International Journal of Hyperthermia

ISSN: 0265-6736 (Print) 1464-5157 (Online) Journal homepage: www.tandfonline.com/journals/ihyt20

Automated subcutaneous fat segmentation with a convolutional neural network in magnetic resonance guided high-intensity focused ultrasound treatment for uterine fibroids

Chenchen Bing, Anna Laaksonen, Kirsi Joronen, Gaber Komar, Teija Sainio, Roberto Blanco Sequeiros, Ari Partanen & Simon Köttgen

To cite this article: Chenchen Bing, Anna Laaksonen, Kirsi Joronen, Gaber Komar, Teija Sainio, Roberto Blanco Sequeiros, Ari Partanen & Simon Köttgen (2026) Automated subcutaneous fat segmentation with a convolutional neural network in magnetic resonance guided high-intensity focused ultrasound treatment for uterine fibroids, International Journal of Hyperthermia, 43:1, 2634734, DOI: 10.1080/02656736.2026.2634734

To link to this article: https://doi.org/10.1080/02656736.2026.2634734

Published online: 26 Feb 2026.

INTERNATIONAL JOURNAL OF HYPERTHERMIA
2026, VOL. 43, NO. 1, 2634734
https://doi.org/10.1080/02656736.2026.2634734

Automated subcutaneous fat segmentation with a convolutional neural network in magnetic resonance guided high-intensity focused ultrasound treatment for uterine fibroids

Chenchen Bing[a,b], Anna Laaksonen[c], Kirsi Joronen[d,e], Gaber Komar[a,e,f], Teija Sainio[d,e], Roberto Blanco Sequeiros[a,e], Ari Partanen[b] and Simon Köttgen[b]

[a]Department of Radiology, Turku University Hospital, Turku, Finland; [b]Profound Medical, Mississauga, Canada; [c]Department of Internal Medicine, University of Turku, Turku, Finland; [d]Department of Medical Physics, Turku University Hospital, Turku, Finland; [e]Department of Clinical Medicine, University of Turku, Turku, Finland; [f]Department of Obstetrics and Gynecology, Turku University Hospital, Turku, Finland

ABSTRACT

Introduction: In MR-guided high-intensity focused ultrasound (MR-HIFU) treatment for uterine fibroids, the subcutaneous abdominal fat layer is prone to unwanted heating, especially during consecutive sonications. Automating its delineation with a deep learning algorithm would enhance treatment safety, efficiency and simplify clinical workflow.

Materials and Methods: The subcutaneous fat layer was manually segmented on MR images from 62 patients treated with MR-HIFU. An attention gated U-Net convolutional neural network (CNN) was trained and validated using the Dice coefficient (DC). Model performance was tested using the DC and 95th percentile Hausdorff distance (HD95). The clinically relevant accuracy was assessed by the average fat layer thickness. The model's transferability was determined on data prepared by a second reader and compared to the interobserver variability.

Results: The model achieved a DC of 0.972 (IQR: 0.951–0.983), and an HD95 of 1.1 (IQR: 0.8–3.2) mm on the held-out test dataset. The mean absolute thickness error between ground truth and the model's prediction was 0.8 ± 0.8 mm for the test dataset, and was 0.7–0.8 mm on the two patients prepared by the secondary reader. The automated segmentation algorithm successfully reduced the segmentation time from 3 min to 3 s.

Conclusion: We established an automatic segmentation algorithm based on an attention-gated U-Net architecture to delineate the abdominal fat layer in MR images of uterine fibroids patients. The model achieved high accuracy, robustly handling both thin and thick fat layers, and performed reliable on data prepared by a second reader.

ARTICLE HISTORY
Received 14 October 2025
Revised 9 February 2026
Accepted 16 February 2026

KEYWORDS
Deep-learning; high-intensity focused ultrasound ablation; uterine fibroids; segmentation; MRI; convolutional neural network

Introduction

Uterine fibroids (also known as uterine leiomyomas) are common growths of the uterus, the most common uterine tumors and are estimated to occur in over 70% of women by the onset of menopause [1]. Although the disease is benign, it is symptomatic in up to 50% of patients. These symptoms can significantly affect the patient's physical and mental health without proper intervention. Current standards of care for uterine fibroids include myomectomy [2,3] and uterine artery embolization (UAE) [4–6].

Magnetic resonance guided high-intensity focused ultrasound (MR-HIFU) is a noninvasive intervention. It utilizes focused ultrasound energy to thermally ablate tissue at the focus and magnetic resonance imaging (MRI) for accurate treatment planning and real-time temperature monitoring. However, detailed knowledge of the tissue structure, especially within the ultrasound beam path, is critical for delivering fast and safe personalized treatments. For example, the abdominal subcutaneous fat layer tends to accumulate heat over long MR-HIFU treatments, leading to the potential risk of undesired tissue damage.

CONTACT Simon Köttgen skottgen@profoundmedical.com Profound Medical, Mississauga, Canada.

In the Sonalleve MR-HIFU system (Profound Medical, Mississauga, Canada), the user is required to manually delineate the subcutaneous fat layer in a single slice. This segmentation is then used to estimate the average thickness of the fat layer, which in turn is given as input to a thermal simulation. Together with other ultrasound parameters including acoustic power, frequency, sonication duration, these factors are considered to determine a safe waiting time between subsequent sonications. Throughout the procedure, the user might need to repeat the delineation due to patient repositioning. This process could be simplified, sped up and made more reproducible with automatic image segmentation.

Traditional segmentation techniques were usually based on edge detection, active contouring and thresholding methods [7,8]. They highly rely on the selected features and efficient pre- and post-processing algorithms. Machine learning models, on the other hand, can adaptively learn high-dimensional features and improve the robustness of the model [9]. Convolutional neural networks (CNN) are a subset of machine learning often used for image classification and object recognition tasks [10,11]. With superior performance on images, CNNs are commonly used in automatic image segmentations. Although CNNs have existed for a long time, their performance was still limited by the size of the available training sets and the size of the networks considered. The first introduction of U-Net architecture was brought up by Ronneberger [12]. The U-Net CNN can work with very few training images [13] and yields more precise segmentations [14], which is critical in biomedical imaging applications.

The objective of this study was to develop a deep-learning model based on MR images commonly acquired in MR-HIFU therapy of uterine fibroids and assess its performance in automatic fat segmentation to simplify and speed up the clinical workflow.

Materials and methods

Data acquisition

This is a retrospective single center study covering a total of 62 patients (60 for model development and test, 2 for interobserver variability and transferability analysis) enrolled between 2020 to 2024 at Turku University Hospital in Turku, Finland. The pretreatment screening and MR-HIFU treatment were conducted using a 3.0T Philips MR scanner (Ingenia, Philips Healthcare, The Netherlands) and a clinical MR-HIFU system (Sonalleve V2, Profound Medical, Mississauga, Canada). The local institutional review board approved the retrospective data collection and study design and granted a waiver for acquiring informed consent from the patients. According to Finnish legislation (552/2019), no separate ethics committee approval was needed.

All uterine fibroid patients in this study were scanned with the default planning MR sequence provided by the Sonalleve MR-HIFU system (3D Turbo Spin-Echo, resolution = $0.625 \times 0.625 \times 1$ mm, matrix = $400 \times 400 \times 200$, TR = 1500 ms, TE = 160 ms, echo train length = 81). The software of the Sonalleve MR-HIFU system currently defines a default tissue boundary curve for segmentation of the subcutaneous fat layer (Figure 1A). The user is required to manually manipulate and adjust the default tissue boundary as part of the clinical workflow (Figure 1B).

Data preparation

Data collected from 60 patients were manually segmented slice by slice by a MR clinical scientist with 10 years of experience (CB) using the Medical Imaging Interaction Toolkit (MITK, v2023.04.2, MITK.org). The segmentation results were then reviewed by a certified radiologist with seven years of experience (AL). To ensure rigorous evaluation and prevent data leakage, we employed a patient-wise split, randomly assigning 60 patients to a training set ($n = 42$), a validation set ($n = 9$), and a held-out test set ($n = 9$). This guaranteed that all images from a single patient were contained within one subset. The model was trained on the training set, and hyperparameter tuning was performed based on the performance on the validation set. The final model's performance was evaluated exclusively on the held-out test set. To assess interobserver variability and transferability, two additional cases were processed blindly and independently by two readers, CB (R1) and a certified radiologist with over fifteen years of experience GK (R2).

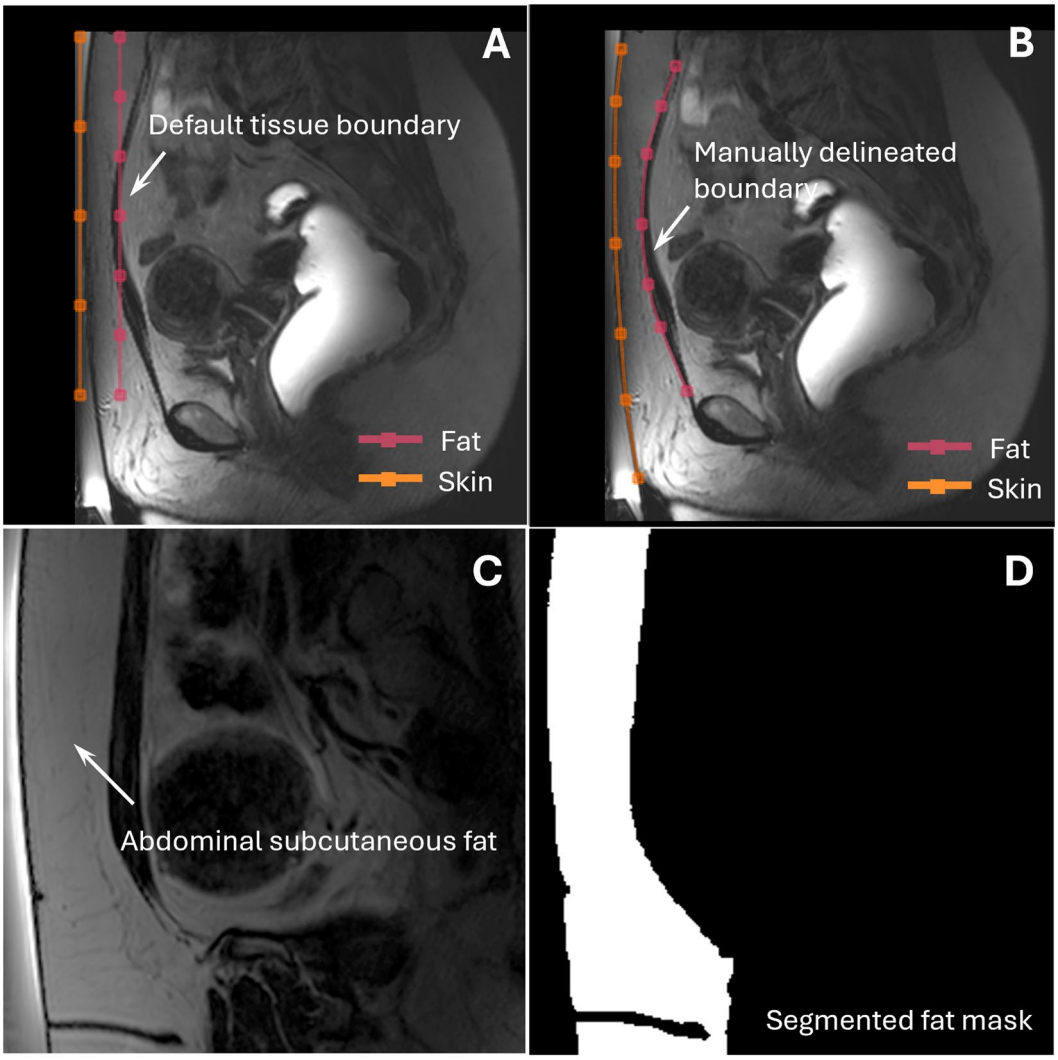

Figure 1. Fat segmentation in Sonalleve MR-HIFU system and in this study. The default tissue boundaries for the skin (orange) and fat (pink) are a straight, anchored line (A). Manual manipulation and adjustment by the user are required to set the proper tissue boundaries (B). Example slices of the data used in this study and the manually segmented fat mask are shown in (C) and (D), respectively.

The image preparation was performed using Python (version 3.10). To ensure that all sagittal slices actually contained subcutaneous fat, 64 sagittal slices in the center of the imaging block covering the primary treatment area were obtained from each patient. The images and labels were resampled to 1 mm isotropic resolution and cropped to a size of 256×256. The image intensity was normalized to a range of 0–255. To enhance the diversity of the training dataset [15], a combination of geometric and photometric augmentations was applied. These include horizonal flipping, random rotation (±20°), motion blurring (blur kernel size of 9×9 pixels), 90% scaling, pixel translation, ±3 degrees shearing, ±20% distortion, and elastic deformation. Additional noise was introduced with Gaussian noise or multiplicative noise or contrast (±5%) and brightness (±5%) adjustment. All the transformations were implemented with Python albumentations module (version 1.3.0).

Convolutional neural network architecture

A U-Net structure with eight encoding and eight decoding blocks was implemented using TensorFlow library (v2.11.0) [16]. An exploration study was conducted by varying the number of initial filters in the first encoder block as it determines whether the model would focus on local features or global features [17]. Three configurations were trained and evaluated: a narrow network starting with

16 filters, a medium network with 32 filters, and a wider network with 64 filters. In all cases, the number of filters in the convolutional layers increased progressively in the encoder path (×1, ×2, ×4, ×8, ×8, ×8, ×8, ×8). The selection of the final number of filters was determined considering the computational cost and the performance on thin subcutaneous fat layers, a known challenging aspect of this task [18].

Each convolution used a 4×4 kernel and 2×2 stride followed by a non-linear leaky rectified linear unit (ReLU) activation function. Starting from the second layer, batch normalization was added after the convolutional layer to stabilize the gradients and enable faster convergence. At the end of the encoder path, the input image was flattened into a 1×1×512 map and connected to the decoder path. In the decoder path, a normal ReLU activation function was applied, followed by the 2D transposed convolution with 4×4 kernel and 2×2 stride. For the first three layers, dropout (rate = 0.3) was applied to prevent overfitting and improve generalization [19]. At the last level of the expansive path, a Sigmoid function was used for segmentation to generate the final output image with the size of 256×256×1.

To enhance our segmentation network for clinical settings, we applied spatial attention mechanisms to address the specific challenges in cases with thin fat layers. As an optimization of the basic model, attention gates were incorporated in the skip function [20]. The gate mechanism computes spatial attention weights through a gating signal pathway that combines these inputs via element-wise addition, followed by ReLU activation and a 1×1 convolution. The resulting attention coefficients (range from 0 to 1) were applied multiplicatively to the encoder features before concatenation with decoder features.

Training and evaluation of the model

The model was trained by minimizing the loss of the Dice coefficient (DC), as a statistic measure for the similarity of two samples [21,22]. We used the Adam optimizer in TensorFlow (v2.11.0) with a learning rate of 0.001 for 80 epochs. The training dataset was shuffled after each epoch to eliminate data order bias. Segmentation performance was evaluated on the independent held-out test dataset consisting of 9 patients (576 images) using DC for similarity and the 95th percentile Hausdorff distance (HD95) for boundary agreement [23]. Metrics were computed on a slice-by-slice basis and subsequently reported as median, 25th and 75th percentile.

Average fat layer thickness was calculated as an additional clinically relevant evaluation metric of the segmentation model. In MR-HIFU treatment for uterine fibroid, patients are typically positioned such that the acoustic window is centered to allow efficient ultrasound energy transmission. Accordingly, fat thickness was quantified along the clinically relevant acoustic pathway rather than over the global abdominal fat distribution.

For each sagittal T2-weighted slice, the thickness calculations were performed on the binary fat mask derived from either reader-annotated ground truth or the model's predicted segmentations. Thickness measurements were performed at the center of the image volume, corresponding to the expected ultrasound beam path. Specifically, 30 image rows at the center on the fat mask were selected, and fat thickness was computed for each row as the anterior-posterior extent of contiguous fat pixels. The thickness values from these 30 rows were averaged to obtain a single slice-level measurement. This procedure was repeated across 64 consecutive slices per patient, and the resulting slice-level thicknesses were averaged to yield a mean fat layer thickness for each patient. The same measurement pipeline was applied identically to ground-truth annotations and model predictions to enable direct comparison.

The thickness error was calculated as the difference between the measurements obtained from the model predictions and ground truth. Errors were computed on a slice-by-slice basis and subsequently averaged across 64 slices to obtain a patient-level mean thickness error. Both actual signed and absolute thickness errors were evaluated.

Interobserver variability and transferability analysis

To assess the robustness and clinical applicability of the model, we performed an analysis of interobserver variability and model transferability using data from two additional patients (128 images). Two readers, a primary reader (R1) and a secondary reader (R2), independently processed the datasets.

Interobserver variability quantifies the agreement between different human experts when determining the ground truth segmentations. It was assessed by comparing the segmentations from R1 and R2 on a per-slice basis. We used the DC to measure geometric overlap and the HD95 to capture boundary discrepancies. This interobserver agreement was reported as median, 25th and 75th percentile for DC and HD95, and served as the performance benchmark for subsequent model evaluations.

Transferability refers to the model's ability to perform effectively on data prepared by a different user. This was tested by applying the final model trained with the annotations done by R1 to the unseen data segmented by R2. DC and HD95 were reported as median, 25th and 75th percentile. Fat layer thickness was measured and compared for the model's prediction as well as the reader-annotated ground truth as a clinically relevant metric.

Statistical analysis

Statistical analysis was performed to evaluate factors affecting segmentation performance and the accuracy of fat thickness estimation. A three-way analysis of variance (ANOVA) was used to evaluate the effects of attention gate implementation (with and without), number of convolutional filters (16, 32, and 64), and fat thickness category (general and thin) on segmentation performance, as measured by DC and HD95. Post-hoc comparisons for the number of filters were conducted using Tukey's test, and significant interactions were further explored using simple effects analysis.

Patient-level thickness errors were summarized using mean ± standard deviation. Agreement between thickness measurements calculated based on ground truth and predictions was assessed using mean signed error, mean absolute error, Pearson correlation, and Bland-Altman analysis. A two-sided significance level of $p < 0.05$ was used for all statistical tests.

Results

Data preparation and training process

In this study, images from 60 patients were used for model training, validation and testing. The images were manually segmented with examples shown in Figure 1C,D. 70% of the whole dataset (42 patients) were included in the training dataset, with 15% (9 patients) allocated to test and validation dataset, respectively.

Statistical analysis revealed significant effects of number of filters and fat thickness on both Dice coefficient (DC) and 95th percentile Hausdorff distance (HD95) (both $p < 0.001$). As anticipated, models performed significantly worse in cases with thin fat layers compared to general cases with thicker fat layers. Regarding the impact of number of filters, post-hoc analysis using Tukey's Honestly Significant Difference test indicated that the primary performance difference stemmed from the 16-filter configuration, which yielded significantly poorer results than both 32- and 64-filter models ($p < 0.001$). No statistically significant difference was observed between 32 and 64 filters ($p = 0.818$), suggesting diminishing returns beyond 32 filters.

While attention gate implementation alone showed no significant effect on performance metrics (DC: $p = 0.700$; HD95: $p = 0.102$), a significant interaction between attention gate and fat thickness was identified ($p < 0.001$ for both DC and HD95). This interaction revealed that attention gates provided selective benefit, substantially improving segmentation accuracy specifically for challenging thin-fat-layer cases. Detailed performance metrics across all experimental conditions are presented in Table 1, and examples of the performance of the model under different conditions are included in Figure 2.

Balancing computational efficiency against performance gains, we selected a U-Net architecture incorporating attention gates with 32 initial filters for subsequent validation testing, as this configuration provided optimal performance without the computational burden. The model's architecture is shown in Figure 3. The DC increased during the initial training phase, reached 0.9 rapidly, and then started to saturate (Figure 4A). The final mean validation DC after 80 epochs was 0.973.

Table 1. Exploration study results for selecting optimal U-net architecture.

	Initial Filters	Training Time (hours)	Validation Dice	General Test ($n=9$; 576 images)		Thin Fat Layer Test ($n=1$; 64 images)	
				DC	HD95 (mm)	DC	HD95 (mm)
Without Attention Gates	64	27.4	0.974	0.973 (0.952–0.983)	1.1 (0.8–3.2)	0.948 (0.937–0.953)	1.1 (0.8–1.8)
	32	7.1	0.972	0.972 (0.954–0.983)	1.1 (0.8–3.2)	0.949 (0.939–0.956)	0.8 (0.8–1.6)
	16	3.2	0.972	0.971 (0.947–0.981)	1.6 (0.8–4.0)	0.925 (0.892–0.942)	2.5 (1.6–6.6)
With Attention Gates	64	37.6	0.974	0.973 (0.955–0.983)	1.1 (0.8–2.9)	0.955 (0.943–0.961)	0.8 (0.8–1.1)
	32	10.1	0.973	0.972 (0.951–0.983)	1.1 (0.8–3.5)	0.951 (0.937–0.961)	0.8 (0.8–1.6)
	16	3.3	0.972	0.970 (0.946–0.982)	1.6 (0.8–4.5)	0.947 (0.936–0.960)	0.8 (0.8–1.6)

Comparison between models with and without attention gates was made, with subcategories for different numbers of initial filters. The model's performance on both the general test dataset and the thin fat layer was included. The median, 25th and 75th percentiles are reported for dice coefficient (DC) and 95th Hausdorff distance (HD95), respectively. Computational cost was demonstrated by training time and increased with more complex model structure.

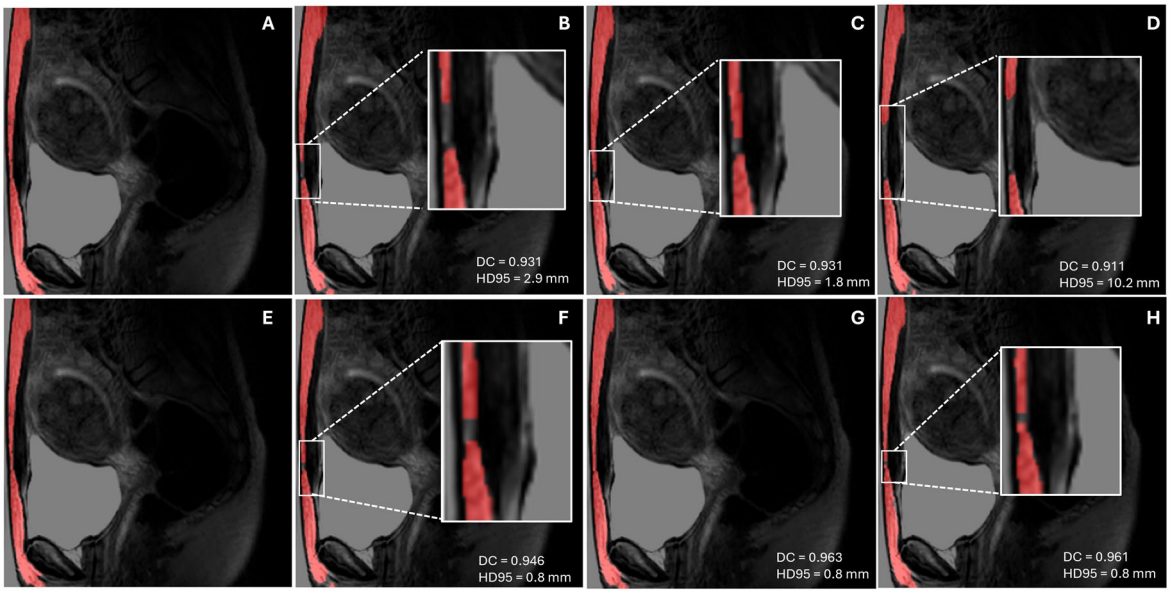

Figure 2. Example of the impact of attention gates and number of filters on the model's performance for a thin fat layer. The segmentation was more accurate with attention gate (E-H) compared to without attentions gates (A–D), demonstrated by a higher median dice coefficient (DC) and a lower 95th percentile Hausdorff distance (HD95). The number of filters has a significant impact after dropping to 16, while no significant difference was observed between 32 and 64 filters despite the integration of attention gates. The zoom-in inserts represent the area where the model failed to include fat pixels. For the model with 16 filters without attention gates (D), a large critical fat region within the ultrasound beam path was missing, leading to a HD95 of 10.2 mm.

Evaluation of the algorithm

As shown in Figure 4B, the median DC was 0.972 (IQR: 0.954–0.983) across 576 test images in 9 patients, and the median HD95 was 1.1 (IQR: 0.8–3.2) mm. The model could successfully segment both thin and thick fat layers (Figure 4C,D). Figure 5 illustrates examples of less accurate cases, including misalignments in the lower abdominal region (Figure 5A,B) and the erroneous inclusion of pubic symphysis (Figure 5C).

Across nine patients in the held-out test dataset, the model demonstrated good agreement with reader-annotated ground truth for fat layer thickness estimation along the clinically relevant acoustic pathway. Ground truth fat thickness was 17.8 ± 7.4 mm, while predicted fat thickness was 18.1 ± 6.8 mm. The mean signed error was 0.2 ± 0.9 mm, indicating slight positive bias. The mean absolute thickness error was 0.8 ± 0.8 mm, roughly equal to the in-plane voxel size (0.65 mm), indicating small deviations exist within the imaging resolution. Notably, a large absolute error (2.94 mm) was only observed in one

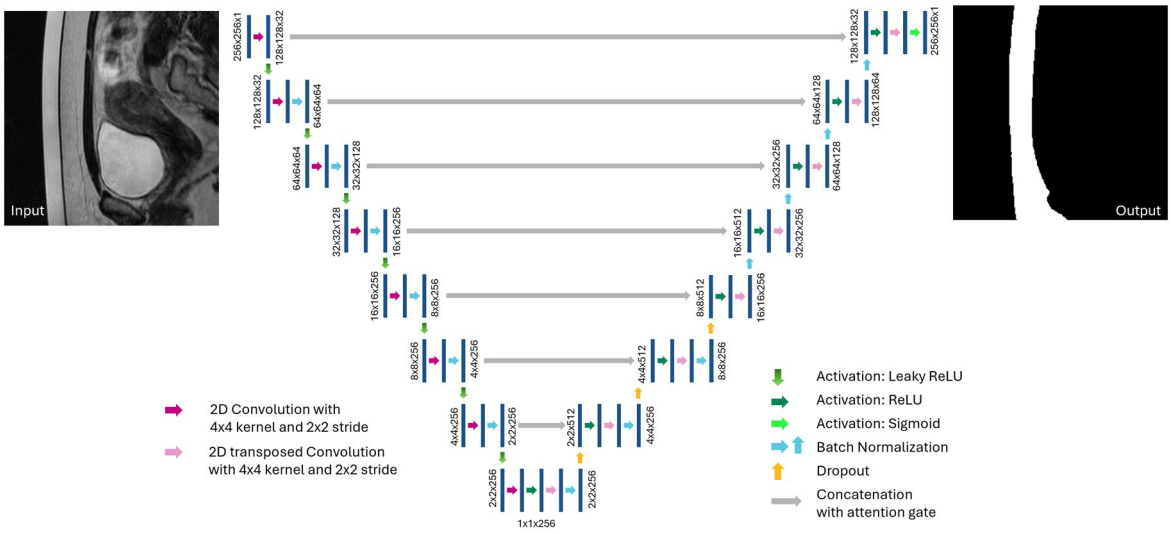

Figure 3. Schematic description of modified 2D U-Net structure consisting of 8 levels of contraction and expansion; input and output images have size 256x256x1; both contraction and expansion path apply convolutions or transposed convolutions with a 4x4 kernel and 2x2 stride together with batch normalization; activation during contraction and expansion is done with a leaky or normal Rectified Linear Unit (ReLU) respectively; first three levels of expansion apply dropout to avoid overfitting. The final activation was performed using a Sigmoid function to generate the output image.

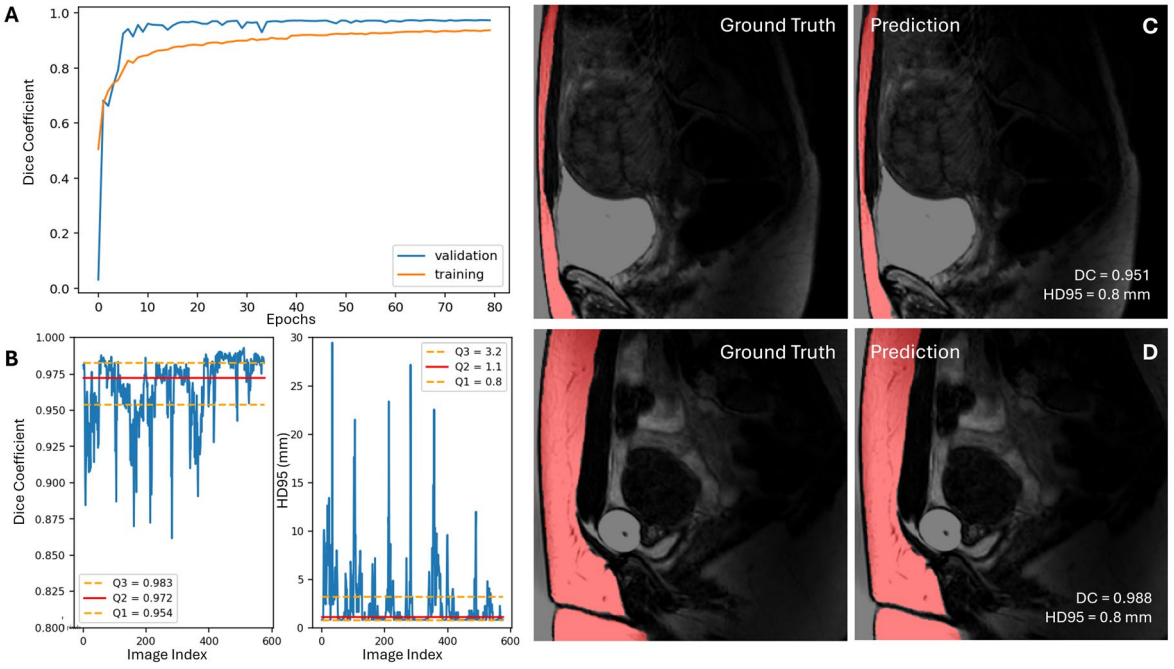

Figure 4. Training curve and model evaluation. The median validation dice coefficient (DC) reached 0.973 after 80 epochs (A). On the held-out test dataset of 576 images, the median DC was 0.972 and the median 95th percentile Hausdorff distance (HD95) was 1.1 mm (B). The model could successfully segment both thin (C; DC = 0.951, HD95 = 0.8 mm) and thick (D; DC = 0.988, HD95 = 0.8 mm) fat layers.

patient with extremely thin fat layer (4.8 mm). Bland-Altman analysis showed a mean bias of 0.23 mm with 95% limits of agreement ranging from −1.55 to 2.02 mm (Figure 6A). The small magnitude of bias and narrow spread relative to the voxel size indicates excellent agreement between predicted and ground truth thickness which was further confirmed by Pearson correlation test (Figure 6B, Pearson $r = 1.00$, $p < 0.05$).

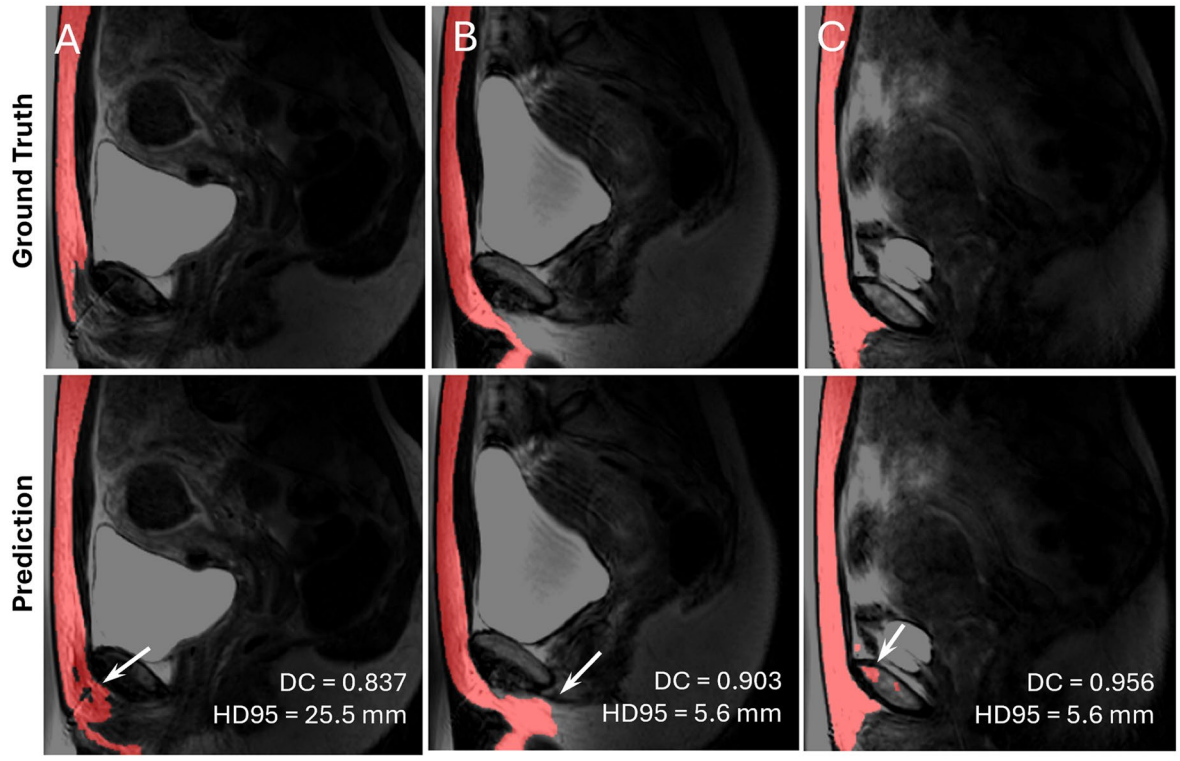

Figure 5. Examples of less accurate segmentations (HD95 > 5 mm) of the subcutaneous fat layer. Both the ground truth (left) and the prediction (right) are shown and compared using the dice coefficient (DC) and 95th percentile Hausdorff distance (HD95). Reasons for less accurate segmentation are discrepancies in the lower abdominal region (A and B, white arrow) and falsely included pubic symphysis (C, white arrow).

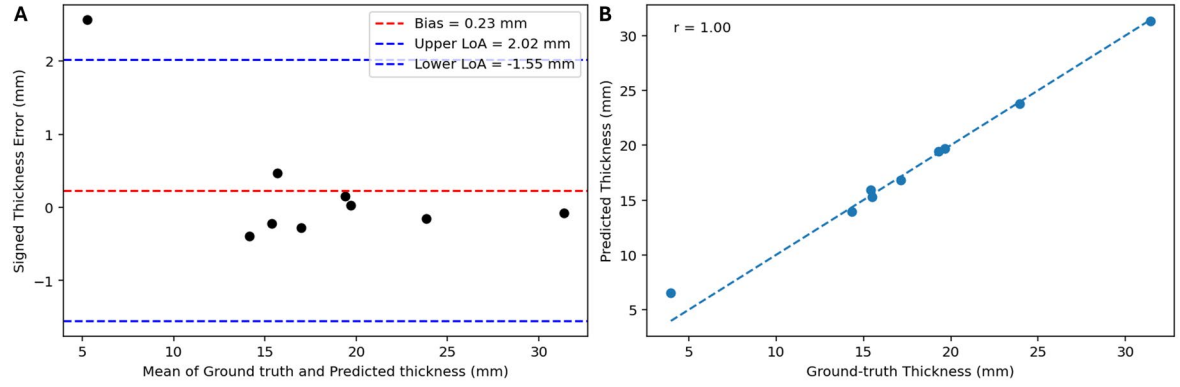

Figure 6. Agreement between predicted and ground-truth fat layer thickness measurements on the held-out test set ($n = 9$). Bland-Altman plot showing the signed difference between predicted and ground truth thickness as a function of their mean (A). The dashed red line indicates the mean signed error (bias), and the upper and lower dashed blue line represent the 95% limits of agreement (LoA). Pearson correlation plot comparing the mean predicted and ground-truth fat thickness indicates strong correlation between the two curves (B, $r = 1.00$). The dashed line represents the identity line ($y = x$).

Interobserver variability and transferability analysis

The median DC between the two readers was 0.934 (IQR: 0.906–0.962), indicating substantial agreement. Median HD95 was 7.6 mm (IQR: 2.8–10.6) mm, reflecting moderate variability in the spatial alignment of boundaries across slices.

Transferability of the model was demonstrated with a median DC of 0.934 (IQR: 0.871–0.952) and a median HD95 of 6.2 (IQR: 4.8–19.8) mm. Figure 7 shows the model's prediction compared to the masks

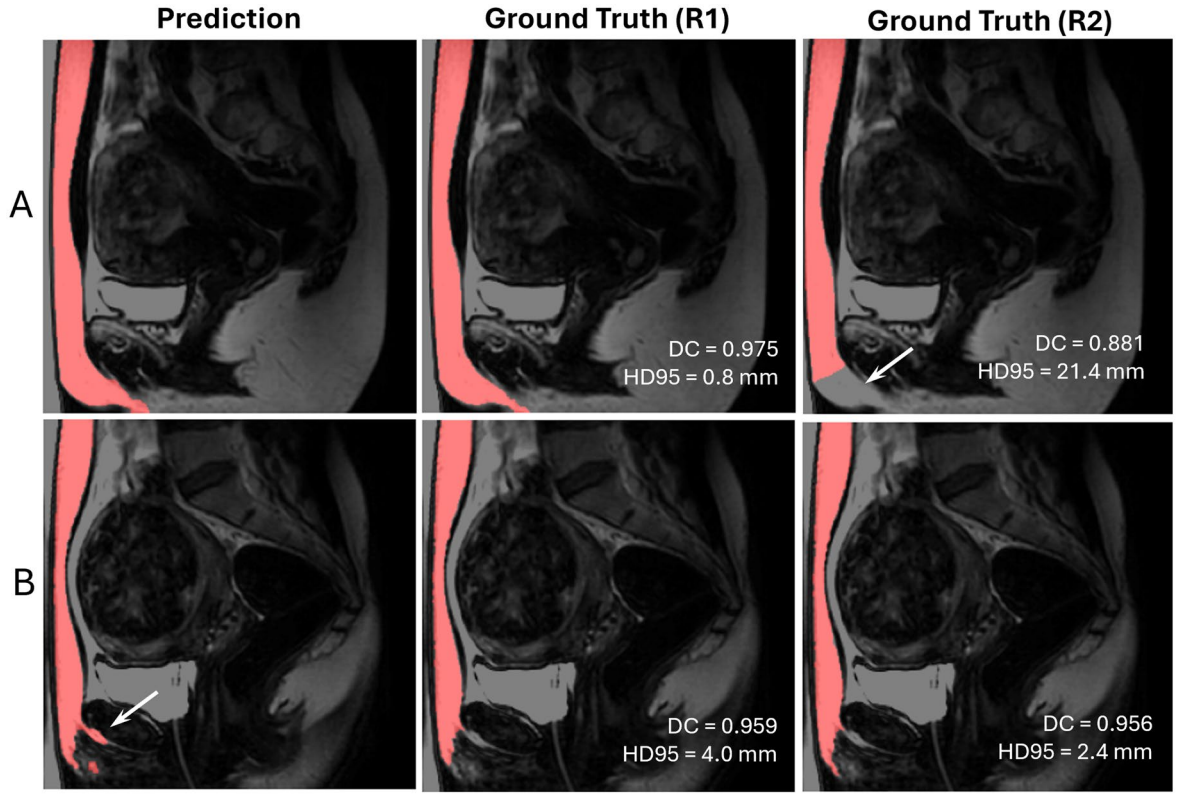

Figure 7. Examples of transferability analysis result. Discrepancies in the lower abdominal region resulted in low dice coefficient (DC) and large 95th Hausdorff distance (HD95) as shown in A (white arrow). Slight overfitting of the prediction results in a relatively high HD95 on the ground truth prepared by both readers (B, white arrow).

annotated by R1 and R2, respectively. The primary discrepancies were the lower abdominal region. In Figure 7A, the high HD95 on R2 mask (21.4 mm) was derived from the lower part where R2 chose not to include that region as it was not relevant to the treatment. A second example is shown in Figure 7B where the model prediction included more tissue in the lower abdominal region, indicating overfitting because of being trained with data annotated by R1.

On the dataset annotated by the second reader, the model demonstrated comparable fat thickness estimates along the clinically relevant acoustic pathway. The fat layer thickness based on the R2 annotated ground truth was 9.9 ± 1.4 mm for patient 1, and 15.8 ± 0.7 mm for patient 2. The model predicted fat thickness was 10.5 ± 1.8 mm for patient 1, and 16.1 ± 1.4 mm for patient 2. The mean absolute thickness error was 0.7 mm for patient 1, and 0.8 mm for patient 2. A small positive bias in fat thickness estimation was observed, with mean signed errors of 0.6 mm for patient 1 and 0.6 mm for patient 2, respectively. These errors were comparable to the in-plane voxel size of 0.65 mm, indicating that the differences were clinically acceptable with the current image resolution for treatment planning.

Discussion

Deep learning has been widely applied in multiple fields including treatment planning [24–26] and tumor delineation [27,28]. In MR-HIFU therapies, there are multiple steps that may benefit from deep learning to achieve streamlined clinical workflow. In a recent study where the researchers developed a deep learning based model for automatic measurement of uterus volume, fibroid volume, and ablated non-perfused volume (NPV) during MR-HIFU treatment of uterine fibroids, the derived DC was 0.90 (uterus), 0.84 (NPV) and 0.74 (fibroid) [29]. These automatic segmentations could further simplify the workflow and improve treatment efficiency. In this study, we successfully developed an attention gated U-Net model which could accurately segment both the thin and thick abdominal subcutaneous fat in

patients who underwent MR-HIFU treatment. The median DC across 576 test images was 0.972, and the HD95 was 1.1 mm. The absolute error of the average fat layer thickness on the model's predicted segmentations was 0.8 mm, comparable to the current in-plane voxel size of 0.65 mm, indicating clinically acceptable discrepancy.

In this study, we investigated a simple U-Net architecture, as well as more advanced models integrating attention gates and deep supervision. In addition, the number of filters was also varied to achieve a balanced model with good performance and acceptable computational cost. Given the specific case of segmenting abdominal subcutaneous fat, deep supervision seems not to be necessary as the fat has well-defined boundaries. Number of filters, attention gate and fat thickness impact the model's performance, but the effect is small or negligible ($\eta^2p < 0.01$) except for the fat thickness, which has a medium impact on Dice coefficient ($\eta^2p < 0.06$). For this specific application, the abdominal fat layer is well defined with good contrast on T2 images, hence moderate number of filters and relatively simplified model structure derived best results without limited overfitting and underfitting. This highlights a tradeoff between efficiency and complexity: while fewer filters suffice for simple structures, more complex anatomies may require higher filter counts and more advanced models to preserve accuracy.

While 3D convolutional networks offer theoretical advantages for capturing spatial continuity in volumetric data, our current 2D slice-by-slice approach remains clinically appropriate and technically sufficient for subcutaneous fat segmentation in MR-HIFU treatment planning. The current 2D model delivers reliable segmentation masks with fat thickness error comparable with image resolution, and strong similarity to the ground truth (DC = 0.972). Future work could explore hybrid approaches, such as 2.5D networks or post-processing with 3D smoothing, to potentially enhance inter-slice consistency while maintaining computational efficiency. However, given the current clinical workflow where physicians review individual slices during treatment planning, and the substantial GPU memory requirements of full 3D U-Nets, our 2D implementation represents an optimal balance of accuracy, efficiency, and clinical integration potential.

A clear definition of boundaries is critical for variability and transferability assessment. In this study, it is difficult to clearly define the fat tissue boundary in the lower abdominal region. The treatment target in MR-HIFU therapies for uterine fibroids are commonly located in the center of the MR images. Hence, the tissue at the edges of the image is not clinically relevant as no ultrasound travels through it. Future studies should prioritize dataset refinement to exclude non-critical regions and better reflect clinical practice.

One limitation in our study is that the original dataset used for model development was prepared by a single reader. Existing literature reports that models trained on annotations from just one expert tend to encode that rater's bias, resulting in inconsistent performance when applied to data labeled by others [30]. In contrast, methods that integrate labels from multiple raters demonstrate stronger calibration and improved segmentation accuracy across observer differences [31]. Meanwhile, diversity in data sources such as different scanners or imaging centers is also important to build a reliable model [32]. Establishing a larger dataset with adequate diversity would further optimize the model's robustness and reliability.

The integration of machine learning into treatment planning and prediction has attracted increasing attention in recent years. Numerous efforts have focused on improving treatment efficiency and accuracy by incorporating AI-based tools into clinical workflows. Machine learning–based image segmentation has been investigated for delineating fibroids, the uterus, and post-treatment non-perfused volume (NPV), thereby streamlining the procedure [29]. The dataset could be formed by not only T2-weighted images, but contrast enhanced T1 images, or other related sequence where the target structure presents as a differentiate intensity or contrast. Retrospective assessment of 30 uterine fibroid treatments performed on the Sonalleve system found that users require on average 30s to 3 min for manual delineation of the subcutaneous fat in a single slice. In our study, we have successfully reduced the time to 3s per image stack on one patient (automatic segmentation) while maintaining the accuracy for clinical acceptance. In addition, this approach opens the potential for a more accurate and personalized, and therefore safer MR-HIFU treatment, by providing information on the shape of the subcutaneous fat layer in the whole treatment region. The present study represents an initial step toward incorporating machine learning into clinical workflow for automated treatment planning. Future work could expand its application across the entire procedure to achieve a more efficient and effective treatment modality.

Conclusion

We developed and validated an attention-gated U-Net algorithm to automatically segment abdominal subcutaneous fat for MR-HIFU treatment planning in uterine fibroid patients. The model demonstrated high segmentation fidelity (median DC: 0.972, HD95: 1.1 mm, absolute error of average fat layer thickness: 0.8 mm) and performed acceptably when compared to a secondary reader, particularly within the clinically relevant treatment region. Despite being trained on single-reader annotations, this approach shows considerable promise for reducing manual segmentation time, standardizing planning workflows, and enhancing treatment safety. As a first step toward full automation in MR-HIFU therapy, this work provides a validated technical foundation ready for further refinement and clinical integration.

Acknowledgements

The authors thank Department of Radiology of University of Turku and Profound Medical Inc. for providing computational resources.

Disclosure statement

SK, AP are employed at Profound Medical Inc., which manufactures an MR-HIFU system. CB was employed at Profound Medical Inc. during the preparation of the manuscript.

Funding

This work was supported by the European Union through the IMAGIO project as part of the Innovative Health Initiative Joint Undertaking (JU) under grant agreement No 101112053. The JU receives support from the European Union's Horizon Europe research and innovation program and life science industries represented by COCIR, EFPIA/Vaccines Europe, EuropaBio and MedTech Europe.

Data availability statement

Patient data used in this study cannot be publicly shared because of the national legislature on the privacy of patient data.

References

[1] Stewart EA, Cookson CL, Gandolfo RA, et al. Epidemiology of uterine fibroids: a systematic review. BJOG. 2017;124(10):1501–1512. doi: 10.1111/1471-0528.14640.
[2] Hu L, Zhao J-S, Xing C, et al. Comparison of focused ultrasound surgery and hysteroscopic resection for treatment of submucosal uterine fibroids (FIGO type 2). Ultrasound Med Biol. 2020;46(7):1677–1685. doi: 10.1016/j.ultrasmedbio.2020.02.018.
[3] Yu S, Bhagavath B, Shobeiri SA, et al. Clinical and patient reported outcomes of pre- and postsurgical treatment of symptomatic uterine leiomyomas: a 12-month follow-up review of TRUST, a surgical randomized clinical trial comparing laparoscopic radiofrequency ablation and myomectomy. J Minim Invasive Gynecol. 2022;29(6):726–737. doi: 10.1016/j.jmig.2022.01.009.
[4] Pyra K, Szmygin M, Szmygin H, et al. Endovascular embolization as a treatment for symptomatic adenomyosis - results of preliminary study. Ginekol Pol. 2022;93(4):273–277. doi: 10.5603/GP.a2021.0136.
[5] Zhang Y, Xu Y, Zhang X, et al. 8Spheres conformal microspheres as embolic agents for symptomatic uterine leiomyoma therapy in uterine artery embolization (UAE): a prospective clinical trial. Medicine (Baltimore). 2023;102(9):e33099. doi: 10.1097/MD.0000000000033099.
[6] Sirkeci F, Moss J, Belli AM, et al. Effects on heavy menstrual bleeding and pregnancy of uterine artery embolization (UAE) or myomectomy for women with uterine fibroids wishing to avoid hysterectomy: the FEMME randomized controlled trial. Int J Gynaecol Obstet. 2023;160(2):492–501. doi: 10.1002/ijgo.14626.
[7] Ximenes Vasconcelos FF, Medeiros AG, Peixoto SA, et al. Automatic skin lesions segmentation based on a new morphological approach via geodesic active contour. Cognit Syst Res. 2019;55:44–59. doi: 10.1016/j.cogsys.2018.12.008.

[8] Mahbod A, Tschandl P, Langs G, et al. The effects of skin lesion segmentation on the performance of dermatoscopic image classification. Comput Methods Programs Biomed. 2020;197:105725. doi: 10.1016/j.cmpb.2020.105725.
[9] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. doi: 10.1038/nature14539.
[10] Obaidi Al-Khafaji SH, et al. An Improved Convolutional Neural Network for Medical Image Segmentation. In *2023 IEEE International Conference on Paradigm Shift in Information Technologies with Innovative Applications in Global Scenario (ICPSITIAGS)* 42–47; 2023. doi: 10.1109/ICPSITIAGS59213.2023.10527722.
[11] Li S, Huang C. Using convolutional neural networks for image semantic segmentation and object detection. Systems and Soft Computing. 2024;6:200172. doi: 10.1016/j.sasc.2024.200172.
[12] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells, WM & Frangi, AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham; 2015. p. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[13] Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact. Removal | IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/8697107.
[14] Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imaging. 2018;37(12):2663–2674. doi: 10.1109/TMI.2018.2845918.
[15] Kozah N, Dornaika F, Charafeddine J, et al. Data augmentation techniques for medical image segmentation – a review. In *2024 International Conference on Computer and Applications (ICCA)* 1–8; 2024. doi: 10.1109/ICCA62237.2024.10927851.
[16] Abadi M, et al. TensorFlow: a system for large-scale machine learning. Preprint at doi: 10.48550/arXiv.1605.08695.; 2016.
[17] Ehab W, Huang L, Li Y. UNet and variants for medical image segmentation. IJNDI. 2024;3(2):100009. doi: 10.53941/ijndi.2024.100009.
[18] Xia Q, Zheng H, Zou H, et al. A comprehensive review of deep learning for medical image segmentation. Neurocomputing (Amst). 2025;613:128740. doi: 10.1016/j.neucom.2024.128740.
[19] Hernandez-Gutierrez FD, Avina-Bravo EG, Ibarra-Manzano MA, et al. Retinal vessel segmentation based on a lightweight U-Net and reverse attention. Mathematics. 2025;13(13):2203. doi: 10.3390/math13132203.
[20] Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207. doi: 10.1016/j.media.2019.01.012.
[21] Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302. doi: 10.2307/1932409.
[22] A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons |. Semantic Scholar. https://www.semanticscholar.org/paper/A-method-of-establishing-group-of-equal-amplitude-S%C3%B8rensen-S%C3%B8rensen/d8d3e6d95b60ec6ac8f91f42a6914a87b13a6bc1.
[23] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Machine Intell. 1993;15(9):850–863. doi: 10.1109/34.232073.
[24] Li L, Lu Z, Jiang A, et al. Swin Transformer-based automatic delineation of the hippocampus by MRI in hippocampus-sparing whole-brain radiotherapy. Front Neurosci. 2024;18:1441791. doi: 10.3389/fnins.2024.1441791.
[25] Lombardo E, et al. Patient-specific deep learning tracking framework for real-time 2D target localization in MRI-guided radiotherapy. Int J Radiat Oncol Biol Phys. S0360-3016(24)03508–9. 2024;122(4):827–837. doi: 10.1016/j.ijrobp.2024.10.021.
[26] Bibault J-E, Giraud P. Deep learning for automated segmentation in radiotherapy: a narrative review. Br J Radiol. 2024;97(1153):13–20. doi: 10.1093/bjr/tqad018.
[27] Dénes-Fazakas L, Kovács L, Eigner G, et al. Enhancing brain tumor diagnosis with L-Net: a novel deep learning approach for MRI image segmentation and classification. Biomedicines. 2024;12(10):2388. doi: 10.3390/biomedicines12102388.
[28] Kunkyab T, Bahrami Z, Zhang H, et al. A deep learning-based framework (Co-ReTr) for auto-segmentation of non-small cell-lung cancer in computed tomography images. J Appl Clin Med Phys. 2024;25(3):e14297. doi: 10.1002/acm2.14297.
[29] Slotman DJ, Bartels LW, Nijholt IM, et al. Development and validation of a deep learning-based method for automatic measurement of uterus, fibroid, and ablated volume in MRI after MR-HIFU treatment of uterine fibroids. Eur J Radiol. 2024;178:111602. doi: 10.1016/j.ejrad.2024.111602.
[30] Shwartzman O, Gazit H, Shelef I, et al. The worrisome impact of an inter-rater bias on neural network training; 2020. Preprint at doi: 10.48550/arXiv.1906.11872.
[31] Lemay A, Gros C, Karthik EN, et al. Label fusion and training methods for reliable representation of inter-rater uncertainty. Preprint at.; 2023. doi: 10.48550/arXiv.2202.07550.
[32] Onofrey JA, Casetti-Dinescu DI, Lauritzen AD, et al. Generalizable multi-site training and testing of deep neural networks using image normalization. Proc IEEE Int Symp Biomed Imaging. 2019;2019:348–351. doi: 10.1109/isbi.2019.8759295.