CoMNeT: A MedNeXt-CorrDiff Framework for Volumetric Brain Tumor Segmentation

Michael L. Evans[1], **MD Fayaz Bin Hossen**[1], **MD Shibly Sadique**[1], **Walia Farzana**[1], **and Khan M. Iftekharuddin**[1,*]

[1]Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA
[*]Corresponding author: kiftekha@odu.edu

ABSTRACT

Accurate brain tumor segmentation from multiparametric magnetic resonance imaging (MRI) is critical for treatment planning, response assessment, and quantitative neuro-oncology research. However, automated segmentation remains a difficult task in computer vision because of variation in tumor appearance and MRI protocols across patient scans. Moreover, clinically important regions such as enhancing tumor (ET) and tumor core (TC) are often small relative to the full brain volume, furthering increasing the difficulty of achieving high voxel-level precision. In this paper, we show that combining a modern 3D convolutional segmentation model with corrective diffusion-based refinement and ensembling improves volumetric glioma segmentation on the UTSW-Glioma dataset. We propose CoMNeT, a MedNeXt-CorrDiff framework that uses four MRI modalities as input and predicts ET, TC, and whole tumor (WT) regions for automated brain tumor segmentation. MedNeXt is used as the primary segmentation model with Global Response Normalization for feature learning, while CorrDiff is trained as a postprocessing residual refinement method to correct errors in the probability maps before final thresholding. Using five-fold cross-validation, CoMNeT achieved the highest Dice score for most tumor regions, with ET, TC, WT, and average Dice scores of 0.7543 ± 0.0261, 0.6806 ± 0.0166, 0.9049 ± 0.0128, and 0.7798 ± 0.0184, respectively. CoMNeT outperformed two selected baseline models: SegResNet (0.7555 ± 0.0190 average Dice) and standalone MedNeXt (0.7697 ± 0.0154 average Dice). Our findings support the use of corrective diffusion and fold-level probability ensembling as practical additions to existing state-of-the-art 3D convolutional models for automated glioma segmentation.

1 Introduction

Gliomas are among the most clinically significant primary brain tumors, and their management depends heavily on accurate interpretation of magnetic resonance imaging (MRI). Manual labeling of the tumor regions is time consuming, has high variability across experts, and difficult to scale across larger studies. Automated brain tumor segmentation therefore remains an important problem in medical image analysis, particularly for quantitative assessment of the enhancing tumor, tumor core, and whole tumor regions.

Multiparametric MRI provides complementary information about tumor anatomy and surrounding tissues. Pre-contrast T1-weighted imaging (T1n) provides anatomical structure, post-contrast T1-weighted imaging (T1c) highlights regions of blood-brain barrier disruption and enhancement, T2-weighted imaging (T2w) captures fluid-sensitive abnormality, and T2-FLAIR suppresses cerebrospinal fluid to improve visualization of edema and infiltrative signal. In this study, we use these four modalities to predict BraTS-style tumor regions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). We use this labeling convention because it separates the most clinically relevant tumor subregions while allowing the model outputs to be interpretable.

Despite recent progress in deep learning-based segmentation, robust glioma segmentation remains challenging. Tumor appearance varies across patients, tumor grade, scanner protocol, and disease presentation. The target regions are also highly imbalanced, since tumor voxels occupy a small fraction of the full brain volume and ET and TC are often much smaller and more difficult to classify than WT. This creates a difficult learning problem where the model may appear stable while missing clinically important enhancing or core tumor voxels.

In contrast to the recent growth of interest around Transformer-based modeling, convolutional encoder-decoder models have been established as the de-facto backbones for data-limited medical imaging tasks because of their spatial inductive bias. The development of ConvNeXt is important in this context because it showed that a pure convolutional network could be modernized using Transformer design principles while preserving the efficiency and locality of the convolution operation[1]. These changes include large-kernel depthwise convolutions, inverted bottlenecks, residual pathways, and improved normalization. MedNeXt extends this direction to 3D medical image segmentation by placing ConvNeXt-style blocks inside a volumetric encoder-decoder network and scaling the model across depth, width, and kernel size[2,3]. This makes MedNeXt an efficient backbone for glioma

segmentation, where the model is expected to learn a combination of local boundary detail, multi-modal MRI contrast, and the larger volumetric context.

In this work, we propose CoMNeT, a MedNeXt-CorrDiff framework for volumetric glioma segmentation on the UTSW-Glioma dataset. The framework uses MedNeXt as the primary segmentation model, CorrDiff for corrective diffusion post-processing, and fold-level ensembling to combine the predictions of independently trained cross-validation models. We compare CoMNeT against SegResNet and standalone MedNeXt to show that our proposed refinement and ensembling strategy improves segmentation performance beyond two strong baselines. The main contributions of this work are as follows:

- Evaluate automated glioma segmentation on the UTSW-Glioma dataset using four MRI modalities and BraTS-style ET, TC, and WT labels on two baseline models.
- Integrate CorrDiff as a corrective diffusion component that learns to refine residual segmentation errors from the MedNeXt probability maps.
- Use fold-level model ensembling so that predictions from each unique cross-validation fold contribute to the final segmentation mask.
- Show that CoMNeT improves five-fold cross-validation Dice Score compared to SegResNet and MedNeXt.

2 Related work

Most modern medical imaging segmentation pipelines build on the encoder-decoder structure first introduced by U-Net, where a contracting path learns semantic context and an expanding path recovers spatial resolution through skip connections[4]. This design was extended to volumetric data by 3D U-Net and V-Net, which process 3D image volumes as a whole rather than analyzing each 2D slice independently[5,6]. These volumetric models are well suited for MRI because tumor appearance is represented as three-dimensional voxels and the relationship between 2D slices contains useful anatomical information.

SegResNet is a strong residual convolutional baseline for brain tumor segmentation and related volumetric segmentation tasks. It follows the same general encoder-decoder pattern, but uses residual blocks to improve gradient flow and feature reuse[7]. Residual segmentation backbones are useful for this task because they can train deeper networks while still preserving low-level boundary information. In this work, we include SegResNet as a baseline to compare CoMNeT against an established 3D residual segmentation model.

The nnU-Net framework later showed that a carefully configured U-Net can remain highly competitive across medical segmentation tasks when preprocessing, patch size, and training settings are adapted to the specific dataset[8]. Transformer and hybrid architectures have also been applied to medical image segmentation because attention can model long-range dependencies[9–11]. However, Transformer-style models often require higher GPU memory and larger training datasets, much more common in natural images. For the UTSW-Glioma dataset, we find that a modernized convolutional backbone remains practical for 3D segmentation at this size.

ConvNeXt was motivated by the idea that the apparent advantage of Transformer-based vision models was partly tied to architectural and training choices rather than the attention mechanism itself. By progressively modernizing a ResNet-style convolutional network, ConvNeXt showed that pure ConvNets can benefit from larger receptive fields, depthwise convolution, inverted bottlenecks, fewer activation functions and normalization layers, and Transformer-inspired scaling while remaining more data efficient[1]. This is particularly relevant for medical imaging, where fully volumetric Transformers can be memory intensive and where annotated datasets are often smaller than natural-image datasets.

MedNeXt builds directly upon the ConvNeXt design for medical image segmentation. Instead of using ConvNeXt as a classification backbone, MedNeXt places 3D ConvNeXt-style blocks into an encoder-decoder segmentation architecture[2]. The design keeps the main strengths of U-Net-like models, including multi-resolution feature learning and skip connections, but replaces conventional ResNet-style convolutional blocks with modernized ConvNeXt blocks. MedNeXt also introduces residual upsampling and downsampling blocks, which helps preserve semantic information as the model changes spatial resolution. This is particularly useful in glioma segmentation because the network must learn small enhancing regions and core tumor boundaries while still maintaining the broader context needed for accurate whole-tumor segmentation.

ConvNeXt V2 then added Global Response Normalization (GRN) to improve feature competition between channels and reduce redundant feature responses[12]. Similarly, MedNeXt-v2 then extended this idea to 3D medical segmentation by incorporating 3D GRN and scaling the architecture for volumetric representation learning[3]. In this paper, we refer to the updated GRN version of MedNeXt-v2 simply as MedNeXt. Our main motivation for using MedNeXt is not only that it is a strong convolutional model, but that it brings ConvNeXt-style large-context feature learning into a 3D encoder-decoder that is practical for medical imaging.

Diffusion models learn to reverse a gradual noising process and have been applied to medical imaging for generation, reconstruction, and segmentation[13,14]. In segmentation, many diffusion-based methods generate a mask directly from the image, which can be computationally expensive and can replace an otherwise strong discriminative segmentation model[15,16]. CorrDiff uses a different approach, instead of predicting the entire segmentation mask from scratch, it learns to correct the systematic errors of a base segmentation model[17]. The authors of CorrDiff used a U-Net segmentation which we replace with MedNeXt, as it provides stronger probability maps. CorrDiff is then used to refine the residual errors produced by the MedNeXt prediction mask.

3 Methodology

In this paper, the CoMNeT framework uses two steps for 3D brain tumor segmentation, shown in Figure 1. Each subject is loaded with four MRI modalities and a segmentation mask. The volumes are standardized through orientation correction, nonzero-voxel normalization, and fixed-size cropping or padding. MedNeXt is trained as the primary segmentation backbone, and CorrDiff then refines the initial prediction mask by modeling the residual error. Finally, fold-specific predictions are ensembled at the probability-map level and thresholded to produce the final ET, TC, and WT masks.

3.1 Dataset and preprocessing

We use the UTSW-Glioma dataset, a curated glioma multi-MRI dataset from the University of Texas Southwestern Medical Center that is publicly available through The Cancer Imaging Archive[18,19]. The dataset contains 625 patients treated at UTSW between 2006 and 2023. Each patient record includes pre-contrast T1-weighted, post-contrast T1-weighted, T2-weighted, and T2-FLAIR images[19]. The dataset also includes demographic, histopathologic, and molecular marker information, including IDH mutation status, 1p/19q codeletion status, MGMT promoter methylation status, tumor type, and tumor grade. These molecular and clinical variables are not used as model inputs in this study, because our goal is segmentation from MRI. However, they may make the dataset valuable for future multimodal analysis of segmentation performance with clinical data.

Each subject is required to include the four MRI modalities and a corresponding segmentation mask to serve as the ground truth. The segmentation labels are converted into three binary tumor-region masks: ET corresponds to enhancing tumor, TC includes the enhancing and core tumor components, and WT includes the full visible tumor, including edema. This conversion gives the model three output channels and keeps the evaluation focused on clinically meaningful tumor regions.

Before training, each subject directory is checked and validated for all required images and the corresponding segmentation mask. Subjects with missing modalities, missing masks, unreadable files, or incompatible image geometry are excluded from training. All volumes are converted to a common canonical orientation. Each modality is then normalized using z-score normalization over nonzero voxels:

$$\hat{x}_i = \frac{x_i - \mu_\Omega}{\sigma_\Omega + \varepsilon}, \tag{1}$$

where x_i is the intensity at voxel i, Ω is the set of nonzero voxels for that modality, μ_Ω and σ_Ω are the mean and standard deviation over nonzero voxels, and ε is a small constant for numerical stability. Normalizing only nonzero voxels prevents the background from corrupting the estimated MRI intensity data distribution.

After normalization, volumes are cropped to a fixed training patch size of $128 \times 160 \times 112$. This patch size preserves sufficient anatomical context while allowing 3D training on our available GPU memory. During training, we use lightweight spatial and intensity augmentation, limited to random flips, random affine transformations, random gamma adjustment, random noise, and random blur. Augmentation is applied only during training, validation inference uses deterministic preprocessing.

3.2 CoMNeT framework

CoMNeT is designed as a multipart segmentation pipeline rather than an architectural level change to MedNeXt. Our proposed framework consists of three core components: a MedNeXt 3D segmentation backbone, a CorrDiff corrective diffusion module, and a fold-level probability ensemble. In the first stage, MedNeXt provides the primary volumetric feature representation, followed with ensembling to improve the stability of the final prediction across independently trained fold models. In the final stage, CorrDiff learns to correct the residual segmentation errors produced by the prediction mask.

MedNeXt segmentation

We use the MedNeXt architecture as the primary segmentation backbone. The motivation for this comes from the original ConvNeXt block and its adaptation to work well on small clinical datasets. ConvNeXt showed that the convolutional design could be substantially improved when standard ConvNets were updated with larger kernels, depthwise spatial filtering, inverted bottlenecks, residual pathways, and improved normalization[1]. MedNeXt adapts these ideas to 3D medical image segmentation by placing ConvNeXt-style blocks inside an encoder-decoder architecture[2]. This is well suited for volumetric brain tumor segmentation from MRI because the model needs to preserve local boundary detail while also using larger anatomical context.

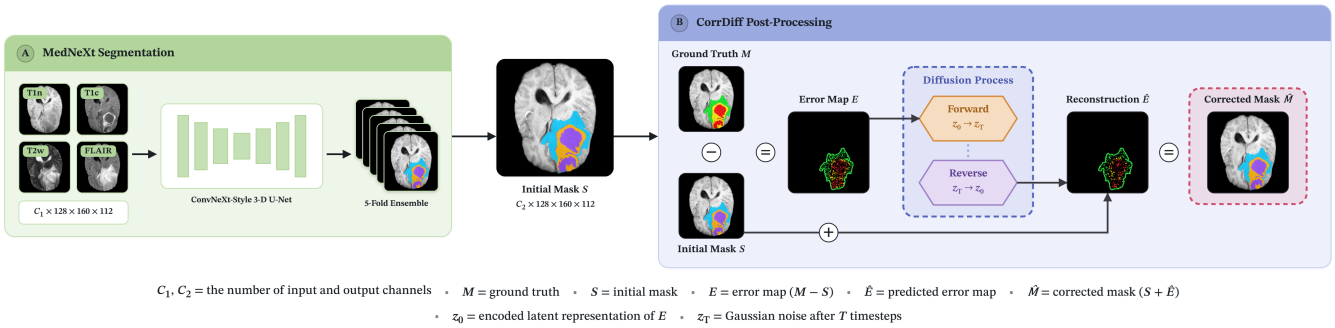

Figure 1. Overview of the proposed CoMNeT pipeline for UTSW-Glioma segmentation. Four MRI modalities are preprocessed and passed through the MedNeXt segmentation backbone. CorrDiff is then used as a corrective diffusion module for residual error refinement. Predictions from the unique fold-specific models are ensembled before final evaluation using BraTS-style tumor subregions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT).

In CoMNeT, MedNeXt follows the same encoder-decoder structure as a U-Net segmentation model. The encoder compresses the input volume into progressively lower-resolution feature maps, allowing the network to learn broader spatial context. The decoder then brings these features back toward the original image resolution and combines them with higher-resolution skip connections from the encoder. This structure is important for tumor segmentation because ET and TC often require fine local detail, while WT depends on the model recognizing the full extent of abnormal tissue across a larger field of view.

The internal MedNeXt blocks are used to make this 3D encoder-decoder more expressive without forfeiting the advantages of the convolution operation. Depthwise 3D convolutions capture spatial patterns within each feature channel, while pointwise convolutions mix information across channels. This block expands the feature representation, applies nonlinear processing and normalization, and then projects the representation back into the residual pathway. This lets each block learn a refinement to the current feature map instead of relearning the full representation. For this reason, MedNeXt can be made deep enough for volumetric segmentation while remaining practical to train using patches.

Global Response Normalization introduced in ConvNeXt V2 is also used to improve inter-channel feature competition and reduce redundant channel responses[12]. MedNeXt extends this idea to 3D medical segmentation by incorporating a 3D GRN module into the ConvNeXt-style backbone[3]. GRN gives the network a global summary of how strongly each feature channel responds across the full 3D patch, and uses that information to modulate the channel responses. For this task, some channels may become sensitive to enhancing tumor, edema, non-enhancing core, or normal anatomic context. We use GRN to help with the spatial imbalance of the tumor regions to avoid a small number of dominant responses to control the learned representation.

Fold-level ensembling

We use five-fold cross-validation for training and evaluation. Each fold produces a unique trained model because it is fit on a different training split and validated on a different held-out split. These fold-specific models often learn a slight variation in decision boundaries. In our proposed framework, each unique fold model generates its own ET, TC, and WT probability maps for a subject. We then average the probability maps voxel-wise for each tumor region and threshold the averaged probabilities to create the final binary masks. Averaging is performed before thresholding because it preserves more information than voting on binary masks. Voxels that are consistently predicted across fold models remain high confidence, while unstable fold-specific predictions are softened. This produces an initial segmentation prediction, shown as S in Figure 1 that uses information from all independently trained models and reduces variance from any single fold.

CorrDiff residual refinement

We propose adding a corrective diffusion model inspired by CorrDiff as a post-processing block to address the residual errors created by S. This choice follows the original CorrDiff implementation, which uses diffusion as a corrective model rather than generating the initial segmentation from the image[17]. This is a different approach from many other diffusion-based segmentation approaches, where the diffusion model is used as the primary mask generator[13–16].

After initial training of MedNeXt, we train and optimze CorrDiff on the UTSW dataset while keeping the MedNeXt model weights frozen. The corrective diffusion model receives the MRI volume, the MedNeXt coarse probabilities, the MedNeXt coarse logits, and an entropy-based uncertainty map. These inputs allow CorrDiff to understand where the segmentation backbone is confident, uncertain, or where the predicted tumor-region boundaries need correction.

The correction target is the residual between the binary tumor-region target and the MedNeXt probability map. Rather than diffusing this residual directly in the full image space, we first encode the residual with a 3D vector quantized variational autoencoder (VQ-VAE). The latent diffusion model then learns to denoise this compact residual representation while conditioned

on the MRI and MedNeXt output. At inference, CorrDiff samples a latent residual correction, decodes it into a correction to the MedNeXt logits, gates that correction using the coarse probabilities and uncertainty map, and adds the gated correction back to the MedNeXt logits before final sigmoid thresholding. CorrDiff is trained with a composite objective function that combines diffusion denoising, residual reconstruction, vector-quantization regularization, and final segmentation supervision:

$$L_{\text{CorrDiff}} = \lambda_{\text{diff}} L_{\text{diff}} + \lambda_{\text{recon}} L_{\text{recon}} + \lambda_{\text{vq}} L_{\text{vq}} + \lambda_{\text{seg}} L_{\text{seg}} \tag{2}$$

The diffusion term L_{diff} is an MSE noise-prediction loss in the latent space. L_{recon} is an L1 loss for reconstructing the target residual $r = y - \sigma(\ell_{\text{coarse}})$. L_{vq} is the VQ-VAE codebook and commitment loss, and L_{seg} combines binary cross-entropy with Dice loss on the corrected logits $\ell_{\text{final}} = \ell_{\text{coarse}} + \Delta\ell$. We use $\lambda_{\text{diff}} = 1$, $\lambda_{\text{recon}} = 1$, $\lambda_{\text{vq}} = 0.25$, and $\lambda_{\text{seg}} = 1$, encouraging CorrDiff to learn corrections that are plausible in the latent diffusion space and beneficial for the final tumor-region segmentation.

3.3 Loss functions

We train MedNeXt with a region overlap aware loss rather than a pure voxel-wise loss. This is useful because tumor regions occupy a small fraction of the full 3D brain volume. Dice-style objective functions directly optimize region overlap. For region r, the soft Dice score is

$$\text{DSC}_r = \frac{2\sum_i p_{r,i} y_{r,i} + \varepsilon}{\sum_i p_{r,i} + \sum_i y_{r,i} + \varepsilon}, \tag{3}$$

where $p_{r,i}$ is the predicted probability for voxel i, $y_{r,i}$ is the binary ground-truth label, and ε is a small constant for numerical stability. The standard Dice loss is defined as the average Dice loss across ET, TC, and WT:

$$L_{\text{Dice}} = 1 - \frac{1}{3} \sum_{r \in \{ET, TC, WT\}} \text{DSC}_r. \tag{4}$$

We train MedNeXt using a DSC++ variant, which replaces the standard Dice denominator with a squared probability term:

$$\text{DSC}_r^{++} = \frac{2\sum_i p_{r,i} y_{r,i} + \varepsilon}{\sum_i p_{r,i}^2 + \sum_i y_{r,i}^2 + \varepsilon}. \tag{5}$$

Our final loss combines DSC++ with binary cross-entropy:

$$L_{\text{MedNeXt}} = L_{\text{DSC}^{++}} + L_{\text{BCE}}. \tag{6}$$

This allows to keep the emphasis on overlap while adding voxel-wise supervision. We still evaluation using the Dice similarity coefficient, due to being the standard overlap metric for ET, TC, and WT segmentation.

We hold the training configuration consistent across SegResNet and MedNeXt experiments to ensure the performance differences can be interpreted primarily as differences between model designs and the CoMNeT framework. MedNeXt is trained with AdamW using a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. The batch size is 1, with gradient accumulation over 2 steps to obtain an effective batch size of 2. Training uses a 5-epoch warmup followed by cosine annealing, a maximum of 150 epochs, early stopping with patience 30, mixed precision, and gradient clipping with maximum norm 1.0. CorrDiff is trained as a second-stage post-processing block with the MedNeXt weights frozen. Core training settings are summarized in Table 1.

3.4 Evaluation metrics

Segmentation performance is evaluated separately for ET, TC, and WT. For each region, the model prediction is converted into a binary mask and compared with the corresponding binary ground-truth mask. Let P_r denote the predicted voxel set for region r, and let G_r denote the ground-truth voxel set. From these masks, we compute the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

The Dice similarity coefficient is the primary metric:

$$\text{DSC}_r = \frac{2|P_r \cap G_r|}{|P_r| + |G_r|} = \frac{2TP}{2TP + FP + FN}. \tag{7}$$

Dice measures the overlap between the predicted and ground-truth tumor region, with 1 indicating perfect overlap and 0 indicating no overlap. We also compute the region-wise average:

$$\text{DSC}_{\text{avg}} = \frac{1}{3} \left(\text{DSC}_{ET} + \text{DSC}_{TC} + \text{DSC}_{WT} \right). \tag{8}$$

This average is used for fold-level model selection and for the overall comparison between SegResNet, MedNeXt, and CoMNeT.

Setting	Value
Segmentation backbone	3D MedNeXt-B with GRN
MedNeXt loss	DSC++ with binary cross-entropy
Deep supervision	Enabled for MedNeXt outputs
Input channels	4
Output channels	3
Patch size	$128 \times 160 \times 112$
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Batch size	1
Gradient accumulation	2 steps
MedNeXt maximum epochs	150
MedNeXt early stopping patience	30 epochs
Diffusion component	CorrDiff residual refinement
CorrDiff maximum epochs	50
CorrDiff early stopping patience	15 epochs
CorrDiff diffusion schedule	200 training timesteps; 25 reverse steps during validation/inference
CorrDiff loss weights	$\lambda_{\text{diff}} = 1$, $\lambda_{\text{recon}} = 1$, $\lambda_{\text{vq}} = 0.25$, $\lambda_{\text{seg}} = 1$
Ensemble size	5-fold

Table 1. Core training settings used for CoMNeT.

Implementation

We build the pipeline in Python using PyTorch for model training. TorchIO is used for medical image preprocessing and augmentation, MONAI utilities support medical imaging workflows and baseline losses, and NumPy and SciPy are used for numerical operations and metric computation. We use CorrDiff as a VQ-VAE and DDPM-based residual refiner. After thresholding, we apply lightweight morphological postprocessing to remove small isolated components and fill holes in each tumor-region channel. Experiments are designed for GPU acceleration when CUDA is available.

4 Results and discussion

We evaluate SegResNet, MedNeXt, and the proposed CoMNeT pipeline on the UTSW-Glioma dataset using five-fold cross-validation. Table 2 summarizes the Dice similarity coefficient for ET, TC, WT, and the region-wise average across all folds. CoMNeT achieves the highest average Dice score, with an average performance of 0.7798 ± 0.0184 across folds. This corresponds to an absolute Dice improvement of 0.0243 over SegResNet and 0.0101 over baseline MedNeXt. CoMNeT also achieves the highest Dice score for both the ET and WT regions.

Dataset	Model	ET	TC	WT	Average
UTSW-Glioma	SegResNet	0.7459 ± 0.0263	0.6717 ± 0.0170	0.8489 ± 0.0131	0.7555 ± 0.0190
UTSW-Glioma	MedNeXt	0.7207 ± 0.0312	$\mathbf{0.6847 \pm 0.0258}$	0.9037 ± 0.0156	0.7697 ± 0.0154
UTSW-Glioma	CoMNeT (proposed)	$\mathbf{0.7543 \pm 0.0261}$	0.6806 ± 0.0166	$\mathbf{0.9049 \pm 0.0128}$	$\mathbf{0.7798 \pm 0.0184}$

Table 2. Five-fold cross-validation segmentation performance on UTSW-Glioma. Values are reported as mean ± standard deviation across folds for enhancing tumor (ET), tumor core (TC), whole tumor (WT), and the region-wise average. Bold values indicate the best performance within the dataset.

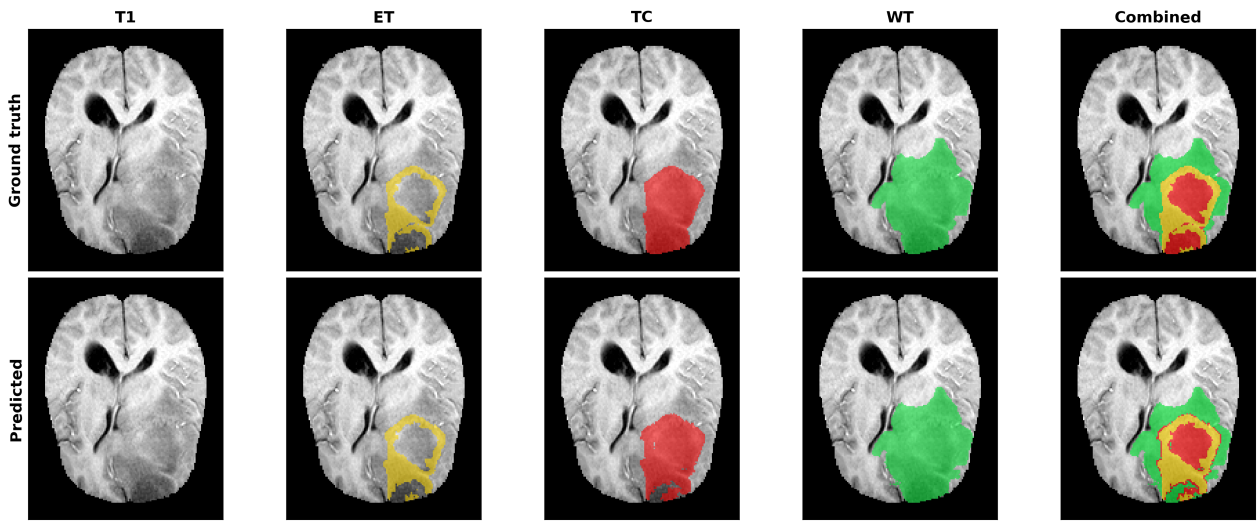

Figure 2. Representative CoMNeT segmentation result for a UTSW-Glioma patient. The top row displays the ground-truth segmentation and the bottom row displays the corresponding prediction from the proposed framework.

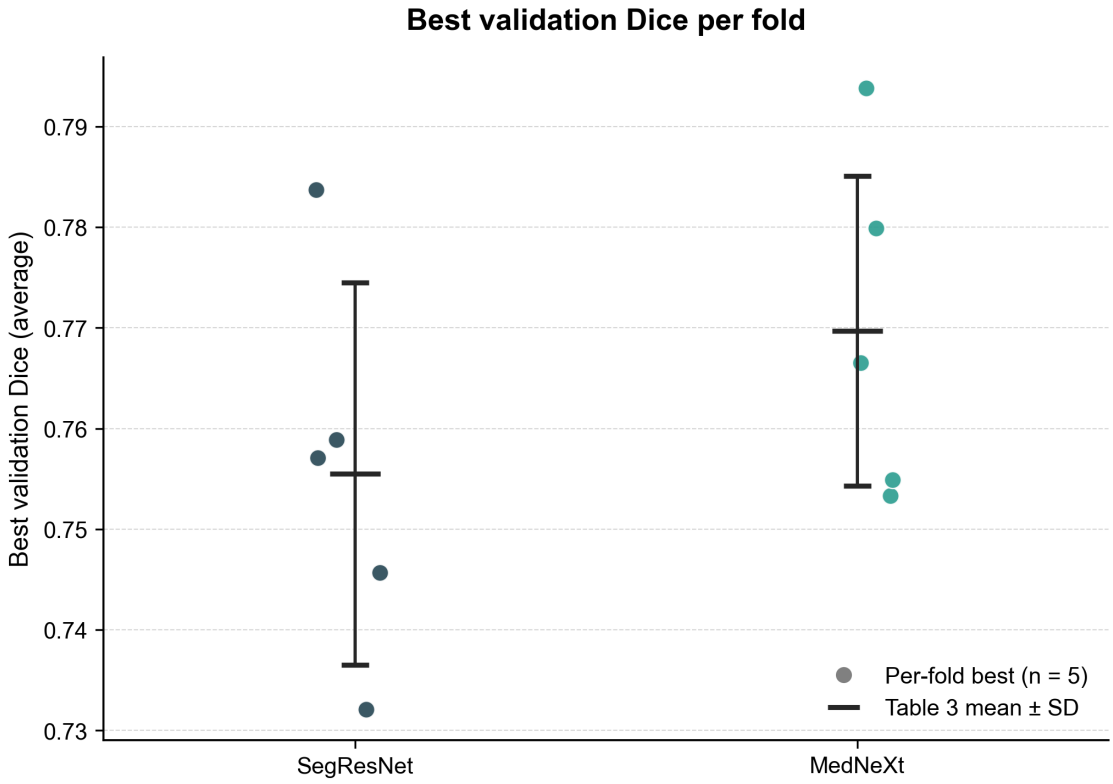

Figure 3. Best validation average Dice for each fold on the UTSW-Glioma dataset. Each point represents the best average Dice from one validation fold, and the horizontal summary markers show the mean ± standard deviation reported in Table 2.

In Figure 4, we show the same comparison by tumor region. SegResNet achieves a higher ET Dice score than MedNeXt, but lower TC and WT performance. Baseline MedNeXt improves WT segmentation substantially, reaching 0.9037 ± 0.0156, but its ET Dice is lower than the SegResNet baseline. Our proposed method improves ET and WT simultaneously over both baseline models, but with a slightly lower TC value compared to MedNeXt. The largest absolute performance increase over SegResNet is for WT, while the largest gain over MedNeXt is observed for ET. The ET and TC regions are usually more

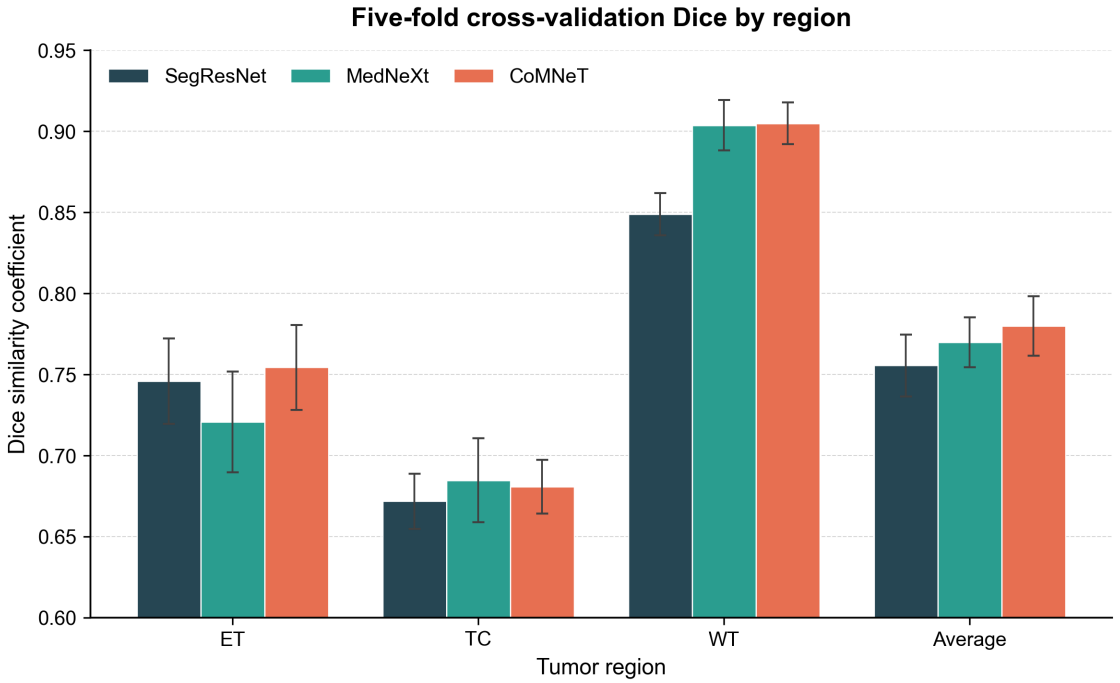

Figure 4. Region-wise Dice comparison for five-fold cross-validation on the UTSW-Glioma dataset. Bars show the mean Dice score for each model, and error bars show the standard deviation across folds. CoMNeT achieves the highest Dice score for ET, WT, and the region-wise average.

difficult than WT due to their smaller size and greater dependence on contrast enhancement, necrosis, and non-enhancing tumor appearance.

Figure 3 displays the best validation average Dice obtained in each fold for both baseline models. The fold-level distribution follows the same trend as Table 2. SegResNet has the lowest mean average Dice and shows wider spread across folds. MedNeXt displays an increase in fold-level performance, but some individual folds still overlap within the SegResNet range, supporting our decision to use MedNeXt as the segmentation backbone. Our pipeline produces the highest mean fold-level distribution, indicating that CoMNeT maintains stable training performance across validation folds.

A qualitative example comparing the CoMNeT prediction with the corresponding ground-truth segmentation is shown in Figure 2. The predicted WT mask closely follows the overall extent of the tumor, and the ET mask captures the main enhancing component with a similar spatial pattern to the ground truth. However, the TC prediction is visually less accurate, with more noticeable boundary differences and local mismatch relative to the ground truth. This qualitative pattern is consistent with Table 2, where TC is the most difficult region for CoMNeT, in which MedNeXt slightly outperforms our proposed method. However, the prediction closely follows the WT and ET regions.

These results suggest that our proposed correction and ensembling strategy improves volumetric brain tumor segmentation over a baseline segmentation backbone. The standalone MedNeXt model improves the average Dice score over SegResNet, mainly through better WT segmentation. This is consistent with the expected advantage of a modern convolutional architecture, since the larger contextual representation helps identify the full tumor context. However, MedNeXt alone underperforms SegResNet for ET, which suggests that the backbone can still struggle with smaller enhancing regions. CoMNeT manages to improve ET while preserving the WT performance from MedNeXt.

The improvement in ET is clinically meaningful because enhancing tumor is often the most visually distinct but also one of the most spatially limited tumor regions. Small false negatives for this region can substantially lower the Dice score. Similarly, small false positives around enhancement-like patterns can also reduce the accuracy of the segmentation. Corrective diffusion is useful for this task, by learning to correct residual errors from the MedNeXt probability maps. This allows the diffusion model to perform a more focused task of refining uncertain or systematically incorrect regions. The improvement in WT also supports the use of corrective refinement.

For the task of segmenting the whole tumor, MedNeXt already performs well with a Dice score of 0.9037 ± 0.0156. CoMNeT increases WT further to 0.9150 ± 0.0128. This improvement is smaller in absolute terms because WT performance is already high, but the reduced standard deviation suggests that our framework also stabilizes whole-tumor segmentation across

folds. Fold-level ensembling also contributes to the final segmentation prediction. Each cross-validation fold produces a unique model trained on a different subset of the data. By averaging the corrected probability maps before thresholding, we can reduce the effect of fold-specific variance and produces smoother masks. This is especially important in a small medical imaging dataset, where each fold may learn relatively substantial differences in the tumor boundaries of difficult cases. Figure 4 shows that CoMNeT has the strongest region-averaged validation Dice score, which supports the use of fold-level ensembling as part of the final pipeline.

In this section, we bring to attention the existing limitations of our study. First, we evaluate a single institutional dataset. UTSW-Glioma is useful because it is well curated and includes expert segmentation masks, but our results do not yet support that CoMNeT generalizes to multiple scanner sites, patient populations, or tumor distributions. Moreover, our current work focuses primarily on Dice-based evaluation. Dice is appropriate for measuring overlap, but we should investigate more detailed boundary metrics, calibration metrics, and qualitative error analysis. Finally, the decrease in Dice score for the TC region compared to baseline MedNeXt should be explored.

5 Conclusion and future work

In this work, we propose CoMNeT, a MedNeXt-CorrDiff framework for volumetric brain tumor segmentation on the UTSW-Glioma dataset. Our method combines a MedNeXt segmentation backbone with GRN, fold-level probability ensembling, and a CorrDiff residual refinement post-processing block. In five-fold cross-validation, CoMNeT achieved the highest Dice score for ET and WT, and the region-wise average, outperforming both baseline models. These results support the use of model ensembling and corrective diffusion as a refinement step for an improved 3D segmentation method. Future work will focus on testing the generalizability of this method beyond the UTSW-Glioma dataset, and improving TC Dice score. We plan to evaluate CoMNeT on the BraTS-GoAT dataset to measure our framework's ability to adapt to changes in tumor type, lesion appearance, scanner setting, and patient population.

Acknowledgements

The authors acknowledge the contributors of the UTSW-Glioma dataset and The Cancer Imaging Archive for making the imaging data available.

Author contributions statement

M.L.E. contributed to the study design, MedNeXt development, CorrDiff integration, model training workflow, result analysis, figure generation, and manuscript preparation. M.F.B.H. contributed to preprocessing design, experimental design, model evaluation, and manuscript preparation. M.S.S., W.F., and K.M.I. contributed to study supervision, methodological guidance, and manuscript review. All authors reviewed the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Data availability: The UTSW-Glioma dataset is publicly available through The Cancer Imaging Archive according to the relevant dataset-access process. Derived predictions, trained model checkpoints, and analysis scripts are available from the corresponding author subject to data-use restrictions and institutional requirements.

References

1. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
2. Roy, S. *et al.* Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 405–415 (Springer, 2023).
3. Roy, S. *et al.* Mednext-v2: Scaling 3d convnexts for large-scale supervised representation learning in medical image segmentation. *arXiv preprint arXiv:2512.17774* (2025).
4. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 424–432 (Springer, 2016).

6. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571 (Ieee, 2016).

7. Myronenko, A. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, 311–320 (Springer, 2018).

8. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* **18**, 203–211 (2021).

9. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

10. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, 272–284 (Springer, 2021).

11. Wang, W. *et al.* Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, 109–119 (Springer, 2021).

12. Woo, S. *et al.* Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16133–16142 (2023).

13. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. neural information processing systems* **33**, 6840–6851 (2020).

14. Kazerouni, A. *et al.* Diffusion models in medical imaging: A comprehensive survey. *Med. image analysis* **88**, 102846 (2023).

15. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P. & Cattin, P. C. Diffusion models for implicit image segmentation ensembles. In *International conference on medical imaging with deep learning*, 1336–1348 (PMLR, 2022).

16. Wu, J. *et al.* Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical imaging with deep learning*, 1623–1639 (PMLR, 2024).

17. Li, W., Huang, W. & Zheng, Y. Corrdiff: Corrective diffusion model for accurate mri brain tumor segmentation. *IEEE J. Biomed. Heal. Informatics* **28**, 1587–1598 (2024).

18. Reddy, D. D. *et al.* Advancing brain tumor analysis: Curating a high-quality mri dataset for deep learning-based molecular marker profiling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2373–2379 (2024).

19. Reddy, D. D. *et al.* The university of texas southwestern glioma dataset-mri, molecular markers and segmentations. *Sci. Data* (2026).