bioengineering

MDPI

Article

Ultra-High Resolution 9.4T Brain MRI Segmentation via a Newly Engineered Multi-Scale Residual Nested U-Net with Gated Attention

Aryan Kalluvila [1,*], Jay B. Patel [2] and Jason M. Johnson [3]

[1] Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL 60201, USA
[2] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA 02129, USA; jbpatel@alum.mit.edu
[3] Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT 06520, USA; jason.johnson@yale.edu
* Correspondence: aryankalluvila2027@u.northwestern.edu

Abstract

A 9.4T brain MRI is the highest resolution MRI scanner in the public market. It offers submillimeter brain imaging with exceptional anatomical detail, making it one of the most powerful tools for detecting subtle structural changes associated with neurological conditions. Current segmentation models are optimized for lower-field MRI (1.5T–3T), and they struggle to perform well on 9.4T data. In this study, we present the GA-MS-UNet++, the world's first deep learning-based model specifically designed for 9.4T brain MRI segmentation. Our model integrates multi-scale residual blocks, gated skip connections, and spatial channel attention mechanisms to improve both local and global feature extraction. The model was trained and evaluated on 12 patients in the UltraCortex 9.4T dataset and benchmarked against four leading segmentation models (Attention U-Net, Nested U-Net, VDSR, and R2UNet). The GA-MS-UNet++ achieved a state-of-the-art performance across both evaluation sets. When tested against manual, radiologist-reviewed ground truth masks, the model achieved a Dice score of 0.93. On a separate test set using SynthSeg-generated masks as the ground truth, the Dice score was 0.89. Across both evaluations, the model achieved an overall accuracy of 97.29%, precision of 90.02%, and recall of 94.00%. Statistical validation using the Wilcoxon signed-rank test ($p < 1 \times 10^{-5}$) and Kruskal–Wallis test (H = 26,281.98, $p < 1 \times 10^{-5}$) confirmed the significance of these results. Qualitative comparisons also showed a near-exact alignment with ground truth masks, particularly in areas such as the ventricles and gray–white matter interfaces. Volumetric validation further demonstrated a high correlation ($R^2 = 0.90$) between the predicted and ground truth brain volumes. Despite the limited annotated data, the GA-MS-UNet++ maintained a strong performance and has the potential for clinical use. This algorithm represents the first publicly available segmentation model for 9.4T imaging, providing a powerful tool for high-resolution brain segmentation and driving progress in automated neuroimaging analysis.

Keywords: magnetic resonance imaging (MRI); GA-MS-UNet++ (Gated Attention, Multi-Scale Residual U-Net++); deep learning (DL); machine learning (ML)

Academic Editor: Hatem Alhadainy

Received: 9 August 2025
Revised: 16 September 2025
Accepted: 22 September 2025
Published: 24 September 2025

Citation: Kalluvila, A.; Patel, J.B.; Johnson, J.M. Ultra-High Resolution 9.4T Brain MRI Segmentation via a Newly Engineered Multi-Scale Residual Nested U-Net with Gated Attention. *Bioengineering* **2025**, *12*, 1014. https://doi.org/10.3390/bioengineering12101014

1. Introduction

Magnetic resonance imaging is a critical tool for characterizing brain anatomy and pathology. It offers important insight into the structural and functional aspects of several

neurological conditions [1]. Accurate segmentation of the brain structures in MRI helps physicians gain critical insights into pathological changes associated with severe neurological conditions. Historically, there have been two leading players in the MRI space: 1.5T and 3T [2]. Large public databases, such as the Human Connectome Project, focus solely on these two sequence types, leading researchers to concentrate on creating algorithms for just those two [3]. Recently, however, ultra-high-field MRI scanners, particularly those operating at 7T and above, have revolutionized the field of brain imaging. They offer significantly better signal-to-noise ratio (SNR) and intensity contrasts, as well as improved visualization of fine anatomical details and better detection of small abnormalities and lesions, such as those seen in multiple sclerosis [4,5]. Unfortunately, high-field scanners are prohibitively costly, and as such, very few hospitals in the United States have been able to deploy 9.4T MRI scanners [6]. These scanners, however, offer unprecedented spatial resolutions down to submillimeter levels (0.6–0.8 mm isotropic) and improve tissue contrast. Existing segmentation models struggle to cope with the 9.4T data because they were trained on lower resolution imaging [7].

1.1. Existing Deep Learning Frameworks

Current state-of-the-art deep learning models include the U-Net, Attention U-Net, Very Deep Super Resolution model (VDSR), R2UNet, and Nested U-Net architecture. The U-Net was the first breakthrough architecture in biomedical segmentation. Proposed by Ronneberger et al., this approach employed a symmetrical encoder and decoder structure with skip connections, preserving spatial information from previous layers [8]. Variations such as the 3D U-Net extend this basic architecture into three dimensions but struggle with GPU efficiency [9]. The Attention U-Net is a good example of how components of the original U-Net can be enhanced to better address specific tasks. Researchers dynamically weighted different regions of the image and focused their computational resources on contrast and anatomical structures [10]. This approach significantly improved the model's ability to delineate subtle boundaries. The Very Deep Super Resolution (VDSR) model was proposed by Kim et al., and it introduced a deep convolutional architecture for single-image super resolution that uses residual learning to help improve training and accuracy. Its feature extraction power has inspired the use of it for segmentation purposes [11]. The R2UNet architecture, proposed by Alom et al., combines the strengths of both recurrent residual convolutional layers with the standard U-Net architecture [12]. Integrating residual units and recurrent connections is a powerful way the R2UNet improves feature representation. It also promotes strong segmentation performance, especially in the cases of complex biomedical images (such as 9.4T). The Nested U-Net, also known as UNet++, further extends the U-Net architecture by utilizing nested and dense skip connections between the encoder and decoder pathways, offering superior accuracy for biomedical segmentation tasks, particularly in brain MRI segmentation [13].

1.2. Volumetry in Brain MRI

Quantitative volumetric analysis aids in the early detection of neurodegenerative diseases. For instance, reduced hippocampal volume is a clear characteristic of Alzheimer's disease and can be one of the first indications prior to cognitive decline [14]. Similarly, changes in brain volume can help differentiate between conditions such as Alzheimer's disease, Parkinson's disease, and another behavioral variant of frontotemporal dementia. In multiple sclerosis (MS) or other diseases involving white matter lesions, brain volume loss correlates with a substantial decline in cognitive ability and disability, making volumetric assessment crucial for evaluating treatment efficacy [15,16]. The integration of automated volumetric analysis tools, such as NeuroQuant, has further revolutionized clinical practice

by providing rapid and reproducible measurements, reducing inter-rater variability, and enhancing diagnostic accuracy [17]. These advancements have made brain volumetry a standard component in the evaluation of patients with cognitive impairments. As mentioned previously, the ultra-high resolution provided by the 9.4T MRI enables the detection of subtle anatomical changes that are often undetectable at lower field strengths. This increased sensitivity enables more precise measurements of brain volumes, providing even greater specificity in detecting neurodegenerative disorders.

Similarly, volumetric analysis plays a pivotal role in the clinical workflow for tumor patients [18,19]. Automated segmentation pipelines using conventional MR such as T1 and T2 imaging have been used in a wide range of applications, including for primary, metastatic, and pediatric brain tumors [20–24]. More recently, work has been performed using less common MR modalities. For instance, a recent systematic review on computer-aided diagnosis using hyperspectral imaging for brain cancer reported a high diagnostic performance despite small datasets and limited external validation [25]. Finally, even though we focus on brain MRI in this manuscript, we note that automated deep learning-based pipelines have use in a diverse range of applications, including but not limited to cardiac MRI, abdominal CT, retinal fundus imaging, and chest X-rays [26–29].

1.3. Clinical Importance of 9.4T

The 9.4T MRI offers ultra-high resolution in brain imaging by capturing a submillimeter resolution [30]. This ultra-high resolution enables clinicians and researchers to identify subtle abnormalities that may have been missed by lower-strength scanners, such as 1.5T and 3T. The enhanced tissue contrast has been shown to improve the detection of small lesions, which is particularly important for the early diagnosis of conditions. Additionally, 9.4T provides stronger and more accurate mapping of white matter pathways, aiding in the understanding of brain connectivity and neurological disorders. The increased sensitivity also enables better functional imaging properties, providing clearer insights into brain activity. Although primarily used for research purposes currently, we hope that, by developing robust segmentation models for 9.4T data, its potential can be realized and further applied in clinical practice. As technology advances and accessibility improves, this tool is likely to become a key component of clinical imaging, helping to detect neurological conditions earlier and support more targeted treatments.

2. Materials and Methods

2.1. Study Design

This study utilizes MRI data that was collected from the UltraCortex dataset [30]. It utilized brain MRI scans at ultra-high resolution, obtained at 9.4T by the Max Planck Institute for Biological Cybernetics. A total of 78 subjects whose MRI scans were analyzed for cortical segmentation were included, comprising data from 78 healthy adult volunteers (28 females and 50 males, aged between 20 and 53 years). The demographic data were collected in accordance with the guidelines and ethical standards of the institution, and all participants provided written informed consent, which the relevant ethics committees of the University Hospital, Tübingen, Germany, approved. Specific subjects were excluded based on the following criteria: (1) known neurological or psychiatric diseases, (2) contraindications to MRI, (3) abnormal vision, and (4) severe motion artifacts or failed technical validation of images.

2.2. Image Acquisition

The MRI scans were obtained using a whole-body 9.4T MRI scanner (produced by Siemens Healthineers, Malvern, PA, USA) at the Department of High Field Magnetic

Resonance in Malvern, PA. These scans employed T1-weighted MP-RAGE and MP2RAGE sequences with spatial resolutions ranging from 0.6 to 0.8 mm, and the total dataset comprised 86 T1-weighted images from 78 unique subjects. For each participant, multiple acquisitions were performed to account for motion artifacts and ensure the most precise sequence was obtained for analysis.

2.3. Dataset Preprocessing

MRI images were stored in NIfTI format, and each scan and respective mask were accessed using the nibabel library. Out of the 78 total subjects, only 12 subjects had associated manual segmentation masks that could be utilized for training and inference. The segmentation masks were binary and represented delineations between white and gray matter boundaries (not including cerebrospinal fluid). This subset was randomly split into eight subjects for training and four subjects for testing. The image and mask paths were stored, with each MRI scan being loaded slice by slice (in a 2D fashion) to promote efficient memory usage. The images were normalized by subtracting the mean and dividing by the standard deviation to ensure consistent scaling across the dataset.

To further improve the model's robustness, data augmentation was applied during training. Random horizontal flips and rotations were applied to both the images and the masks to ensure that the model learned to generalize across various orientations. These augmentations were only applied to the training set, whereas the testing set remained unchanged to ensure valid evaluation. Each axial slice was resized to 256 $\times$ 256 pixels using bicubic interpolation. This resizing was applied in the image (pixel) space rather than the physical (millimeter) space. Given that the original voxel spacing ranged from 0.6 mm to 0.8 mm, the resulting voxel spacing after resizing is proportionally adjusted relative to the original dimensions of each image volume. They were converted into PyTorch tensors (version 2.7.1) and returned alongside their corresponding masks. This pipeline ensured that the MRI images and masks were processed correctly and augmented, providing high-quality data for training and testing the models.

While incorporating the complete 3D volume could potentially enhance anatomical consistency, GPU memory limitations restricted us from inputting entire volumes into the model. Furthermore, our goal was to develop a fast, easily deployable model tailored explicitly for 9.4T brain MRI. Therefore, we opted to work with 2D axial slices instead. Even though central axial slices of the brain have a more rich anatomical detail than peripheral slices, we did not find any benefit in using a slice sampling scheme weighted towards central slices. Thus, in our implementation, every available axial slice (including peripheral slices with little to no ground truth segmentation present) from each subject's T1-weighted volume was used for training.

On the 74 patients which did not have manual segmentations, we utilize SynthSeg to generate automatic segmentations of the target region [31]. We note that SynthSeg only runs on 1 mm isotropic resolution data. Thus, we first downsample our ultra high-resolution images to 1 mm isotropic to use with SynthSeg, and subsequently upsample the output segmentations via nearest-neighbor interpolation back to the original imaging resolution, which is between 0.6 and 0.8 mm. These outputs served as a reference point for validating our model's performance in the absence of extensive manual ground truth annotations. The SynthSeg-generated ground truth labels were also produced in the same format as the manual ground truth labels provided with the dataset, ensuring consistency in structure and compatibility. To further validate their suitability, all generated segmentations were subjected to a visual quality control step, during which we manually inspected the outputs to confirm anatomical plausibility and to ensure that no gross errors or artifacts were present.

2.4. Neural Network Engineering

For the deep learning portion of this study, we expanded on the traditional Nested U-Net architecture. It integrates multiple scales for multi-feature learning and has demonstrated exceptional performance in brain MRI segmentation tasks, as it can handle both fine-grained and large-scale anatomical details. In this paper, we introduce the Gated Multi-Scale Nested U-Net (GA-MS-UNet++) (Figure 1). This is an improvement over the standard UNet++ with three key innovations: (1) multi-scale residual blocks, (2) attention mechanisms, and (3) gated skip connections. The architecture still follows the encoder–decoder structure, where the encoder path progressively downsamples the input using stacked Multi-Scale Blocks (MSBlocks).

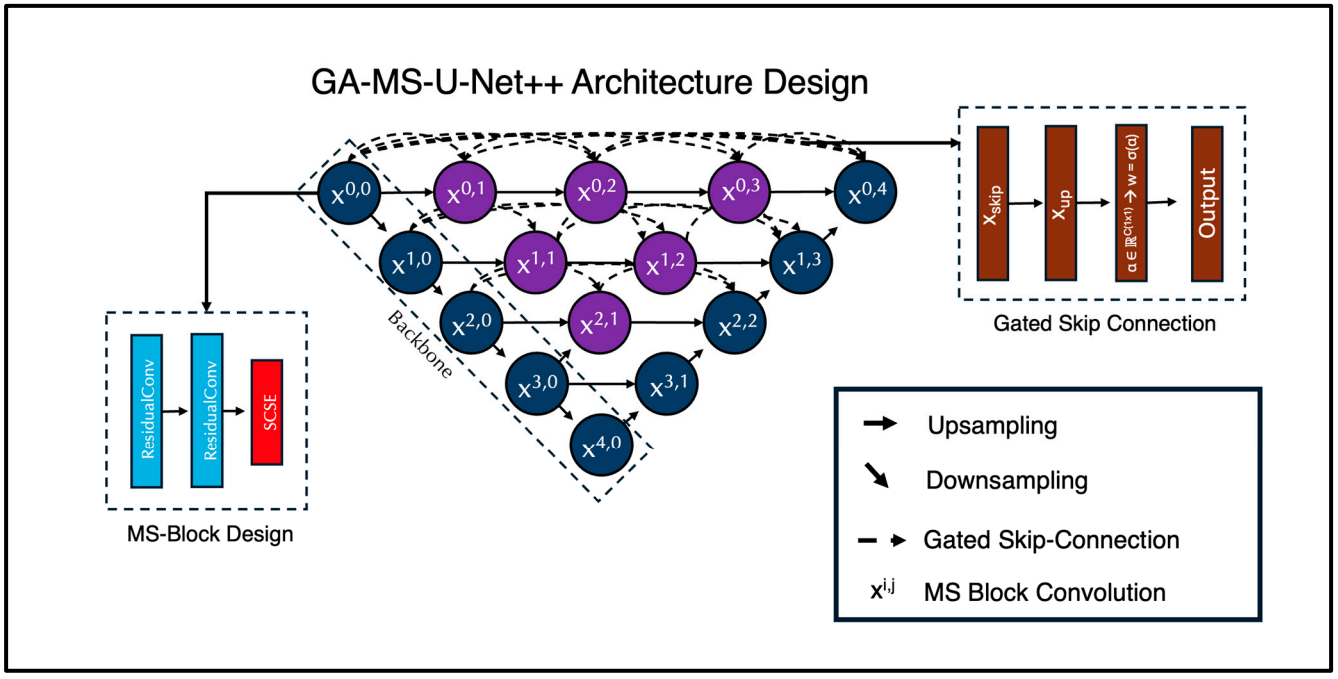


Figure 1. Proposed GA-MS-UNet++ Architecture with Multi-Scale-Block and Gated Skip Connection designs mapped out. Our novel architecture includes these components in order to bolster performance and improve efficiency of the model. We modified the original Nested U-Net architecture by injecting these components within the skip connections and the standard convolutions.

Each MSBlock contains two Residual Convolution Units (RCUs) and one Spatial Channel Squeeze and Excitation (SCSE) attention mechanism. Each RCU is composed of two convolutional layers with kernel size 3. We modify the second convolutional layer to use a dilation value of 2, broadening the receptive field of the network without increasing the number of trainable parameters. This multi-scale context aggregation enables the network to capture additional fine details and global structure. We utilize group normalization (GN) instead of BatchNorm (to stabilize training on small batches) and Leaky ReLU activation functions. A residual connection is applied by adding the input to the output of the convolutional sequences. If the input and output channels differ, a 1×1 convolution is used to align the dimensions. Following the two RCU layers, a Spatial Channel Squeeze and Excitation (SCSE) attention mechanism is used to recalibrate feature maps. The channel attention part uses global average pooling to help condense each channel into a scalar, then passes this vector through 1×1 convolutions with a reduction ratio r = 4 and applies a sigmoid activation to produce channel-wise attention weights. Mathematically, it is represented as follows:

$$cSE(x) = x \cdot \sigma(W2 \cdot ReLU(W1 \cdot GAP(x)))$$

Here, *GAP(X)* performs the global pooling average, and W1 and W2 are the learnable weights of the convolutions. The spatial attention then applies the 1×1 convolution to produce a single spatial attention map:

$$sSE(x) = x \cdot \sigma(Ws * x)$$

The final output of the SCSE block is the sum of the added tensors:

$$SCSE(x) = cSE(x) + sSE(x)$$

After each set of two MSBlocks per layer, downsampling in the encoder is handled via a 2×2 max pooling operation. The decoder is symmetrical to the encoder, using two MSBlocks per layer to reinforce multi-scale context for all stages of reconstruction. Upsampling in the decoder is handled via a bilinear interpolation operation.

We leveraged the nested skip design path from the regular Nested U-Net, which generates multiple intermediate decoder outputs. However, unlike the traditional UNet++, which concatenates skip features, the GA-MS-UNet++ utilizes gated skip connections to help learn how much information to retain from the encoder features versus the upsampled decoder features. Each gate has a learnable parameter, which is initialized to zero, and the skip connection is computed as follows:

$$GatedSkip(xskip, xup) = \sigma(\alpha) \cdot xskip + (1 - \sigma(\alpha)) \cdot Conv1x1(xup)$$

Here, $\sigma(\alpha)$ generates a per-channel attention mask that facilitates the blending of the encoder and decoder signals. The final segmentation map is produced by applying a 1×1 convolution to the final decoder output.

Each skip connection pathway had a learnable gate parameter that determined how much information to pass from the encoder feature map versus the upsampled decoder feature. These gates were initialized to 0.5, ensuring that, at the start of training, the contribution from the encoder and decoder pathways was balanced equally. The gate parameters were then optimized with the rest of the network using backpropagation, allowing the model to learn whether to favor encoder features, decoder features, or a mixture of both.

2.5. Training Parameters

A batch size of 1 was used due to memory constraints and the large size of the input images, with training performed for 100 epochs on a NVIDIA H100 GPU. The model was optimized using the Adam optimizer, with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}, minimizing binary cross-entropy loss between the predicted masks and the ground truth. The predictions were thresholded at 0.5 to generate the binary segmentation outputs. After training, the model's performance was evaluated using the Dice coefficient, precision, accuracy, recall, and Structural Similarity Index Measure (SSIM). These metrics were computed from both the raw predictions and their binarized forms to assess both pixel-wise accuracy and perceptual quality.

As an additional validation step, we also compared all 12 manually annotated MRI scans against their corresponding SynthSeg-generated segmentations. This comparison produced an average Dice coefficient of 0.9480 with a standard deviation of 0.0050, which demonstrated a strong concordance between the automated and expert-derived labels.

3. Results

3.1. Dice and SSIM Evaluation

The first experiment we performed was a quantitative analysis using Dice and SSIM to determine the segmentation accuracy across the models. We evaluate our model and baselines using 4 subjects with manual segmentations and 78 subjects with SynthSeg segmentations. To ensure appropriate fairness and relevance, we excluded any slice with a Dice or SSIM value of precisely 0 or 1, as these typically correspond to empty or non-informative slices (common at the beginning or end of MRI volumes). After this filtering, we analyzed the remaining slices for each subject. Dice and SSIM were both included in this study, as Dice captures the overlap, whereas SSIM focuses on the entire image quality. We split the SSIM and DICE results into two categories: manual segmentation ground truth (Figure 2) and SynthSeg segmentation ground truth (Figure 3).

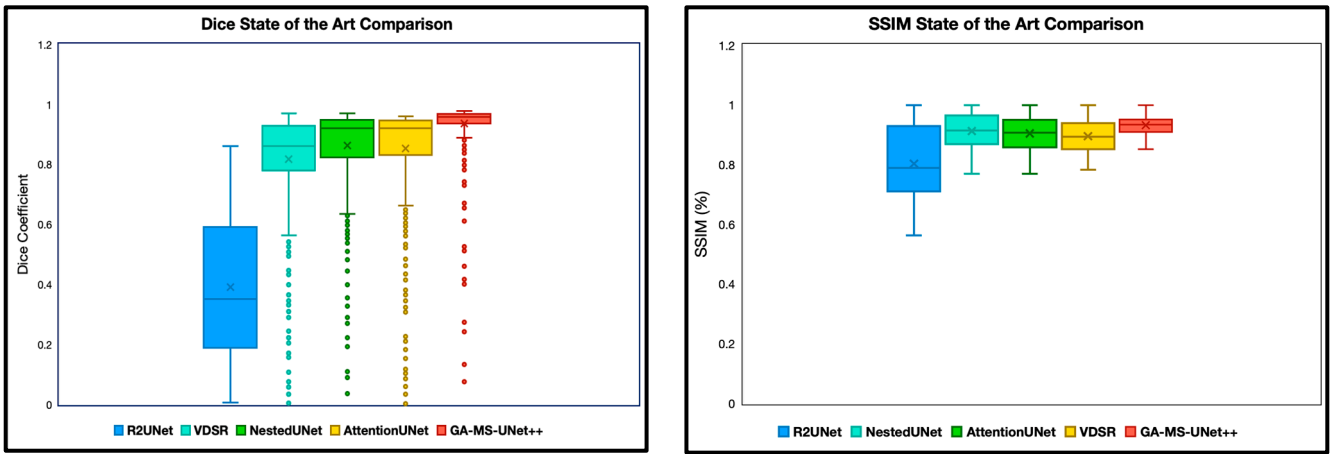


Figure 2. Dice coefficient and SSIM comparison of segmentation performance across five deep learning models on 9.4T MRI scans using manual segmentation as ground truth. Dice scores quantify the spatial overlap between predicted segmentation masks and ground truth annotations, with higher values indicating better segmentation accuracy. SSIM scores were computed between predicted segmentations and ground truth masks, assessing perceptual similarity in terms of luminance. The X notation in between each box and whisker plot represents the mean. GA-MS-UNet++ consistently outperformed the baseline models—including Nested U-Net, Attention U-Net, R2UNet, and VDSR—demonstrating the highest mean Dice score and lowest variance across evaluated slices. These results highlight the model's robustness in accurately delineating brain structures at ultra-high resolution.

3.2. Whole Brain Volumetry

For the volumetry experiment, we evaluated the accuracy of whole-brain volume predictions generated solely by the GA-MS-UNet++ across the four subjects (Figure 4) Each subject's binarized predicted segmentation output was calculated by counting the nonzero pixels and converting them into cubic millimeters (mm^3). These predicted volumes were then compared to the ground truth manual segmentations. As shown in the side-by-side bar graph, the GA-MS-UNet++ produced volumes that were close to the ground truth for all four cases. To quantify this relationship further, we performed a regression analysis, yielding a high correlation coefficient of 0.9092, indicating a strong agreement between the predicted and ground truth volumes.

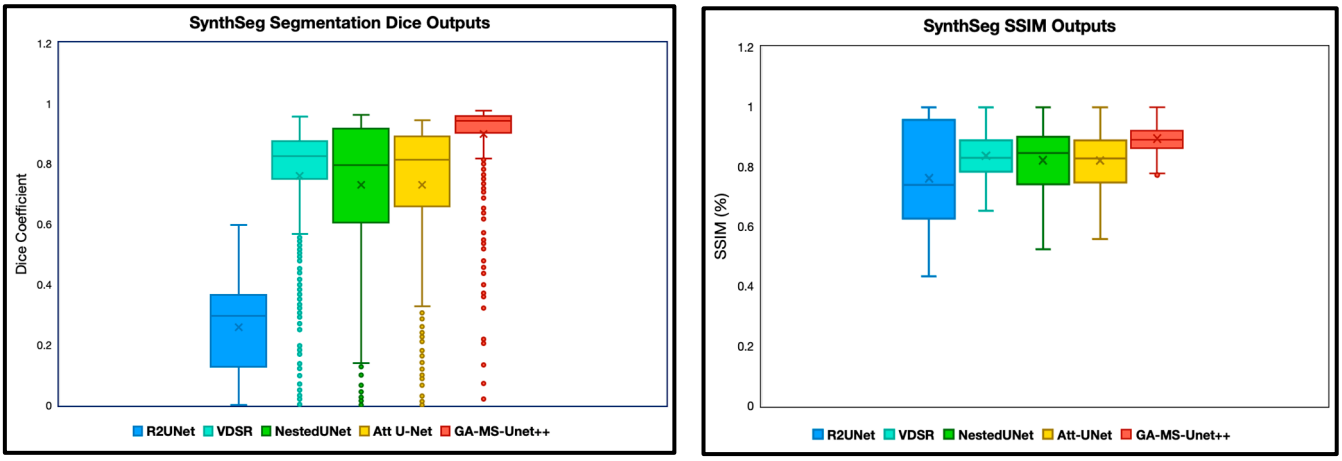
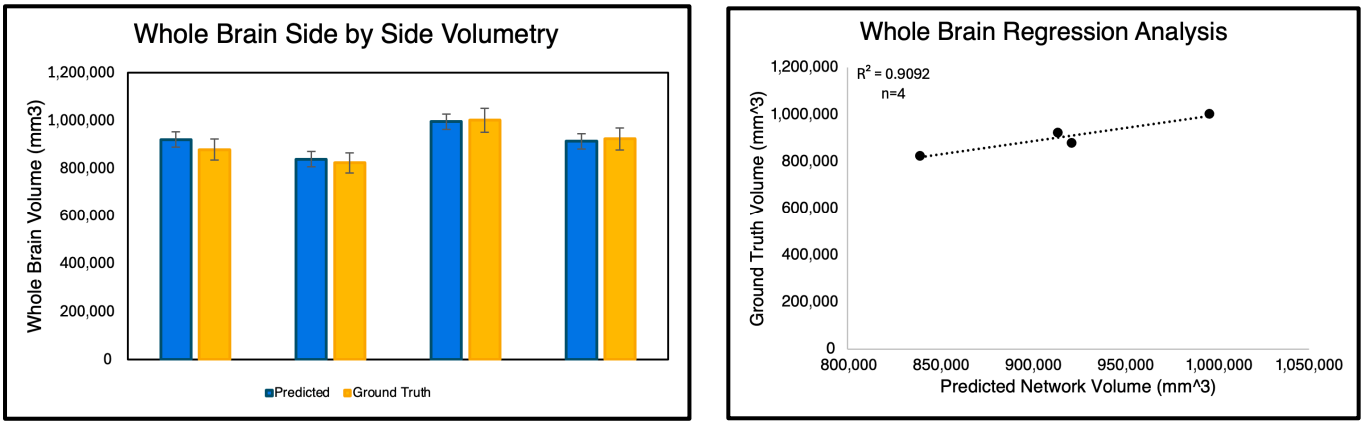


Figure 3. Dice and SSIM comparison across five segmentation models using 9.4T MRI data but with SynthSeg labels as ground truth. Models were evaluated on the remaining scans without manual segmentation ground truth labels. The X notation in between each box and whisker plot represents the mean.



(**A**) (**B**)

Figure 4. Comparison of predicted and ground truth whole-brain volumes across four manually segmented 9.4T MRI subjects. The left panel (**A**) displays a side-by-side bar graph showing the predicted brain volumes generated by GA-MS-UNet++ and the corresponding manual segmentation ground truth in cubic millimeters (mm^3) for each subject. The right panel (**B**) presents a regression analysis of predicted versus ground truth volumes, demonstrating a strong positive correlation ($R^2 = 0.9092$). These results confirm the volumetric accuracy and reliability of the GA-MS-UNet++ model in high-resolution brain segmentation.

3.3. Qualitative Analysis

To complement the quantitative evaluation, we display the outputs of our model on a random test subject. Figure 5 shows the binarized prediction from all methods for a single brain slice, illustrating that our method yields the most accurate segmentation results. Figure 6 provides a deeper dive into the output of the GA-MS-UNet++, highlighting the residual difference between the ground truth and the output.

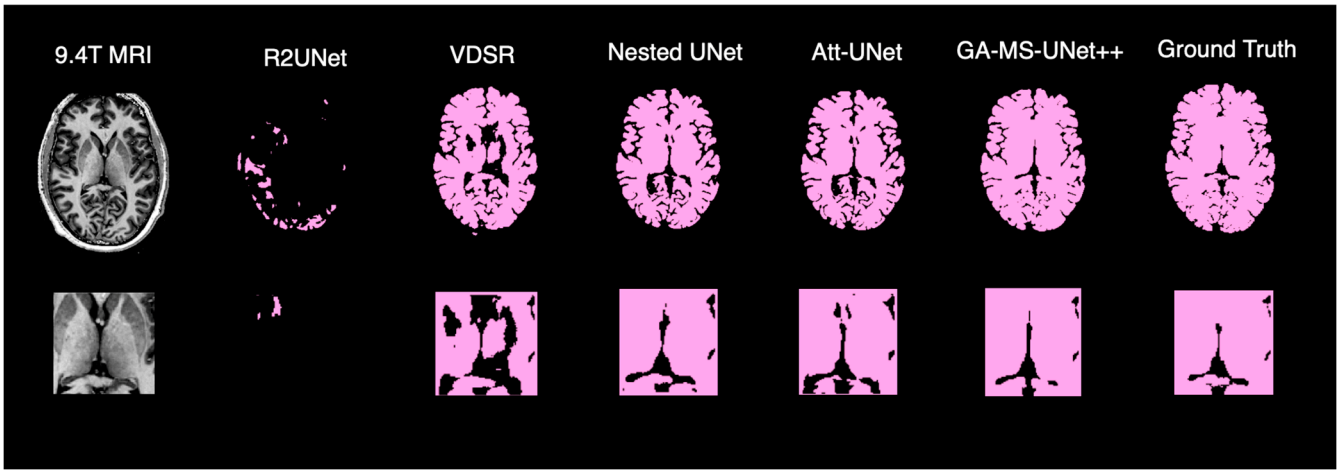


Figure 5. Qualitative comparison of brain segmentation outputs from five different models.

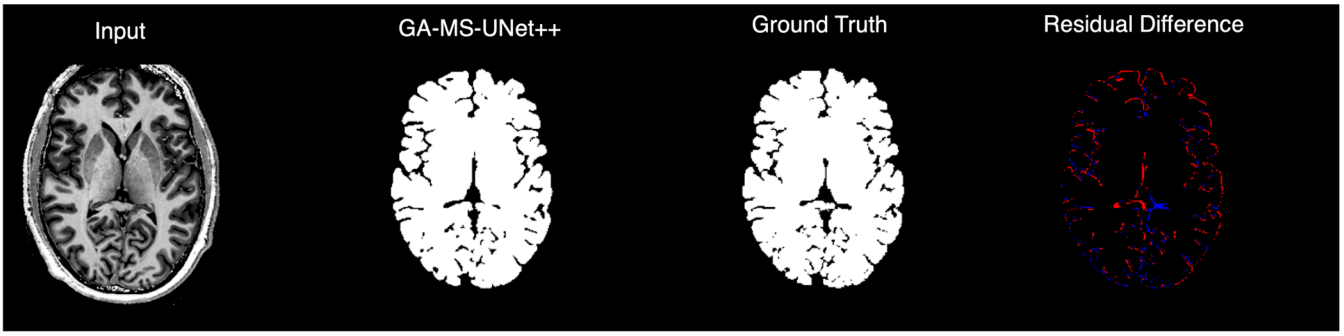


Figure 6. GA-MS-UNet++ segmentation compared to ground truth with residual error map visualization.

3.4. Statistical Analysis and Memory Efficiency

We evaluated segmentation performance using accuracy, precision, and recall to provide a more comprehensive assessment among the different error types (Table 1). These metrics were only calculated for the manually segmented images. We conducted a statistical analysis using nonparametric methods. First, a Kruskal–Wallis rank sum test was used to assess the overall differences in Dice scores across all the segmentation models and reference types, making no strict assumptions about the normality of the data. We also used a Wilcoxon signed-rank test for the pairwise comparison between our GA-MS-UNet++ model and each baseline segmentation model, a robust test for matched differences that does not require normal distributions (Table 2). We also reported each model's inference time, parameter count, memory footprint, and floating-point operations (FLOPs) to benchmark efficiency alongside accuracy (Table 3).

Table 1. Quantitative comparison of segmentation performance across models (manual segmentation output analysis, n = 4). Red-highlighted values represent the highest performing models in each class, respectively.

	R2UNet	VDSR	Nested U-Net	Attention U-Net	GA-MS-UNet++
Accuracy	0.8039	0.9081	0.9434	0.9305	0.9729
Precision	0.8877	0.9721	0.9692	0.9725	0.9002
Recall	0.2175	0.4406	0.6469	0.5525	0.9400

Table 2. Wilcoxon signed t-test for paired samples (SynthSeg output dice score analysis, n = 78).

GA-MS-UNet++	*p*-Value
R2UNet	$<1 \times 10^{-5}$
VDSR	$<1 \times 10^{-5}$
Nested U-Net	$<1 \times 10^{-5}$
Attention U-Net	$<1 \times 10^{-5}$

Shapiro–Wilk normality test: $p = 0.0000$ (not normal); Kruskal–Wallis test result, H-value: 26,281.98.

Table 3. Computational efficiency comparison of segmentation across models (manual segmentation output, n = 4).

	R2UNet	VDSR	Nested U-Net	Attention U-Net	GA-MS-UNet++
Inference (ms)	105.07	93.87	97.07	98.93	119.21
Parameters (M)	39.09	0.66	9.16	34.88	14.66
FLOPs	152.70	43.56	34.62	66.46	20.83
GPU Memory (GB)	0.36	0.07	0.20	0.31	0.24

Abbreviations: floating-point operations (FLOP).

As a benchmark, we compared model outputs and SynthSeg outputs against manual segmentations, confirming SynthSeg as a strong baseline with an overall Dice performance of 0.9192 ± 0.1601, while our model achieved 0.9311 ± 0.1710. Per-subject results showed consistent gains in three of four cases, supporting our hypothesis that our architectural modifications enhance segmentation quality.

To evaluate our model's generalization to conventional 3T MRI, we tested it on the open-source OASIS dataset [32], where it achieved a Dice score of 0.8651 ± 0.1523 on a set of 446 patients. These findings demonstrate that GA-MS-UNet++ not only performs robustly on ultra-high-field 9.4T data but also transfers effectively to widely used 3T MRI. This cross-field generalization is critical for broader clinical applicability.

4. Discussion

In this study, we introduce the GA-MS-UNet++ architecture for brain segmentation at an ultra-high field strength and resolution (9.4T). We conducted thorough statistical testing and rigorous comparisons against state-of-the-art algorithms to ensure our algorithm can be effectively utilized in clinical practice.

We first evaluated our model via Dice scores and SSIM. These evaluations were completed to ensure that the GA-MS-UNet++ performs competitively against other state-of-the-art models in standard quantitative metrics. Our model had a higher Dice score than any of the other four state-of-the-art models and, furthermore, had the smallest standard deviation and spread. The inconsistency of a few of the datapoints was mainly due to slices that were near the edges of the brain. These are challenging to produce accurate segmentations for and are of little use to radiologists.

GA-MS-UNet++ also outperformed the other state-of-the-art algorithms when evaluated via SSIM, albeit with a slightly higher spread. This is expected as SSIM values are generally constrained more. Thus, the other algorithms have a more substantial overlap with the proposed algorithms. SynthSeg segmentation ground truth experiments were performed in addition to the manual segmentation ground truth labels to expand the testing set. Similar results were demonstrated here, with the GA-MS_UNet++ having a higher average and lower spread than the other baselines for both the Dice score and SSIM.

While our model was solely trained on 2D axial slices, we recognize that using partial 3D context could help improve the spatial continuity in segmentation. The 2.5D strategies or overlapping slab-based approaches may help better capture the inter-slice relationships

without requiring full 3D volume processing. These methods help preserve anatomical context across the slices and can help enhance the boundary consistency within complex structures. However, the choice to use a 2D design was ultimately motivated by two factors: (1) memory efficiency and faster inference speed for ultra-high-resolution 9.4T data, and (2) clinical emphasis on deployability, where lightweight yet accurate models are favored. Still, future work could compare the GA-MS-UNet++ with 2.5D variants to see whether the spatial continuity gains justify the added computational complexity.

While this study focused on static 2D slices from the 9.4T MRI, the segmentation challenges are even more pronounced in temporally dynamic or multi-modal imaging data. Prior work on multi-modal, vision-based classification and sequence attention modeling suggests that attention-driven architectures like Transformer-based designs are well suited for capturing such dependencies [33]. The gated and residual mechanisms introduced in the GA-MS-UNet++ are thought to be a foundation for extending segmentation frameworks into these more complex domains.

As mentioned previously, brain volumetry calculations are crucial in determining whether an algorithm can be helpful in clinical settings, specifically for diagnosing neurological disorders. In our volumetry experiment, we aimed to compute the whole-brain volume of the ground truth segmentation output and the respective output of the GA-MS-UNet++. Since the algorithm outputs binary segmentation maps, the nonzero voxels were computed in mm^3. The regression results revealed an R^2 correlation value of 0.90, demonstrating strong agreement between the true volume and predicted volume for the four subjects with manual segmentations. It is important to note that the volumetric outputs of the study were accounted for when resizing to 256×256. Another key limitation of this study is the small number of patients with a manual ground truth. Therefore, we examined individual metrics to identify any outliers. Figure 4 shows each of the four subjects individually with their respective volume outputs. Across all four, there were no significant differences in volume. To ensure the reliability of our Dice score results, we conducted a paired Wilcoxon signed-rank test between our proposed GA-MS-UNet++ model segmentation Dice scores and those from the baseline approaches. The results clearly showed that the GA-MS-UNet++ performed significantly better in all pairwise comparisons, with p-values less than 1×10^{-5} (Table 2).

In our volumetric validation, GA-MS-UNet++ predictions demonstrated a strong correlation with ground truth brain volumes ($R^2 = 0.90$). However, this analysis was based on only four manually segmented subjects and should therefore be interpreted with caution. While the result provides encouraging preliminary evidence that a higher Dice overlap translates into reliable volumetric estimates, it does not, by itself, establish clinical applicability. To mitigate this limitation, we supplemented the analysis with two additional evaluations: (1) large-scale comparison against SynthSeg-derived segmentations across 78 subjects, where GA-MS-UNet++ maintained strong volumetric fidelity, and (2) external testing on the OASIS 3T dataset, which demonstrated generalization with a Dice score of 0.8651 ± 0.1523. Taken together, these complementary findings suggest that the model's volumetric accuracy is not limited to the small, manually annotated cohort, although larger studies with expert annotations will be necessary to confirm its role in the clinical monitoring of neurodegenerative disease.

Beyond just looking at technical performance numbers, we wanted to understand how better computer accuracy helps in real medical practice. Our volume measurements showed that GA-MS-UNet++ matched human expert measurements very closely (90% correlation), proving that when computers become better at identifying brain structures, the volume calculations become more trustworthy.

Accurate measurements of brain volume are crucial for tracking diseases that affect the brain. In Alzheimer's disease, certain brain areas like the hippocampus shrink before patients show obvious symptoms, so accurate measurements help with early detection. For multiple sclerosis patients, doctors track both brain lesions and overall brain shrinkage to monitor disease progression and see how well treatments are working.

Upon review, we also observed that the GA-MS-UNet++ demonstrates a slight tendency towards over-segmentation. This behavior explains the higher recall but comparatively lower precision. The model favors including ambiguous boundary regions rather than excluding them. Practically, this results in volumetric outputs that are marginally larger than the ground truth. This conservative bias is likely seen to avoid missing true positives. While this feature may inflate false positives, it also reduces the likelihood of underestimating subtle structures, which can be clinically advantageous in early disease detection.

To test generalizations from the 9.4T to the standard 3T MRI, we evaluated GA-MS-UNet++ on the OASIS dataset. Using SynthSeg to generate reference segmentations, the model achieved a mean Dice score of 0.8651 ± 0.1523, indicating that it can transfer effectively to lower-field data. To confirm that SynthSeg provided reliable pseudo-labels for this purpose, we also compared both SynthSeg outputs and model outputs against a subset of manually segmented cases. SynthSeg achieved a Dice score of 0.9192 ± 0.1601, while our model achieved 0.9311 ± 0.1710, supporting that SynthSeg is a strong baseline and that its labels are appropriate for use as the ground truth in OASIS. Together, these results show that GA-MS-UNet++ performs robustly across different field strengths and that SynthSeg-derived labels offer a practical way to extend evaluation to large external datasets.

Additionally, we applied the Kruskal–Wallis test to assess the differences across all of the models. The test produced an H-Value of 26,281.98, which strongly supports our assertion that models do not perform equally and that the GA-MS-UNet++ stands out in terms of segmentation accuracy. Given that the Dice score is one of the most important indicators of segmentation performance, these statistical results prove the effectiveness of our proposed model and its superiority over other approaches. We also investigated the computational efficiency of our model compared to other state-of-the-art models. We reported a speed of 119.21 m/s, which falls within the range of other algorithms in terms of the inference time to parameter ratio. The accuracy, precision, and recall scores of the model were above 90%, with accuracy and recall being the strongest among all models, and precision performing competitively.

Finally, we perform a qualitative assessment of the binarized segmentation masks from the GA-MS-UNet++ against other state-of-the-art models. Across the axial views, the GA-MS-UNet++ achieves a near-exact match with the ground truth and accurately delineates brain structures while minimizing false positives and false negatives. In contrast, the VDSR and R2UNet exhibit underestimation in key brain regions, such as the frontal lobe. While the Nested U-Net and Attention U-Net offer stronger results, they still exhibit noise and minor inaccuracies across tissue boundaries. When focusing on 9.4T segmentation accuracy, the ventricles and gray–white matter interfaces are preserved with the highest fidelity. The segmented output maintains sharp structural boundaries and captures the fine-scale anatomy and broader brain contours that are characteristic of the 9.4T resolution.

Despite the promising performance of the GA-MS-UNet++ architecture in 9.4T brain MRI segmentation, we recognize that this study has several important limitations. Firstly, the dataset size was limited, with only 12 manually annotated subjects used for training out of the total 86 available scans. This small sample size may compromise the model's generalizability, particularly when applied to a diverse population. While we tried to incorporate data augmentation, the model's performance should be validated on larger and more heterogeneous datasets. Secondly, the model was trained on 2D MRI slices

rather than on whole 3D volumes. While this reduces the computational demand for both training and inference, it may limit the model's ability to fully utilize the volumetric spatial context, which is crucial for segmenting complex brain structures. Finally, the dataset consisted of scans from a single 9.4T scanner using imaging parameters. This raises concerns about the model's adaptability to data from other ultra-high-resolution scanners or even lower-resolution scanners. External validation on these datasets would improve the model's confidence in clinical applicability. Addressing all three of these limitations in future research will be crucial for translating the GA-MS-UNet++ into reliable clinical tools for ultra-high-resolution brain MRI analysis.

5. Conclusions

In this study, we introduce a new algorithm called GA-MS-UNet++, a novel deep learning model specifically designed for high-accuracy brain MRI segmentation on ultra-high-field 9.4T MRI data. Our architecture integrates three main components: (1) multi-scale residual blocks, (2) gated skip connections, and (3) spatial channel attention mechanisms to capture both fine anatomical details and larger structural patterns. The model was ultimately evaluated on metrics such as the Dice score, accuracy, and SSIM, and consistently outperformed the four state-of-the-art segmentation models. Most importantly, the Dice scores were validated through rigorous testing and statistical analyses, including the Wilcoxon signed-rank tests ($p < 1 \times 10^{-5}$) and a Kruskal–Wallis test (H = 26,281.98), confirming that the GA-MS-UNet++ delivers statistically significant improvements in segmentation accuracy. Qualitative analysis further supports these findings. Compared to the ground truth, GA-MS-UNet++ achieved superior boundary outlining and anatomical resolution, particularly in the ventricles and gray–white matter interfaces. This level of accuracy is crucial, particularly in conditions where small volumetric assessments can indicate neurodegenerative diseases. Volumetric validation also demonstrated a high correlation ($R^2 = 0.90$) between the model's predictions and the ground truth, indicating the potential utility of this approach in monitoring neurodegenerative diseases. Despite there being only a few labeled 9.4T scans, our model performed robustly and maintained consistency across test subjects. Ultimately, as the 9.4T MRI becomes increasingly accessible in research and clinical settings, we hope our model can serve as a foundational tool in enabling precise and automated segmentation. In the future, we even hope our model will have clinical use cases, contributing to improved diagnostic accuracy, streamlined workflows, and better patient outcomes for ultra-high resolution neuroimaging applications.

Author Contributions: A.K. conceptualized the study, implemented the model, and wrote the manuscript. J.B.P. contributed to experimental design, code optimization, and revised the manuscript. J.M.J. supervised the project, assisted with statistical analysis, and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study involved a retrospective analysis of publicly available de-identified imaging data and does not meet the definition of human subject research. As such, it was not subject to institutional IRB review or approval by federal regulations (45 CFR 46).

Informed Consent Statement: Informed consent was obtained by the original investigators at the time of data collection. All data used in this study were de-identified and publicly available through the UltraCortex dataset.

Data Availability Statement: The UltraCortex dataset used in this study is publicly available at https://openneuro.org/datasets/ds005216/versions/1.0.0 (URL accessed on 1 September 2025).

Acknowledgments: The authors would like to acknowledge Northwestern University's Quest Computing System for allowing us to use their GPUs. We would also like to thank the peer reviewers for their time in providing feedback for our manuscript.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Marek, J.; Bachurska, D.; Wolak, T.; Borowiec, A.; Sajdek, M.; Maj, E. Quantitative brain volumetry in neurological disorders: From disease mechanisms to software solutions. *Pol. J. Radiol.* **2025**, *90*, e299–e306. [CrossRef] [PubMed] [PubMed Central]
2. Kabasawa, H. MR Imaging in the 21st Century: Technical Innovation over the First Two Decades. *Magn. Reson. Med. Sci.* **2022**, *21*, 71–82. [CrossRef] [PubMed] [PubMed Central]
3. Glasser, M.F.; Smith, S.M.; Marcus, D.S.; Andersson, J.L.; Auerbach, E.J.; Behrens, T.E.; Coalson, T.S.; Harms, M.P.; Jenkinson, M.; Moeller, S.; et al. The Human Connectome Project's neuroimaging approach. *Nat. Neurosci.* **2016**, *19*, 1175–1187. [CrossRef] [PubMed] [PubMed Central]
4. Balchandani, P.; Naidich, T.P. Ultra-High-Field MR Neuroimaging. *AJNR Am. J. Neuroradiol.* **2015**, *36*, 1204–1215. [CrossRef] [PubMed] [PubMed Central]
5. Ladd, M.E.; Bachert, P.; Meyerspeer, M.; Moser, E.; Nagel, A.M.; Norris, D.G.; Schmitter, S.; Speck, O.; Straub, S.; Zaiss, M. Pros and cons of ultra-high-field MRI/MRS for human application. *Prog. Nucl. Magn. Reson. Spectrosc.* **2018**, *109*, 1–50. [CrossRef] [PubMed]
6. Vaughan, T.; DelaBarre, L.; Snyder, C.; Tian, J.; Akgun, C.; Shrivastava, D.; Liu, W.; Olson, C.; Adriany, G.; Strupp, J.; et al. 9.4T human MRI: Preliminary results. *Magn. Reson. Med.* **2006**, *56*, 1274–1282. [CrossRef] [PubMed] [PubMed Central]
7. Donnay, C.; Dieckhaus, H.; Tsagkas, C.; Gaitán, M.I.; Beck, E.S.; Mullins, A.; Reich, D.S.; Nair, G. Pseudo-Label Assisted nnU-Net enables automatic segmentation of 7T MRI from a single acquisition. *Front. Neuroimaging* **2023**, *2*, 1252261. [CrossRef] [PubMed] [PubMed Central]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *arXiv* **2015**, arXiv:1505.04597. [CrossRef]
9. Du, W.; Yin, K.; Shi, J. Dimensionality Reduction Hybrid U-Net for Brain Extraction in Magnetic Resonance Imaging. *Brain Sci.* **2023**, *13*, 1549. [CrossRef] [PubMed] [PubMed Central]
10. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [CrossRef]
11. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks (Version 2). *arXiv* **2016**, arXiv:1511.04587. [CrossRef]
12. Alom, M.Z.; Yakopcic, C.; Hasan, M.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* **2019**, *6*, 014006. [CrossRef] [PubMed] [PubMed Central]
13. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. [CrossRef] [PubMed] [PubMed Central]
14. Schuff, N.; Woerner, N.; Boreta, L.; Kornfield, T.; Shaw, L.M.; Trojanowski, J.Q.; Thompson, P.M.; Jack, C.R., Jr.; Weiner, M.W.; Alzheimer's Disease Neuroimaging Initiative. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* **2009**, *132 Pt 4*, 1067–1077. [CrossRef] [PubMed] [PubMed Central]
15. Marciniewicz, E.; Pokryszko-Dragan, A.; Podgórski, P.; Małyszczak, K.; Zimny, A.; Kołtowska, A.; Budrewicz, S.; Sąsiadek, M.; Bladowska, J. Quantitative magnetic resonance assessment of brain atrophy related to selected aspects of disability in patients with multiple sclerosis: Preliminary results. *Pol. J. Radiol.* **2019**, *84*, e171–e178. [CrossRef] [PubMed] [PubMed Central]
16. Lauer, A.; Speroni, S.L.; Patel, J.B.; Regalado, E.; Choi, M.; Smith, E.; Kalpathy-Kramer, J.; Caruso, P.; Milewicz, D.M.; Musolino, P.L. Cerebrovascular disease progression in patients with ACTA2 Arg179 pathogenic variants. *Neurology* **2021**, *96*, e538–e552. [CrossRef]
17. Stelmokas, J.; Yassay, L.; Giordani, B.; Dodge, H.H.; Dinov, I.D.; Bhaumik, A.; Sathian, K.; Hampstead, B.M. Translational MRI Volumetry with NeuroQuant: Effects of Version and Normative Data on Relationships with Memory Performance in Healthy Older Adults and Patients with Mild Cognitive Impairment. *J. Alzheimer's Dis.* **2017**, *60*, 1499–1510. [CrossRef] [PubMed] [PubMed Central]
18. Dorfner, F.J.; Patel, J.B.; Kalpathy-Cramer, J.; Gerstner, E.R.; Bridge, C.P. A review of deep learning for brain tumor analysis in MRI. *npj Precis. Oncol.* **2025**, *9*, 2. [CrossRef] [PubMed]
19. Patel, J.; Chang, K.; Ahmed, S.R.; Jang, I.; Kalpathy-Cramer, J. Opportunities and challenges for deep learning in brain lesions. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*; Lecture Notes in Computer Science; Crimi, A., Bakas, S., Eds.; Springer: Cham, Switzerland, 2022; Volume 12962, pp. 25–36. [CrossRef]

20. Patel, J.; Chang, K.; Hoebel, K.; Gidwani, M.; Arun, N.; Gupta, S.; Aggarwal, M.; Singh, P.; Rosen, B.R.; Gerstner, E.R.; et al. Segmentation, Survival Prediction, and Uncertainty Estimation of Gliomas from Multimodal 3D MRI Using Selective Kernel Networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2020)*; Lecture Notes in Computer Science; Crimi, A., Bakas, S., Eds.; Springer: Cham, Switzerland, 2021; Volume 12659. [CrossRef]
21. Prasanna, P.; Patel, J.; Partovi, S.; Madabhushi, A.; Tiwari, P. Radiomic features from the peritumoral brain parenchyma on treatment-naïve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings. *Eur. Radiol.* **2017**, *27*, 4188–4197. [CrossRef] [PubMed]
22. Beig, N.; Patel, J.; Prasanna, P.; Partovi, S.; Varadan, V.; Madabhushi, A.; Tiwari, P. Radiogenomic analysis of hypoxia pathway reveals computerized MRI descriptors predictive of overall survival in glioblastoma. In *Medical Imaging 2017: Computer-Aided Diagnosis*; Armato, S.G., Petrick, N.A., Eds.; SPIE: Bellingham, WA, USA, 2017; Volume 10134, p. 101341U. [CrossRef]
23. Patel, J.; Ahmed, S.R.; Chang, K.; Singh, P.; Gidwani, M.; Hoebel, K.; Kim, A.; Bridge, C.; Teng, C.; Li, X.; et al. A Deep Learning Based Framework for Joint Image Registration and Segmentation of Brain Metastases on Magnetic Resonance Imaging. In *Machine Learning Research, Proceedings of the 8th Machine Learning for Healthcare Conference, New York, NY, USA, 11–12 August 2023*; PMLR: Brookline, MA, USA, 2023; Volume 219, pp. 565–587. Available online: https://proceedings.mlr.press/v219/patel23a.html (accessed on 1 September 2025).
24. Peng, J.; Kim, D.D.; Patel, J.B.; Zeng, X.; Huang, J.; Chang, K.; Xun, X.; Zhang, C.; Sollee, J.; Wu, J.; et al. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-Oncology* **2022**, *24*, 289–299. [CrossRef] [PubMed]
25. Leung, J.-H.; Karmakar, R.; Mukundan, A.; Lin, W.-S.; Anwar, F.; Wang, H.-C. Technological Frontiers in Brain Cancer: A Systematic Review and Meta-Analysis of Hyperspectral Imaging in Computer-Aided Diagnosis Systems. *Diagnostics* **2024**, *14*, 1888. [CrossRef]
26. Mukundan, A.; Gupta, D.; Karmakar, R.; Wang, H.C. Transforming pediatric imaging: The role of four-dimensional flow magnetic resonance imaging in quantifying mesenteric blood flow. *World J. Radiol.* **2025**, *17*, 106582. [CrossRef] [PubMed] [PubMed Central]
27. Robinson-Weiss, C.; Patel, J.; Bizzo, B.C.; Glazer, D.I.; Bridge, C.P.; Andriole, K.P.; Dabiri, B.; Chin, J.K.; Dreyer, K.; Kalpathy-Cramer, J.; et al. Machine learning for adrenal gland segmentation and classification of normal and adrenal masses at CT. *Radiology* **2022**, *306*, 285–295. [CrossRef]
28. Ilesanmi, A.E.; Ilesanmi, T.; Gbotoso, G.A. A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks. *Healthc. Anal.* **2023**, *4*, 100261. [CrossRef]
29. Çallı, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep learning for chest X-ray analysis: A survey. *Med. Image Anal.* **2021**, *72*, 102125. [CrossRef] [PubMed]
30. Mahler, L.; Steiglechner, J.; Bender, B.; Lindig, T.; Ramadan, D.; Bause, J.; Birk, F.; Heule, R.; Charyasz, E.; Erb, M.; et al. Submillimeter Ultra-High Field 9.4 T Brain MR Image Collection and Manual Cortical Segmentations. *Sci. Data* **2025**, *12*, 635. [CrossRef]
31. Billot, B.; Greve, D.N.; Puonti, O.; Thielscher, A.; Van Leemput, K.; Fischl, B.; Dalca, A.V.; Iglesias, J.E. ADNI SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* **2023**, *86*, 102789. [CrossRef] [PubMed] [PubMed Central]
32. Marcus, D.S.; Wang, T.H.; Parker, J.; Csernansky, J.G.; Morris, J.C.; Buckner, R.L. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **2007**, *19*, 1498–1507. [CrossRef]
33. Chen, H.; Zendehdel, N.; Leu, M.C.; Yin, Z. A gaze-driven manufacturing assembly assistant system with integrated step recognition, repetition analysis, and real-time feedback. *Eng. Appl. Artif. Intell.* **2025**, *144*, 110076. [CrossRef]