a dropout rate of 0.3.

Decoder: Three blocks that decode. Each block has two 3x3 convolutional layers, one for batch normalisation and one for ReLU activation. The feature maps from the encoder blocks that go together are combined (skip connection) and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

enlarged by 2x2. The dropout rates are 0.2, 0.1, and 0.0.

Output: A convolutional layer that is 1×1 in size and uses sigmoid activation. This gives an output with only one channel. There are about 1.95 million parameters in this model.

The images are shrunk down to 128×128 pixels. After that, they become images with only one channel. The range for normalising pixel intensity is [0,1]. Binarizing masks and using nearest neighbour interpolation helps keep the labels in resized images correct.

To add data, we use random horizontal and vertical flips, random 90-degree rotations in k = {0,1,2,3}, random brightness changes of ±10%, and random contrast changes of ±10%.

Loss Function: The BCE loss and the Dice loss are combined. Loss is calculated for each sample and then averaged.

$$L = L_{BCE} + (1 - DSC) \quad (1)$$

The dataset has a lot more background pixels than tumour pixels, but this combination fixes that.

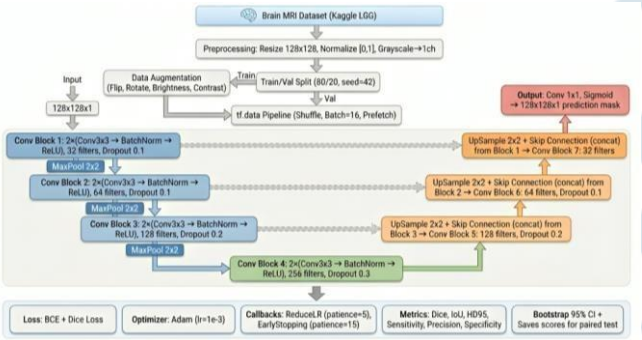

Figure 1. U-Net Architecture Diagram

C. SAM with Bounding Box Prompt

We use the SAM with the ViT-B (Vision Transformer Base) backbone, with 91 million parameters. The pretrained checkpoint is used as is, and hence there is no fine-tuning and the evaluation is performed without any training.

Prompt Generation: A bounding box prompt is generated for each image based on the ground truth mask associated with the image. A bounding box is the smallest box that can accommodate all the non-zero pixels of the mask. A 2-pixel space is added on each side of the bounding box. This is an oracle prompting scenario, i.e., it is the best possible prompting scenario because it is based on information that is not available in a clinical scenario.

Inference: For SAM's image encoder, it is required that images be resized to 96x96 pixels and in 3-channel format. If there is no tumor in the image, then it is predicted that there is no tumor. When images contain tumors, then the output of SAM is used in multi-mask output format, and the mask is chosen that has the highest confidence score predicted by SAM. For comparison with U-Net output, images are resized to 128x128.

No Tumor Handling: If the ground truth mask is empty, then there is no bounding box, and it is predicted that there is no tumor. This implies that the model is achieving a perfect Dice score of 1.0 on these slices.

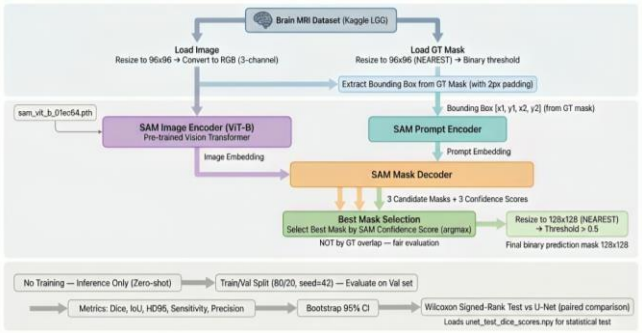

Figure 2. SAM with Bounding Box Prompt Architecture Diagram

D. Evaluation Metrics

We evaluate using five complementary metrics:

Dice Similarity Coefficient (DSC):

$$DSC = \frac{2\,|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \quad (2)$$

Intersection over Union (IoU / Jaccard Index):

$$IoU = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \quad (3)$$

95th Percentile Hausdorff Distance (HD95): This is a measure of boundary quality. It is calculated by finding the 95th percentile of symmetric surface distance between the predicted and ground truth boundaries. It is measured in pixels. A smaller number indicates better boundary alignment.

Sensitivity (Recall): TPR = TP / (TP + FN) - This is the proportion of true tumor pixels that were correctly detected.

Precision (Positive Predictive Value): PPV = TP / (TP + FP) - This is the proportion of detected tumor pixels that were actually part of the tumor.

The results are given in terms of mean ± standard deviation with bootstrapped 95% confidence interval (2,000 resamples). In this paper, we present results for two scenarios: (a) all slices and (b) tumor slices. Tumor slices are defined as slices where there is at least one pixel in the ground truth.

E. Statistical Testing

To assess whether observed differences are statistically significant, we apply the Wilcoxon signed-rank test [7] on the paired per-sample Dice scores from both models. This non-parametric test is appropriate because: (1) the samples are paired (both models evaluated on identical images), and (2) Dice score distributions are non-normal. We report two-sided p-values and consider $p < 0.05$ as statistically significant.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

V. EXPERIMENTAL SETUP

A. Dividing the Data

Using a fixed random seed (random_state=42), the dataset is split into 80% training and 20% validation/test. Data loading uses a deterministic alphabetical order for patient directories and filenames to make sure that both models are tested on the same set of 786 slices (266 with tumours and 520 without).

B. Setting Up U-Net Training

- Loss Function: BCE + Dice loss for each sample
- Optimizer: Adam with a starting learning rate of 10^{-3}
- LR Scheduling: ReduceLROnPlateau (factor0.5, patience 5, min LR 10^{-6})
- Early Stopping: Wait 15 on validation loss and restore the best weights.
- Epochs: 100 (the most)
- Size of Batch: 16
- TensorFlow 2.x / Keras is the framework.
- Size of input: 128×128×1 (black and white)
- Augmentation: flips, rotations, and brightness/contrast (for training only).

C. Setting up SAM

- Model: SAM ViT-B with 91 million parameters
- Checkpoint: sam_vit_b_01ec64.pth (trained on SA-1B before)
- Type of Prompt: Bounding box (oracle, based on the ground truth mask)
- Choosing a Mask: The best mask according to SAM's own confidence score
- PyTorch is the framework.
- The size of the SAM input is 96×96×3 (RGB).
- Size of the evaluation: 128×128 (changed for comparison)

D. Hardware

Tests were done on a system with an NVIDIA GeForce RTX 3070 Ti Laptop GPU. TensorFlow with CUDA/cuDNN acceleration was used for U-Net training, and PyTorch with CUDA was used for SAM inference.

VI. RESULTS

A. U-Net Training Dynamics

Figure 3 shows the training and validation curves for the trained model. We can see that the total loss is still going down. The training loss is now 0.77 and the validation loss is now 0.81. The training Dice coefficient rises to 0.25, and the validation Dice coefficient rises to 0.22 during the training process. These numbers are based on the soft Dice coefficients that were calculated during training. The post-threshold evaluation metrics are much higher because we are making the predictions more accurate by setting a threshold of 0.5.

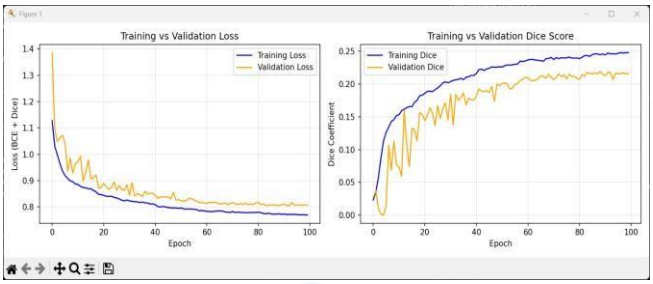

Figure 3. U-Net training and validation curves over 100 epochs. Left: Combined BCE + Dice loss. Right: Dice coefficient.

B. Results in Numbers

Table I shows a full comparison of all the evaluation metrics.

Overall performance: When tested on all slices, SAM does better than U-Net on all five metrics. The Dice advantage is very large (+0.132) because SAM handles non-tumor slices perfectly and has better boundary quality.

Performance on tumor-only slices: The picture is more complicated when it comes to slices that contain tumours that are clinically important. The Dice scores for both models are almost the same (U-Net: 0.638, SAM: 0.637), but this similarity hides very different error patterns, which are discussed in Section VII-A.

C. Statistical Significance

The Wilcoxon signed-rank test confirms that the differences between models are statistically significant:

All slices: $W = 21{,}905$, $p = 1.65 \times 10^{-9}$

Tumor slices only: $W = 13{,}606$, $p = 9.53 \times 10^{-4}$

Both comparisons exceed the significance threshold ($p < 0.05$), confirming that the per-sample differences between models are consistent and not attributable to chance, even on tumor slices where the mean Dice scores are nearly identical.

D. U-Net Segmentation Quality

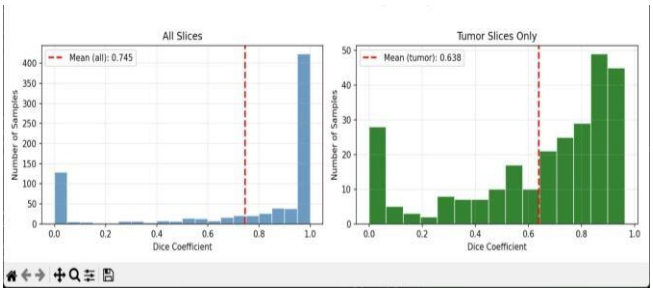

Figure 4. U-Net Dice score distribution. Left: All slices (mean = 0.745)

Right: Tumor slices only (mean = 0.638, n=266).

Fig. 4 presents the U-Net Dice score distribution. The all-slices distribution is bimodal: a dominant peak near 1.0 (non-tumor slices correctly predicted as empty) and a spread of lower scores (tumor slices). On tumor slices, the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

distribution ranges broadly from near 0 to above 0.9, with a notable cluster of near-zero scores representing tumors the model fails to detect, alongside a concentration of high scores (0.7–0.95) for successfully segmented tumors

Table 1. Comprehensive Comparison of U-Net and SAM + BBox on the Evaluation Set (786 Slices, 266 Tumor-Containing). All Values Reported as Mean ± Std [95% CI]. Bold Indicates the Better Value for Each Metric. † Indicates Lower is Better (HD95)

Category	Metric	U-Net	SAM + BBox	Better
All Slices	Dice	0.745 ± 0.376 [0.719, 0.773]	0.877 ± 0.195 [0.864, 0.890]	SAM
	IoU	0.709 ± 0.387 [0.681, 0.738]	0.826 ± 0.262 [0.809, 0.844]	SAM
	HD95†	29.21 ± 64.04 [24.60, 33.47]	1.82 ± 3.10 [1.60, 2.03]	SAM
	Sensitivity	0.885 ± 0.240 [0.869, 0.902]	0.996 ± 0.017 [0.994, 0.997]	SAM
	Precision	0.783 ± 0.363 [0.757, 0.809]	0.829 ± 0.262 [0.811, 0.846]	SAM
Tumor Only	Dice	0.638 ± 0.297 [0.602, 0.676]	0.637 ± 0.158 [0.617, 0.656]	Tied
	IoU	0.530 ± 0.286 [0.496, 0.566]	0.487 ± 0.171 [0.466, 0.508]	U-Net
	HD95†	15.53 ± 39.84 [10.87, 20.33]	5.37 ± 3.05 [5.00, 5.75]	SAM
	Sensitivity	0.661 ± 0.307 [0.625, 0.699]	0.987 ± 0.027 [0.984, 0.990]	SAM
	Precision	0.750 ± 0.273 [0.717, 0.783]	0.494 ± 0.182 [0.472, 0.516]	U-Net

E. SAM Segmentation Quality

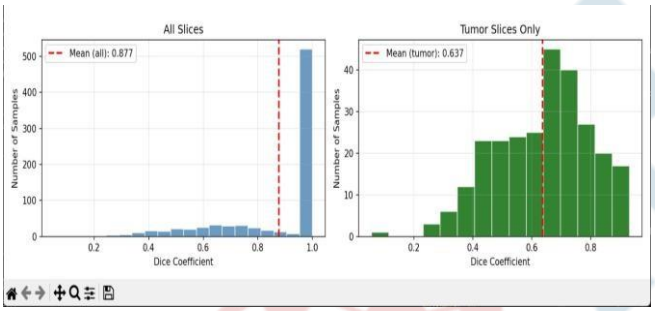

Figure 5. SAM + BBox Dice score distribution. Left: All slices (mean = 0.877). Right: Tumor slices only (mean = 0.637, n=266)

Fig. 5 shows the SAM Dice distribution. Compared to U-Net, SAM's tumor-only distribution is notably tighter (std = 0.158 vs. 0.297) with virtually no near-zero scores. This indicates SAM consistently detects tumors when given a bounding box prompt, whereas U-Net occasionally misses tumors entirely. The majority of SAM's tumor scores cluster in the 0.5–0.8 range.

F. Qualitative Results

Figures 6 and 7 illustrate the results of qualitative segmentation. In both models, correct results were obtained for non-tumor slices. U-Net produces correct but incomplete results for slices with tumors. In Figure 6, row 3 shows that the model produces results with good boundaries, given that Dice = 0.861. In row 4, however the model produces results that are under-segmented for complex morphology structures, given that Dice = 0.541. In SAM, more areas of tumors are covered. However, over-segmentation occurs because the bounding box is filled more than needed for the tumor boundary, as seen in row 4 where Dice = 0.385.

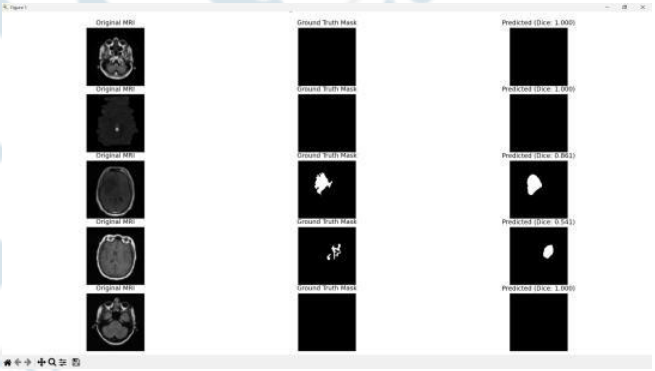

Figure 6. U-Net qualitative results. Columns: original MRI, ground truth, prediction with per-sample Dice

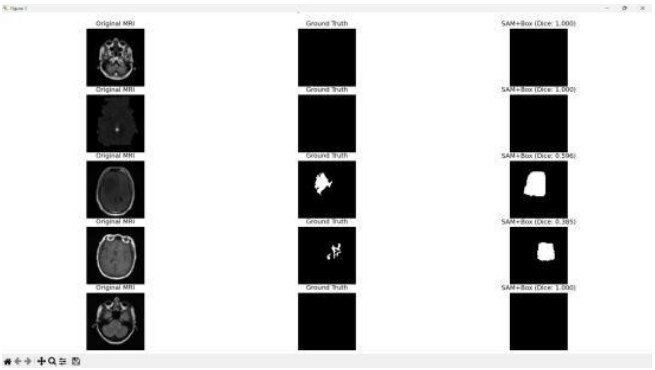

Figure 7. SAM + BBox qualitative results. Columns: original MRI, ground truth, prediction with per-sample Dice

VII. DISCUSSIONS

A. Analysis of the Error Profile

The most interesting thing is that U-Net and SAM get almost the same tumour Dice scores (0.638 vs. 0.637) even though they make mistakes in very different ways:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 6, June 2026

SAM over-segments: With a sensitivity of 0.987, SAM finds almost all tumour pixels. But its accuracy of 0.494 means that about half of the pixels it marks as tumour are healthy tissue. This over-segmentation happens because SAM is a general-purpose model that tends to fill the bounding box area too much. It doesn't have the domain-specific knowledge to tell the difference between subtle tumour boundaries and surrounding brain tissue.

U-Net doesn't segment enough: U-Net's tumour predictions are mostly correct, with a precision of 0.750. But its sensitivity of 0.661 means it misses about one-third of the tumour pixels. The model has learned to be cautious and only predicts tumours in areas where it is very sure, which means that diffuse or low-contrast tumour boundaries are not segmented well enough.

This precision-sensitivity trade-off has direct clinical implications: SAM is better when missing tumour tissue is very risky (like when planning a surgical margin), and U-Net is better when false positives are expensive (like when you want to avoid unnecessary intervention).

B. Boundary Quality

Even though the Dice scores are similar, SAM has much better boundary quality on tumour slices (HD95: 5.37 vs. 15.53 pixels, a 2.9× improvement). This means that when SAM's segmentation does overlap with the tumour, the edges are very close to the real contour. U-Net's higher HD95 means that there are sometimes big changes in the boundary, which is probably because it missed tumour areas that create distant surface points. The edge that SAM has over other models is that it was trained on more than 1.1 billion high-quality masks that cover a wide range of object boundaries.

C. Oracle Prompt Advantage

The bounding box prompts for SAM come from ground truth masks, which show an idealised situation. In clinical practice, these prompts would have to come from either a radiologist's manual input or an automated detector, both of which would not be as accurate as oracle prompts. So, SAM's reported performance is the best it can be. Even though the trained U-Net has this big advantage, it still gets a similar Dice on tumour slices. This shows that well-designed task-specific models can still compete with foundation models that are prompted in a good way.

D. Architectural Trade-Offs

U-Net (task-specific, 1.95M parameters):

- Needs labelled training data, which is about 3,143 pairs of images and masks.
- Inference that is completely automatic and doesn't need any human input.
- More accuracy, less sensitivity.
- Performance limited by the amount and variety of training data.

SAM (a base model with 91 million parameters):

- Doesn't need any training data for specific tasks (zero-shot).
- Needs a spatial prompt (bounding box) when deciding.
- More sensitivity, less accuracy.
- Uses information from 11 million different images and 1.1 billion masks.

E. Clinical Applicability

These models have different but important roles in clinical settings:

- **Automated screening:** U-Net is the only option that works and doesn't need any human input while still being very accurate.
- **Interactive refinement:** SAM works best when a radiologist can give a rough bounding box, which gives almost complete tumour coverage for detailed analysis.
- **Hybrid pipelines:** The best way to work would be to use both U-Net for automatic detection at first and then SAM for interactive boundary refinement, where doctors draw bounding boxes around areas of interest.

VIII. LIMITATIONS

1. Oracle bounding box: The prompts used in SAM come from ground truth and can be thought of as an idealised situation. When using prompts made by people or by a program, performance is likely to be lower.
2. Single dataset: This study is confirmed using solely the LGG MRI dataset. It is anticipated that performance will vary across different tumour types (e.g., HGG, metastases).
3. 2D processing: Each model is using its own 2D slices. This doesn't consider the possible benefit of using 3D information to make segmentation smoother.
4. Single evaluation split: We only used one 80/20 split in this work. In general, it is best to use k-fold cross-validation. Even though no hyperparameter tuning is done on the evaluation set, k-fold would give better performance estimates.
5. Single metric for statistical testing: In this study, we exclusively utilised the Dice scores for conducting the Wilcoxon test. Multivariate statistical testing is generally advised.

IX. CONCLUSION

This paper compared U-Net and SAM for brain MRI tumour segmentation on the LGG dataset. It used five metrics and statistical significance testing to compare the two methods on the same set of 786 slices (266 of which had tumours).

SAM with oracle bounding box prompts has a much better overall segmentation quality (Dice: 0.877 vs. 0.745, $p < 0.001$) and a much better boundary precision (HD95: 5.37 vs. 15.53 pixels). On slices with tumours, both models get

similar Dice scores (0.637 vs. 0.638, $p < 0.001$), but they make different kinds of mistakes. SAM gets the most sensitive score (0.987) but the least precise score (0.494), while U-Net gets the most precise score (0.750) but the least sensitive score (0.661).

These results show that foundation models like SAM have a lot of potential for medical image segmentation, especially in interactive clinical workflows where a doctor gives spatial guidance. A well-trained task-specific model with only 1.95 million parameters can match the tumor-level Dice of a foundation model with 91 million parameters running under perfect conditions. The key benefit is that it can run completely on its own. The two paradigms are not in competition; they work well together. Future clinical systems should use both.

Future research should investigate: (a) the optimisation of SAM on medical imaging data, as illustrated by MedSAM [8] and the Medical SAM Adapter [14]; (b) the utilisation of SAM 2's [9] video segmentation features for volumetric MRI propagation; (c) the incorporation of transformer-based U-Net variants [10], [17] and self-configuring frameworks [15] to establish more robust task-specific baselines; and (d) hybrid pipelines that merge automatic detection with prompt-based refinement.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Cham: Springer International Publishing, pp. 234–241, 2015.

[2] A. Kirillov et al., "Segment Anything," in *Proc. IEEE/CVF ICCV*, pp. 4015–4026, 2023.

[3] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Computers in Biology and Medicine*, vol. 109, pp. 218–225, 2019.

[4] M. A. Mazurowski et al., "Segment Anything Model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[5] S. He et al., "Computer-vision benchmark Segment-Anything Model (SAM) in medical images: Accuracy in 12 datasets," *arXiv preprint arXiv:2304.09324*, 2023.

[6] Y. Huang et al., "Segment Anything Model for medical images?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[7] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[8] J. Ma et al., "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[9] N. Ravi et al., "SAM 2: Segment Anything in Images and Videos," *arXiv preprint arXiv:2408.00714*, 2024.

[10] J. Chen et al., "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[11] M. Abrar et al., "Enhancing brain tumor segmentation using attention based convolutional UNet on MRI images," *Scientific Reports*, vol. 15, no. 1, p. 36603, 2025.

[12] C. Jin and H. Ibrahim, "Advancements in brain tumor segmentation: a literature survey of U-Net variants," *Neural Computing and Applications*, vol. 38, no. 5, p. 110, 2026.

[13] W. Zhang et al., "GoodSAM: Bridging domain and capacity gaps via Segment Anything Model for distortion-aware panoramic semantic segmentation," *arXiv preprint arXiv:2403.16370*, 2024.

[14] J. Wu et al., "Medical SAM Adapter: Adapting Segment Anything Model for medical image segmentation," *Medical Image Analysis*, vol. 102, p. 103547, 2025.

[15] F. Isensee et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[16] O. Puldajoov, Pooja, and A. Aggarwal, "Brain tumor segmentation capabilities of 3D deep learning architectures (U-Net, V-Net, Attention U-Net, ResNet-based U-Net, Transformer-based model)," *Neural Computing and Applications*, vol. 37, no. 32, pp. 27137–27149, 2025.

[17] T. M. Angona and M. R. H. Mondal, "An attention based residual U-Net with Swin Transformer for brain MRI segmentation," *Array*, vol. 25, p. 100376, 2025.

[18] W. Jiangtao, N. I. R. Ruhaiyem, and F. Panpan, "A comprehensive review of U-Net and its variants: advances and applications in medical image segmentation," *IET Image Processing*, vol. 19, no. 1, p. e70019, 2025.