

Article

Interactive Confidence Thresholding in Virtual Reality for AI-Assisted 3D MRI Segmentation of Mandibular Glands

Nastaran Rasouli ¹, Lotta Orsmaa ¹, Mikko Saukkoriipi ², Jari Kangas ¹, Jorma Järnstedt ^{1,3,4},
Jaakko Sahlsten ², Kimmo Kaski ² and Roope Raisamo ^{1,*}

¹ Faculty of Information Technology and Communication Sciences, Tampere University, Kalevantie 4, 33014 Tampere, Finland; nastaran.rasouli@tuni.fi (N.R.); lotta.orsmaa@tuni.fi (L.O.); jari.a.kangas@tuni.fi (J.K.); jorma.jarntstedt@pirha.fi (J.J.)

² Department of Computer Science, School of Science, Aalto University, Otakaari 1B, 00076 Espoo, Finland; mikko.1.saukkoriipi@aalto.fi (M.S.); jaakko.sahlsten@aalto.fi (J.S.); kimmo.kaski@aalto.fi (K.K.)

³ Department of Radiology, Tampere University Hospital, Wellbeing Services County of Pirkanmaa, Elämänaukio 2, 33520 Tampere, Finland

⁴ Faculty of Medicine and Health Technology, Tampere University, Kalevantie 4, 33014 Tampere, Finland

* Correspondence: roope.raisamo@tuni.fi

Abstract

Standard Three-dimensional (3D) Magnetic Resonance Imaging (MRI) segmentation models typically rely on a fixed threshold or argmax-based class selection, providing little or no insight into model uncertainty, which can limit clinician trust. In this work, we developed a framework that integrates a calibrated 3D U-Net into a Virtual Reality (VR) system, enabling clinicians to manipulate confidence thresholds in real time and observe how the segmentation changes. The network was trained to segment mandibular structures (sub-mandibular glands and mandibular canal) on 55 T2-weighted MRI scans from the AAPM RT-MAC 2019 dataset and post hoc calibrated using temperature scaling. The VR application, built in Unity with OpenXR, offers two interaction modes: a single upper-threshold mode and a range mode that controls both lower and upper bounds. A user study with six clinicians was conducted to evaluate both modes using the UMUX-Lite questionnaire (from which a System Usability Scale (SUS)-equivalent score was derived), interaction logs, and qualitative feedback. Both modes were rated as highly usable, and narrower confidence intervals, focusing on the most reliable predictions, were more popular among users. The results suggest that enabling real-time adjustment of confidence thresholds improves the clarity of segmentation outputs and may support clinicians' perceived confidence, as indicated by participants' qualitative feedback and stated preferences.

Keywords: 3D medical visualization; semantic segmentation; virtual reality; confidence calibration; user-adjustable decision boundaries; interpretability; clinical trust



Academic Editor: Douglas
O'Shaughnessy

Received: 9 June 2026

Revised: 9 July 2026

Accepted: 11 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Radiological images, such as Magnetic-Resonance Imaging (MRI), Computed Tomography (CT), or Cone Beam Computed Tomography (CBCT), are a fundamental part of medical diagnostics and treatment processes [1]. Image annotation—identifying and marking regions such as nerve canals or tumors in radiological images—is traditionally performed manually by radiologists, making it a time-consuming process and inherently subjective [2]. To address the limitations of manual annotation, Artificial Intelligence (AI)-based radiological image segmentation has gained significant attention, enabling the

automatic identification and highlighting of regions of interest within imaging data [3]. If proper training and validation are used, AI can segment radiological images with high accuracy and reliability, while reducing manual labor [4]. Deep learning models based on U-Net architectures have been shown in some cases to achieve expert-level performance in head and neck organ segmentation, comparable to independent clinical experts [5]. Despite these advances, such AI methods are typically applied and evaluated using conventional 2D desktop displays, which can limit the effective perception of complex three-dimensional anatomical structures, although 3D images are used. This limitation can make spatial relationships more difficult to interpret and reduce the intuitiveness of interaction with volumetric data [6].

There is a critical need for transparent and interpretable AI systems in medical imaging, where misinterpretations can have notable consequences. Transparency here refers to the extent to which users and stakeholders can access and understand the internal processes of an AI system [7], which, in the case of deep learning models especially, might otherwise be completely hidden from the user. Interpretability refers to how well end users can understand and make sense of an algorithm's output [8]. The lack of interpretability and interactive uncertainty exploration can reduce the clinical adoption of AI-based medical imaging systems [9]. Both transparency and interpretability fall under the umbrella of eXplainable AI, which aims to provide users with the reasoning needed to make an AI system's operation easy to understand [10].

Voxel-wise segmentation outputs can provide insight into model behavior, but they are often created by a single binary mask using a fixed threshold or argmax rule, which assigns each voxel to the class with the highest predicted probability. Uncertainty is often presented as static heatmaps without interactive controls for clinicians. Thus, two limitations continue to exist: the reliability of a binary segmentation mask that conceals information about model uncertainty, and the difficulty of inspecting complex 3D anatomy on a conventional 2D display. Conventional segmentation systems present a fixed-threshold binary segmentation, whereas our system allows clinicians to interactively adjust the confidence threshold(s) and observe the resulting segmentation in real time.

Virtual Reality (VR) enables stereoscopic depth perception and improved perception of complex anatomical relationships, which are difficult to achieve on flat 2D displays [11]. Additionally, the 3D image can be manipulated more intuitively in VR [12], with increased possibilities for interaction techniques [13] compared to traditional desktop environments. Previous studies with promising results have explored the feasibility of VR for interactive AI segmentation editing, allowing radiologists to point at areas that should be added to or removed from an existing segmentation (either from 2D slices [14] or from the 3D image itself [15]). Consequently, since VR has shown promise as a platform for interactive AI segmentation editing, it is reasonable to also explore its suitability for interactive confidence thresholding.

Previous work has explored uncertainty visualization using static heatmaps and 2D overlays, while VR-based systems have focused primarily on anatomical inspection and segmentation editing (either manual [16] or AI prompting rather than AI interpretability [15]). These approaches lack interactive control over probabilistic outputs and do not leverage calibrated confidence measures. While prior work has explored uncertainty visualization and VR-based medical imaging separately, no existing system combines calibrated uncertainty, real-time interaction, and immersive 3D exploration for segmentation outputs.

To address these challenges, we propose a framework that integrates interactive 3D visualization with calibrated AI outputs. Our approach employs VR headsets to enable immersive inspection of MRI scans and segmentation results, offering better spatial depth compared to traditional 2D interfaces. Additionally, rather than providing binary segmen-

tation masks, our system equips clinicians with tools to explore segmentation outputs with dynamic, user-defined confidence thresholds, revealing more detailed model behavior. This framework could potentially enhance trust by allowing clinicians to efficiently navigate the high- and low-confidence segmentation regions in an intuitive 3D environment. By integrating calibrated predictions derived through temperature scaling with immersive VR visualization, our approach connects complex AI outputs to the practical needs of clinical applications.

In the current study, we present the design, implementation, and evaluation of our VR-based visualization framework, applied to the mandibular region using the AAPM RT-MAC 2019 dataset. Although the segmentation model covers both the submandibular gland and the mandibular canal as a single foreground class, the user study focuses on the submandibular gland because it provides a clearer soft-tissue target in MRI, whereas the mandibular canal is generally better visualized using CBCT. This allowed us to evaluate the proposed interaction technique on a single anatomical structure before extending it to multi-class segmentation. We assess its effectiveness through a user study measuring usability, interaction behavior, and qualitative clinician feedback on trust and interpretability. Our findings indicate that dynamic confidence thresholding in a VR environment can enhance interpretability and may support clinicians' perceived confidence. Whether these benefits improve clinical decision-making remains to be investigated. By introducing a scalable, interactive visualization tool, this work establishes a new paradigm for uncertainty-aware medical visualization, with potential applications that extend to other domains that require intuitive data exploration.

The paper is organized as follows: Section 2 reviews related work. Section 3 details the AI algorithm and VR system used. Section 4 covers the experimental methods, including participants, tasks, conditions, and study design. Section 5 presents the quantitative and qualitative results. These findings are discussed in Section 6, and the paper concludes in Section 7.

2. Related Work

This section reviews prior work in medical image segmentation, confidence calibration, and virtual reality visualization, and highlights the contributions of this work.

2.1. 3D Medical Image Segmentation

Medical image segmentation involves identifying and outlining anatomical structures or other clinically relevant regions to support specific diagnostic or therapeutic tasks. It is applied to a wide range of tasks, including accurate tumor localization, organ boundary identification, and analysis of anatomical structures in volumetric data [17]. Clinical areas such as oncology, neurology, dentistry, and orthopedics depend on segmentation to analyze patient-specific anatomy with accuracy and consistency. As high-resolution imaging has become more common, there is a greater need for automated methods that can process large datasets efficiently while preserving structural detail [17].

The emergence of deep learning has led to the widespread adoption of convolutional neural network architectures, most notably U-Net-based models [18,19] and their three-dimensional variants [20,21]. These approaches now routinely achieve performance levels comparable to the agreement observed between expert radiologists [22,23]. More recently, transformer-based architectures have further pushed the boundaries of segmentation performance [23,24]. Despite these advances, even state-of-the-art models can produce clinically relevant errors, underscoring the continued need for human oversight and validation by radiologists [25].

U-Net models and their variants are considered the standard deep learning models for image segmentation. These models can achieve high accuracy by extracting both local and global features through an encoder–decoder architecture. While their extensions, such as nnU-Net [26], have further improved performance across a wide range of medical image analysis tasks, most segmentation systems still rely on a fixed threshold, usually argmax-based classification, to produce a binary output. This limits the interpretation of model confidence and does not allow users to explore how predictions change across different confidence levels.

2.2. Confidence Calibration

Confidence calibration aims to align a model’s predicted probabilities with the true likelihood of correctness. Modern neural networks are often poorly calibrated and tend toward overconfidence, but post hoc techniques such as temperature scaling can improve calibration without modifying the trained model or its predicted class [27].

In medical imaging, calibrated confidence is especially important for safety-critical decision-making [28]. This is particularly relevant for interactive thresholding: if the underlying probabilities are miscalibrated, a user-selected threshold no longer corresponds to a consistent level of reliability. Calibrating the confidence map before exposing it to user-controlled thresholds ensures the values clinicians filter on are meaningful.

2.3. VR in Medical Visualization

Virtual reality (VR) is increasingly used for the visualization of medical data, including surgical planning, training, and diagnostic support. VR allows users to interact with volumetric data in an immersive way and improves spatial understanding compared to traditional 2D displays.

However, most VR systems focus only on visualization rather than interaction with AI-generated predictions. Segmentation results are usually presented as fixed outputs, without access to underlying confidence information or the ability to adjust model behavior.

Previous studies have explored uncertainty visualization, VR-based anatomical visualization, and AI-assisted segmentation editing separately, but these capabilities have not been integrated into a single framework.

Table 1 summarizes the representative approaches discussed above and highlights the combination of capabilities provided by the proposed framework.

Table 1. Comparison of representative approaches discussed in the related work.

Approach	VR	AI	Conf.	Inter.	Calib.
Static uncertainty visualization [29]	No	Yes	Yes	No	No
VR anatomical visualization [11,12]	Yes	No	No	No	No
VR-based AI segmentation editing [14–16]	Yes	Yes	No	No	No
Proposed framework	Yes	Yes	Yes	Yes	Yes

Bold indicates the framework proposed in this study.

3. Framework

This section describes the proposed framework, including the medical segmentation model, confidence calibration, and the VR system used for interactive visualization.

3.1. Medical Segmentation

The segmentation backbone is a 3D U-Net [18,19] trained for binary segmentation, in which the submandibular gland and the mandibular canal are unified into a single foreground class versus background. The network follows the standard encoder–decoder

structure with skip connections, which preserves fine-grained spatial information while capturing multi-scale contextual features. The encoder progressively downsamples the input volume through stacked convolutional blocks, whereas the decoder reconstructs the segmentation through upsampling and feature concatenation. Batch normalization and ReLU activations are applied throughout the network to improve training stability and convergence. The model takes a single 3D MRI volume as input and produces a voxel-wise probability map representing the likelihood of each voxel belonging to the foreground structure.

Before training, all MRI volumes were preprocessed to ensure consistent geometry and intensity. The original anisotropic volumes (voxel spacing of $0.5 \times 0.5 \times 2$ mm) were resampled to an isotropic spacing of 0.5 mm using linear interpolation and cropped to a fixed spatial size of $384 \times 384 \times 384$ voxels, corresponding to a 192 mm field of view that focuses on the head and neck region. Non-zero voxel intensities were normalized to the range $[0, 1]$, and no aggressive smoothing was applied so that anatomically relevant boundaries were preserved. Binary ground-truth masks were used for training, with a voxel value of 1 indicating the target structure and 0 indicating background.

The model was implemented in PyTorch (version 2.2.1) [30] and trained on an NVIDIA V100 GPU (NVIDIA Corporation, Santa Clara, CA, USA) with 32 GB of memory. A composite objective combining the Dice loss and a (weighted) cross-entropy loss was used to address the strong foreground–background class imbalance and to improve boundary accuracy [31]. Optimization was performed with the Adam optimizer at an initial learning rate of 10^{-3} . Training was run for up to 300 epochs with early stopping when no improvement in validation performance was observed for 100 consecutive epochs, and the checkpoint with the highest validation Dice Similarity Coefficient (DSC) was retained for inference. To increase robustness to spatial and intensity variability, data augmentation was applied during training, including affine transformations (rotation, scaling, translation), axis flipping, contrast and intensity adjustments, Gaussian noise, and light smoothing.

At inference, the softmax output yields voxel-wise probabilities in the range $[0, 1]$, which are stored as 3D confidence maps in NIfTI format for use in the VR application. A connected-component analysis step retains only the largest foreground regions corresponding to the target structures, suppressing false positive predictions and improving usability in the immersive environment. This post-processing is applied only at inference and not during training. The resulting voxel-wise confidence maps provide a continuous representation of model certainty and form the basis for the interactive thresholding described below.

The segmentation model was evaluated on the held-out test cases across the five cross-validation folds. Across the full test set ($n = 55$), the model achieved a Dice Similarity Coefficient of 0.82 ± 0.09 , Intersection over Union of 0.71 ± 0.12 , precision of 0.83 ± 0.09 , and recall of 0.83 ± 0.12 . These results indicate that the confidence maps used in the VR interface were generated from segmentations with reasonable agreement with the expert annotations.

3.2. Temperature Scaling

Raw softmax probabilities produced by deep segmentation networks are frequently miscalibrated, meaning that the predicted confidence does not match the empirical likelihood of correctness; this can create a false sense of certainty and lead clinicians to over- or under-trust the output. To address this, we applied temperature scaling as a post hoc, single-parameter calibration step that adjusts the sharpness of the probability distribution without modifying the trained model weights or the predicted class.

Given the network logits z , temperature scaling divides the logits by a scalar temperature $T > 0$ before the softmax, as shown in Equation (1):

$$\hat{p} = \text{softmax}\left(\frac{z}{T}\right), \quad (1)$$

so that $T = 1$ recovers the original probabilities, $T > 1$ softens (less confident) and $T < 1$ sharpens (more confident) the distribution. The optimal temperature T was obtained by minimizing the negative log-likelihood (NLL) on a held-out calibration set, separate from the training, validation, and test data. As shown in Equation (1), increasing or decreasing the temperature adjusts the confidence values without changing the predicted class. Because T does not affect the argmax decision, temperature scaling adjusts the confidence values used for thresholding while leaving the underlying segmentation accuracy unchanged. The calibrated voxel-wise probabilities are then exported as the confidence maps that drive the VR thresholding interface.

Calibration was evaluated on the postprocessed voxel-wise confidence maps exported to the VR application, as these were the confidence values available to clinicians during the user study. Across the 55 test cases, the deployed confidence maps had an Expected Calibration Error (ECE) of 0.0005 and a Brier Score of 0.0007.

3.3. VR Interface

We developed an interactive VR system using Unity (version 2019.4.12f1; Unity Technologies, San Francisco, CA, USA) and OpenXR (version 1.0.0; The Khronos Group, Beaverton, OR, USA) [32], compatible with standard headsets, for interactive 3D visualization of MRI segmentations. The interface displays segmentation maps of the mandibular structures within the volumetric MRI data. Clinicians can manually adjust lower and upper confidence thresholds using sliders, allowing dynamic exploration of segmentation regions within the selected confidence range in an immersive 3D environment. This enables users to directly observe how the segmentation changes in real time based on different confidence values. The system provides two interaction modes. In the single-threshold mode, the user adjusts one threshold to control the visible segmentation. In the range-based mode, the user defines both minimum and maximum thresholds, allowing exploration of a specific confidence interval. This interaction makes it easier to inspect uncertain regions and boundaries. The interface is implemented using the Mixed Reality Toolkit (MRTK version 2.5.1; Microsoft Corporation, Redmond, WA, USA) and supports real-time interaction, including volume navigation, threshold adjustment, and visualization updates. The menus are designed to be simple and accessible, allowing users to interact with the system without interrupting the immersive experience.

An overview of the application is shown in Figure 1. Each MRI volume and its corresponding confidence map are imported as 3D textures and rendered with a custom raycasting shader that supports opacity blending, clipping, and threshold filtering. Confidence values are visualized through intensity-based color mapping, where higher-confidence voxels appear more opaque and lower-confidence voxels more transparent. Rendering performance was optimized using adaptive sampling and GPU-based volume caching to maintain real-time frame rates on a desktop RTX GPU as well as on a Meta Quest 3 headset (Meta Platforms, Inc., Menlo Park, CA, USA).

The VR environment includes two main menus, designed to support intuitive interaction while keeping the workspace uncluttered. The first menu (Figure 2) contains the core functional controls: Show Raycast Volume, a mode switch between the 3D constructed view and the radiological rendering mode; Show Segmentation, a visibility toggle to hide or reveal the segmentation; Toggle Segment Mode, which selects different visualization states for the start and end layers of the segmentation; X-ray Flashlight, a dynamic clipping

plane for inspecting internal structures; Save Threshold Values, which records the current threshold settings for analysis; Rate, two rating buttons for evaluating the tool and the overall system; and Next Image, which loads the next scan and its segmentation.

The second menu, the image adjustment menu (Figure 3), contains sliders for brightness, contrast, and threshold control, together with a reset button that restores the default values. In the single-threshold mode, the panel exposes a single threshold slider (Figure 3a), whereas in the range mode it exposes both a minimum and a maximum threshold slider (Figure 3b). Both modes are implemented with MRTK sliders linked to the shader's `_ThresholdMin` and `_ThresholdMax` parameters, providing immediate visual feedback as the thresholds change. To ensure valid interaction, the minimum threshold is constrained so that it cannot exceed the maximum, and a small gap is enforced between the two values to prevent overlap. When either slider is adjusted, the updated threshold range is propagated to the rendering system through an event-driven mechanism. Both menus are floating panels positioned within comfortable reach, and users can reposition the menus and the volume freely to obtain a comfortable working layout.

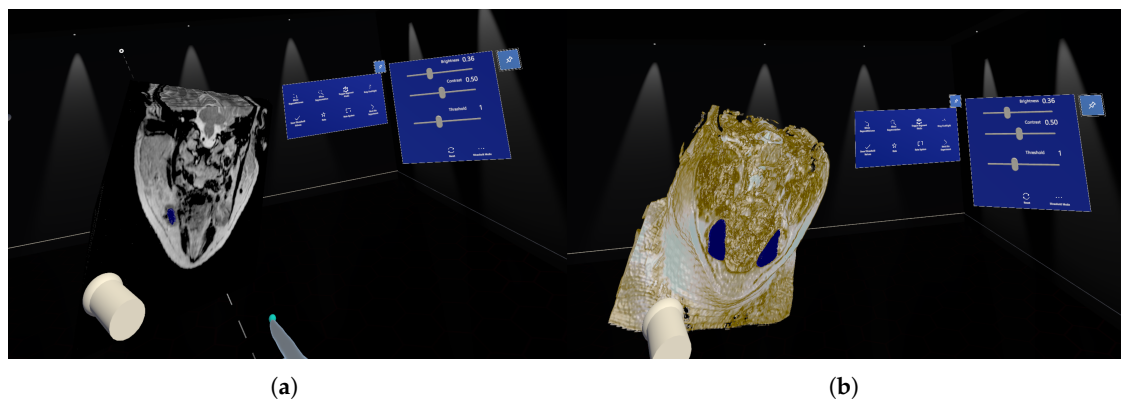


Figure 1. Overview of the VR application. (a) Radiological rendering mode showing an MRI slice view with the predicted segmentation and the floating control panels. (b) 3D reconstructed view of the head-and-neck volume with the segmented mandibular structures highlighted.

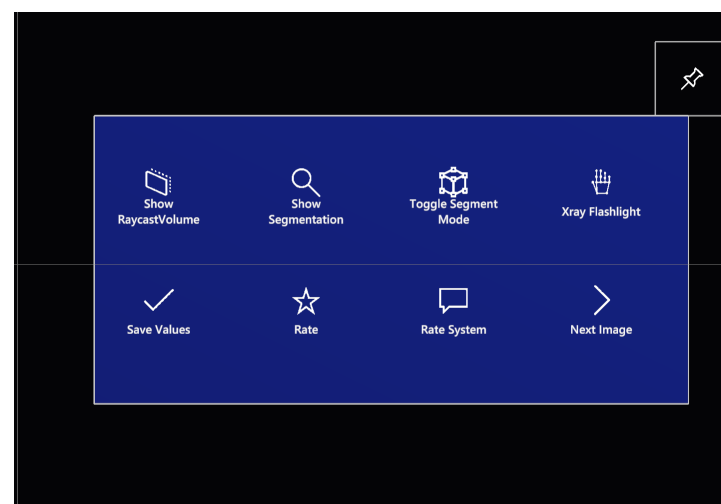


Figure 2. Interaction menu containing the core functional controls of the VR application.

3.4. Dataset

We utilize the AAPM RT-MAC Grand Challenge 2019 dataset [33], which contains 55 T2-weighted MRI scans from head and neck patients for radiotherapy planning. Our study targets the mandibular region, unifying the submandibular gland and mandibular canal into a single foreground class versus background, leveraging clinicians' anatomical

knowledge to distinguish these structures. The dataset provides submandibular gland annotations, while mandibular canal segmentations were manually created by a radiologist with 25 years of experience. Since the primary objective of this work was to evaluate the proposed VR interaction technique, the user study focused on a single confidence-thresholding workflow. Future work will extend the framework to multi-class segmentation.

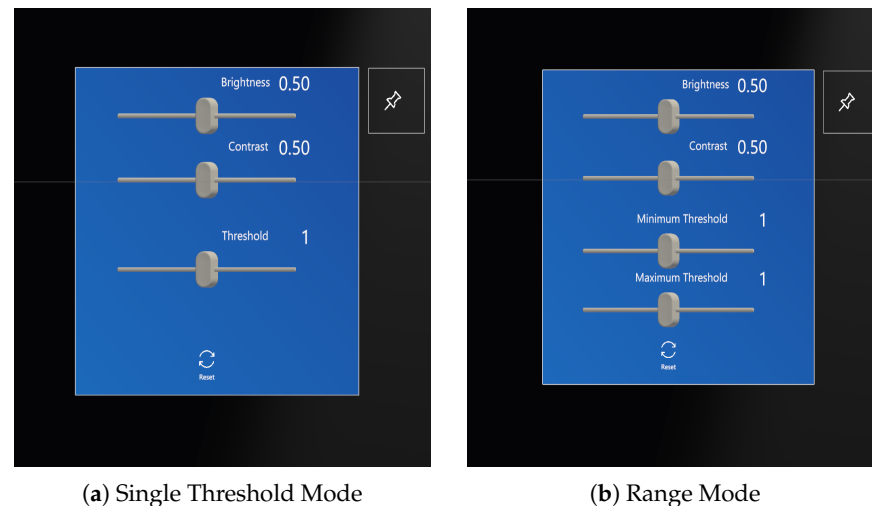


Figure 3. Image adjustment menu. (a) Single-threshold mode with a single threshold slider. (b) Range mode with separate minimum- and maximum-threshold sliders.

We employ 5-fold cross-validation, splitting each fold into 22 images for training, 11 for validation to guide augmentation, hyperparameter tuning, and model selection, 11 for calibration to perform temperature scaling, and 11 for testing. Aggregating folds yields a test set covering all 55 MRI images, ensuring robust evaluation. For the usability study, all six clinicians evaluated 50 of the 55 images; five images were excluded to keep the session duration manageable, limiting VR-related fatigue and maintaining the reliability of the usability ratings.

4. Experiments

This section describes the experimental setup and the usability study used to evaluate the proposed framework.

4.1. Participants

Six clinicians with expertise in medical and dental imaging participated, with clinical experience ranging from 1 to over 40 years. The group included four women and two men, aged between 30 and 65 years, all right-handed, with professional backgrounds spanning radiology, dentistry, dental instruction, and medical engineering. Most participants had tried VR once or a few times before, while one participant had long-term experience with immersive technologies. This mix of expertise allowed us to assess usability across a range of familiarity with both medical imaging and VR. To preserve participant anonymity, the results are presented using anonymized user IDs rather than individual professional roles or years of clinical experience. All participants signed a consent form prior to the experiment, and participation was uncompensated.

4.2. Usability and Trust Study

The user study used a within-subjects design. Each participant evaluated 50 MRI cases under both threshold interaction modes (single-threshold and range), with the order of conditions counterbalanced across participants to control for learning and order effects.

For each case, participants were instructed to explore the volume freely, interact with the VR system by adjusting the threshold(s), and identify the most clinically reasonable segmentation boundary based on the visualized confidence.

Usability was measured using the UMUX-Lite questionnaire (two items: perceived ease of use and usefulness, each on a 7-point scale), administered for each evaluated image. From the two UMUX-Lite items, we additionally computed a predicted System Usability Scale (SUS) score using the established UMUX-Lite-to-SUS regression equation [34]; this provides an SUS-equivalent value on the familiar 0–100 scale but is not the original 10-item SUS instrument. Because ratings were collected per image, each participant contributed a distribution of scores in each interaction mode (25 per mode). These repeated observations were analyzed within participants to describe participant-specific differences between the single- and range-threshold modes. The per-image paired tests were not used as population-level evidence across clinicians. For each participant, per-image scores were compared using a paired *t*-test when the paired differences satisfied a Shapiro–Wilk normality check, and a Wilcoxon signed-rank test otherwise. Error bars in Figure 4 therefore represent the standard deviation of per-image scores within each participant. Additional metrics include interaction time, selected threshold values, and user preference. Qualitative feedback was collected through a think-aloud protocol to better understand user experience, trust, and interpretation of the model outputs. Here, we use the term trust to refer to participants' expressed confidence in, and willingness to rely on, the segmentation output, assessed qualitatively through their think-aloud comments and stated mode preference rather than with a standardized trust questionnaire.

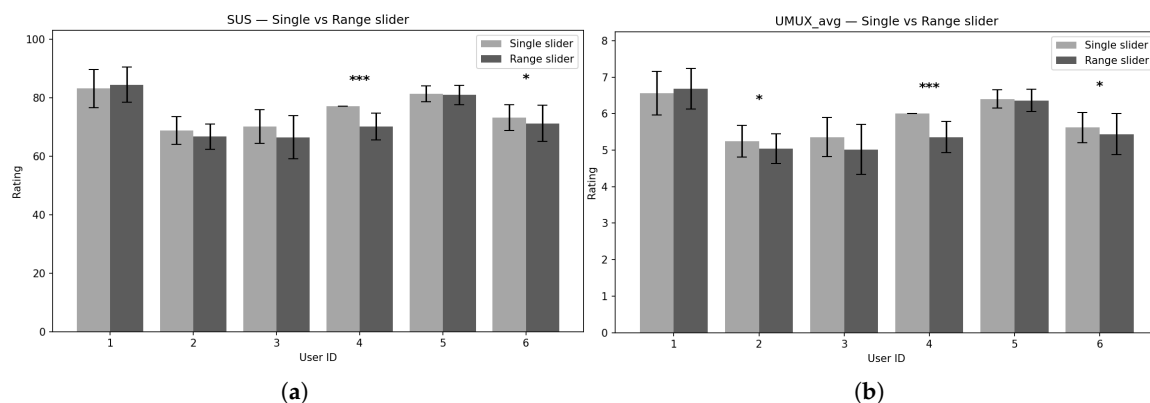


Figure 4. Usability ratings comparing single and range slider modes. (a) System Usability Scale (SUS) scores per participant; error bars represent standard deviation. (b) UMUX-Lite average scores per participant. Asterisks indicate statistically significant within-participant differences between modes (paired *t*-test or Wilcoxon signed-rank test, as appropriate): * $p < 0.05$, *** $p < 0.001$.

4.3. Implementation

The segmentation model is implemented in PyTorch and trained using GPU hardware. The VR system is developed in Unity using OpenXR and the Mixed Reality Toolkit (MRTK), allowing interactive 3D visualization of MRI data and segmentation results. The system supports real-time interaction, including volume navigation, threshold adjustment, and dynamic updates of the segmentation based on user input.

5. Results

This section reports the findings of the user study. Given the limited number of participants, the results should be interpreted as preliminary evidence of usability and interaction behavior rather than broadly generalizable conclusions. We first present UMUX-Lite ratings

and estimated SUS scores, followed by interaction time, threshold-selection behavior across the two interaction modes, and qualitative feedback from the think-aloud sessions.

5.1. Usability Ratings

Both interaction modes were rated as usable. Estimated SUS scores derived from the UMUX-Lite responses were generally high across participants. Differences between the single-threshold and range modes were small for most participants. Two participants showed a slight preference for the range mode, while two participants reported lower usability ratings for the range mode. One participant rated the single-threshold mode significantly higher than the range mode ($p < 0.001$). UMUX-Lite ratings followed a similar pattern, with mean values ranging from 5.0 to 6.7 on the 7-point scale, indicating satisfactory perceived usefulness and ease of use across both modes. The per-participant estimated SUS and UMUX-Lite scores are shown in Figure 4.

5.2. Interaction Time

Mean interaction time per case ranged from approximately 85 to 190 s across participants, with substantial individual variation that reflected differences in exploration behavior and prior VR experience. Two participants completed the task significantly faster with the range mode (156 s $\rightarrow$ 72 s, $p = 0.020$; 108 s $\rightarrow$ 84 s, $p = 0.004$), while the remaining participants showed comparable durations across modes. Importantly, the additional control offered by the range mode did not impose a consistent time penalty: at the group level, providing two thresholds did not slow participants down. The per-participant interaction times are shown in Figure 5. Interaction time should not be interpreted as a measure of task quality. Faster interaction may reflect either greater efficiency or less thorough exploration. Future studies should combine interaction time with objective measures of task performance and segmentation quality.

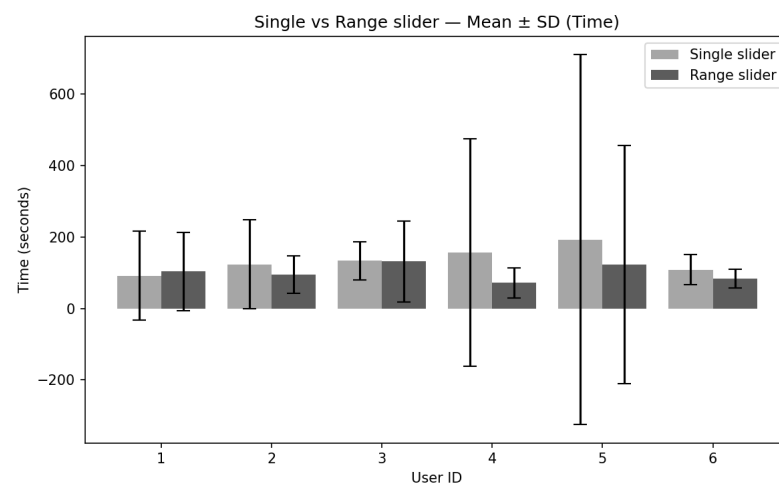


Figure 5. Average time per case (in seconds) under the single and range slider conditions. Error bars represent standard deviation.

Some participants showed greater variation in interaction time, likely reflecting differences in how they explored the images and the difficulty of individual cases. Future studies should examine these variations in more detail and consider reporting median interaction times in addition to mean values.

A summary of the usability and threshold-selection results across the two interaction modes is given in Table 2.

Table 2. Summary of usability and threshold-selection results across the two interaction modes ($N = 6$). Threshold values are reported on the normalized $[0, 1]$ slider scale.

Metric	Single-Threshold	Range
Estimated SUS score (Mean $\pm$ SD)	75.63 $\pm$ 5.90	73.31 $\pm$ 7.59
UMUX-Lite score (Mean $\pm$ SD)	5.86 $\pm$ 0.55	5.65 $\pm$ 0.70
Time per case (s, Mean $\pm$ SD)	134.47 $\pm$ 36.22	101.57 $\pm$ 22.84
Minimum threshold (Mean $\pm$ SD)	0.00 (fixed)	0.29 $\pm$ 0.17
Maximum threshold (Mean $\pm$ SD)	0.64 $\pm$ 0.27	0.82 $\pm$ 0.17
Range width (Mean $\pm$ SD)	0.64 $\pm$ 0.27	0.54 $\pm$ 0.28

5.3. Threshold Selections

Because the thresholds were applied to the model's voxel-wise confidence maps, the recorded values represent subjective filtering behavior rather than segmentation accuracy. Most voxel confidence values across the 50 images fell below 0.35, and only a small proportion exceeded 0.70. The thresholds chosen by participants therefore indicate how they navigated and filtered the probabilistic output of the model.

Minimum threshold values showed a clear shift once participants were given control over the lower bound. In the single-threshold mode, the minimum was fixed at zero, whereas in the range mode, all participants selected a non-zero minimum (user-wise means from 0.15 to 0.50; Figure 6a), which removed low-confidence voxels. Statistical analysis showed significant increases in *thr_min* for five of the six participants ($p < 0.001$), which, for these participants, appears associated with a controllable lower threshold changing how they engaged with the confidence map when they could suppress low-confidence regions.

Maximum threshold values were also consistently higher in the range mode than in the single mode (Figure 6b), with statistically significant increases for three participants ($p < 0.001$). Because confidence values above 0.70 represent the upper end of the model output, these higher upper bounds may suggest that, within this exploratory sample, participants tended to concentrate on regions where the model expressed its greatest certainty; in the single-threshold mode, the inability to adjust the lower bound appears to have been compensated by a more conservative upper bound.

The width of the selected interval (Figure 6c) was generally broader in the single-threshold mode, where the minimum was fixed at zero, and narrower in the range mode, where participants typically focused on intermediate-to-high confidence voxels. Significant reductions in range width were observed for three participants ($p < 0.01$), reflecting more selective filtering of uncertainty when both limits could be controlled. Consistently, the average threshold (Figure 6d) was higher in the range mode for all participants ($p < 0.01$). Given that the confidence distribution is heavily skewed toward lower values, average threshold values between roughly 0.42 and 0.71 already represent selective filtering of the more reliable predictions, and several participants noted that the range mode let them “clean up” the visualization and more confidently isolate the intended structure. The selected thresholds were not compared with expert annotations, so they should not be considered the best thresholds for clinical use. Future studies should examine how the selected thresholds relate to clinically meaningful anatomical boundaries.

5.4. Qualitative Results

Qualitative feedback suggests that participants felt more confident when interpreting AI-generated segmentations with confidence visualization enabled. Participants report that adjusting thresholds helps them better understand model behavior and identify uncertain regions. Overall, five of six participants expressed a preference for the interactive confidence-thresholding approach.

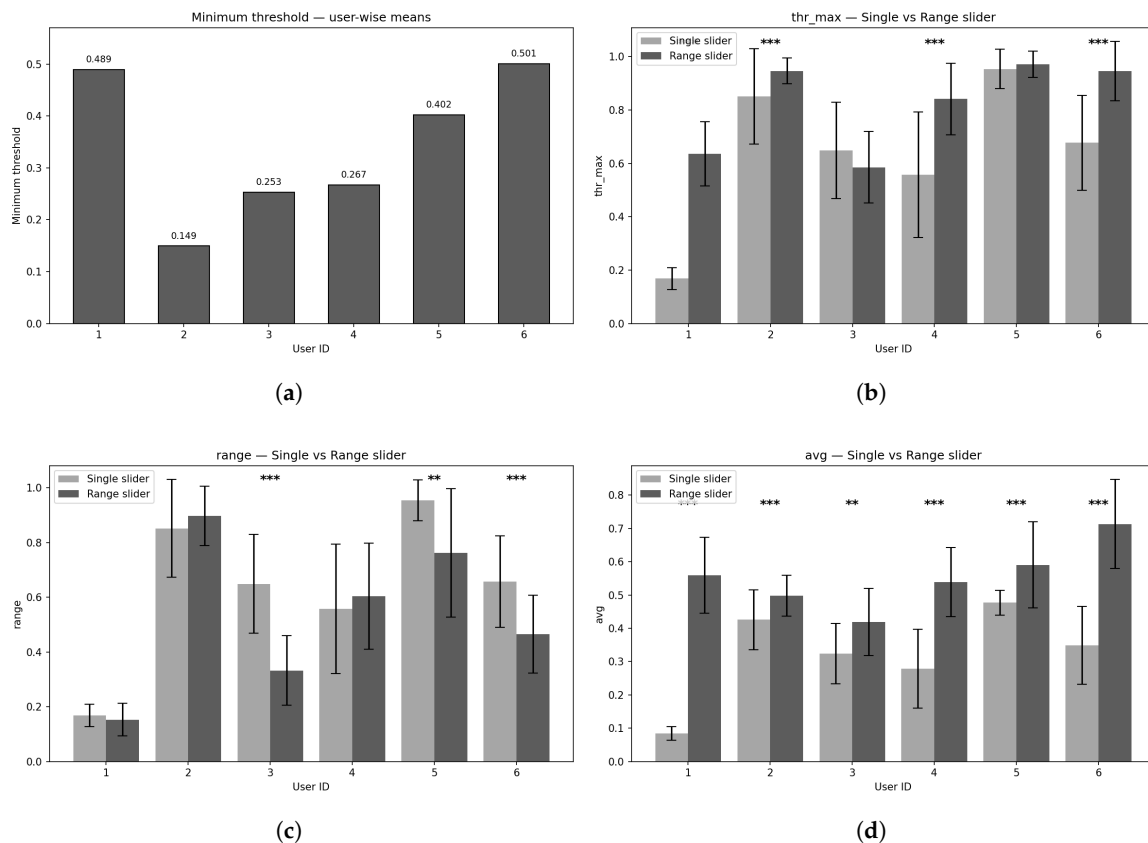


Figure 6. Threshold-selection behavior across the six participants. (a) Minimum threshold (thr_{min}) values in the range slider mode. (b) Maximum threshold (thr_{max}) comparison between single and range slider modes. (c) Threshold range width comparison between modes. (d) Average threshold comparison between modes. Asterisks indicate statistically significant within-participant differences between modes (paired t -test or Wilcoxon signed-rank test, as appropriate): ** $p < 0.01$, *** $p < 0.001$.

The participants' think-aloud data were also analyzed through a thematic analysis [35]. Based on the coded data, two main categories were formed: Threshold Tool and Medical Image. In the Threshold Tool category, the three subcategories found are the following: Usability, Feasibility, and Preferences. In the Medical Image category, the two subcategories found are Adjustments and Interpretability. The results are discussed below.

5.4.1. Threshold Tool

The participants had some unified views regarding the threshold tool's usability. Three out of six participants commented that the threshold usage was either easier (P4), faster (P5), or logical (P1) after they gained some experience in using the tool. P2 assumed that the segmentation clarity might depend on the method of how the thresholding is done. They also commented that it is easier to manipulate the sliders with the primary hand, highlighting that the interaction needs enough precision to place the threshold values into a correct setting.

For the threshold tool's feasibility, three out of six participants had conflicting views. P5 said that the tool seems "pretty useful", whereas P4 commented that the segmentation was already in a good threshold setting without modifications:

"I think segmentations are already in good threshold." (P4)

Furthermore, P2 felt that there was not enough distinguishable difference between the threshold values:

“I don’t feel much difference in what I am doing regarding of confidence values.” (P2)

Two out of six participants had strong preferences regarding the threshold tool. P1 stated that the threshold should be just barely visible; otherwise, the area visibility is compromised. They continued further:

“Seeing the faint border between mandibular gland and fat tissue is what I do, really thin border.” (P1)

P3 expressed to prefer the tool that only showed one threshold, as it was quicker and easier to use, “working wonderfully”, although they felt that the tool with both the minimum and maximum threshold sliders might be more usable in some work tasks. They elaborated:

“I like the tool where just one threshold is shown because its quick and easy. Even though the min max (tool with minimum and maximum threshold sliders) is more precise it’s also more complicated and takes time.” (P3)

5.4.2. Medical Image

For the medical image adjustments, P2 commented that in approximately half of the images, they “realized again how to use flashlight tool”, which is a tool for slicing the medical 3D image in different angles by grabbing the tool and pointing the light towards the image. They further elaborated that they “somehow just did not bother to use it”. P3 expressed their frustration related to the grabbing order of the flashlight and the medical image:

“Everytime the image is grabbed it makes it hard to grab the flashlight, because of this it messes up the anatomical angles.” (P3)

P3 proposed that the flashlight would be the only grabbable object by default, and that there would be a toggle button turning the image grabbing on and off.

Regarding the image angles, P4 expressed that toggling between the two render modes (slice view, 3D model view) helps with deciding whether the image angle is correct or not. P3 ideated that each anatomical angle would have its own button in order to make the angle changes more easier:

“It would be the best if there were buttons for axial, sagittal, and coronal angles so I would not have to mess up with the image rotation. For now, I will just stick to the axial direction that I managed to place the image, although the opposite view would be better. Otherwise I fear that I ruin the image placement again.” (P3)

Furthermore, P1 stated that when there is a more complex case, the image adjustments (e.g., brightness, contrast, image angle) take more time, making the evaluation task more difficult. P4 also noted that the brightness settings had more visible change to the image than contrast.

Two participants out of six mentioned problems related to image interpretability. P4 was unsure about whether the image angle they were looking at was correct:

“Am I looking at the right angle?” (P4)

P4 was also unsure about what is the segmentation in the image, whereas P1 mentioned that the segmentation is sometimes completely in a wrong place, which makes it harder to evaluate the image.

6. Discussion

6.1. Clinical Impact

The results suggest, in this exploratory study, that interactive confidence thresholding may improve the interpretability of AI-generated segmentations. Prior research has demonstrated that visualizing uncertainty can improve trust and decision-making in AI-assisted systems [36]. However, simply adding uncertainty to visualizations can be insufficient and may even lead to misinterpretation [37]. Our findings extend this literature by showing that interactive manipulation of uncertainty, rather than passive visualization alone, could enable more effective interpretation. In our software, the clinicians can explore how the segmentation changes across different confidence levels. This allows them to better understand model behavior and identify uncertain regions. The VR environment further supports this process by providing an intuitive way to interact with volumetric data. The combination of immersive visualization and user-controlled thresholding may help users better interpret the model output. However, whether this translates into improved clinical decision-making remains outside the scope of the present study.

In the context of medical imaging, uncertainty has been recognized as a critical factor influencing clinical decisions [29], yet practical tools for interacting with it remain limited. The behavior observed in our study, where users selectively filtered low-confidence predictions and focused on higher-confidence regions, aligns with theories of confidence-based trust calibration, which suggest that optimal decision-making relies on thresholding AI confidence [38].

Our results also suggest that interactive, immersive approaches to uncertainty may be a key enabler for clinically usable eXplainable AI systems, consistent with evidence that interactive visualization tools are closely tied to trustworthiness and reliable interpretation in healthcare contexts [39].

The confidence values represent calibrated voxel-wise probabilities rather than segmentation quality. Lower confidence is expected in uncertain regions, and temperature scaling produces more reliable confidence estimates.

However, interactive uncertainty tools alone are not sufficient to ensure clinically usable AI. Combined with other elements of explainable AI, such as interactive segmentation editing that allows users to contest and refine AI outputs [15], these approaches can support more comprehensive human–AI collaboration and facilitate appropriate trust calibration. VR seems to be not only a suitable platform for 3D visualizations but also for interactive AI workflows [14,15]. The findings of this study suggest that VR can be a suitable platform for interactive confidence visualizations.

The qualitative findings also identified several usability issues, including slider precision, image orientation, and flashlight interaction. These observations provide clear directions for future interface improvements and should be considered in the next iteration of the system.

At the same time, interactive confidence thresholding also carries potential risks. If users set the minimum threshold too high, low-confidence regions corresponding to clinically relevant anatomy or pathology may be hidden from view, potentially reinforcing confirmation bias by emphasizing only predictions the model considers highly confident. The proposed system is intended as an exploratory visualization tool rather than a diagnostic or decision-support system, and users can freely adjust the thresholds at any time to inspect both low- and high-confidence regions. Future studies should investigate how different confidence visualization strategies influence user behavior, confirmation bias, and clinical decision-making.

In particular, the range-based interaction enables more detailed exploration of uncertain areas compared to single-threshold approaches. This suggests that flexible interaction with confidence is an important factor in perceived confidence and usability.

6.2. Limitations

This study has several limitations. First, the number of participants was relatively small, which limits the generalizability and statistical power of the findings. Recruiting clinicians with relevant imaging expertise was challenging, and relatively small participant groups are common in this type of usability research. Although each clinician evaluated 50 MRI images across the two interaction conditions, providing repeated observations over a diverse set of cases and reducing the influence of image-specific effects, these repeated evaluations do not replace a larger group of independent participants. The findings should be considered preliminary and require validation in a larger group of clinicians. Second, the evaluation focused on a single anatomical region, and the findings may not directly transfer to other segmentation tasks. Interpretation of medical image segmentations is inherently subjective and may vary between clinicians. Future work should include multiple independent expert evaluations to better assess inter-observer variability.

Because each participant evaluated multiple images, the observations were not fully independent. Future studies with larger samples should use mixed-effects models or participant-level analyses to account for this.

Because the study included only six clinicians, subgroup analyses based on clinical experience were not statistically meaningful and could also have reduced participant anonymity.

Additionally, the study focuses primarily on usability and perceived confidence rather than quantitative improvements in segmentation accuracy. The selected thresholds were not compared with expert reference annotations using objective segmentation metrics (e.g., Dice Similarity Coefficient). This study does not determine whether interactive thresholding improves anatomical boundary accuracy or clinical decision-making. Because the objective of this study was to evaluate the usability of the proposed VR interaction techniques, a conventional 2D viewing condition or fixed-threshold baseline was not included. Future work will incorporate these comparisons to assess their impact on clinical decision-making. Future work will compare user-generated segmentations with reference annotations using these metrics.

Although calibration was evaluated using ECE and Brier Score on the postprocessed confidence maps used in the VR interface, the target structures occupy only a small fraction of the full 3D volumes. Because these metrics are computed over all voxels, they are dominated by the background class, which greatly outnumbers foreground voxels, and should not be taken as evidence of near-perfect calibration in the clinically relevant gland and canal regions. Therefore, the reported values should be interpreted as an overall characterization of the deployed confidence maps. Future work should include region-specific calibration analyses. Perceived confidence was also not measured using a standardized trust questionnaire, so these findings should be interpreted as qualitative observations rather than validated measures of trust.

In addition, effect sizes and confidence intervals were not reported in the present study. Future studies should include these measures alongside significance tests to better quantify the practical relevance of the observed differences.

6.3. Future Work

Future work should include larger user studies with more diverse clinical participants to validate the findings. Extending the system to multi-class segmentation and different

anatomical regions would also improve its applicability. In addition, integrating other uncertainty estimation methods and exploring alternative interaction techniques could further enhance the usability and effectiveness of the system.

7. Conclusions

This work presents a framework for interactive confidence visualization in VR for medical image segmentation. By combining calibrated voxel-wise confidence with user-controlled thresholding, the system allows users to explore segmentation results beyond fixed binary outputs.

The results suggest that this approach may improve interpretability and may support clinicians' perceived confidence, although its effect on segmentation accuracy and clinical decision-making remains to be evaluated. Users can better understand model behavior and identify uncertain regions by interacting with confidence values in real time.

Overall, enabling interaction with model confidence, rather than only visualizing it, shows promise as an approach for supporting decision-making in AI-based medical imaging systems.

Author Contributions: Conceptualization, N.R., L.O., J.K., M.S. and R.R.; methodology, N.R., M.S., J.S. and K.K.; software, N.R. and M.S.; validation, N.R., L.O. and J.J.; formal analysis, N.R.; investigation, N.R.; resources, J.J., K.K. and R.R.; data curation, M.S. and J.S.; writing—original draft preparation, N.R.; writing—review and editing, M.S., L.O., J.K., J.J., J.S., K.K. and R.R.; visualization, N.R.; supervision, J.K. and R.R.; project administration, R.R.; funding acquisition, K.K. and R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Council of Finland grant number 80/31/2023, as part of the MEDALLION project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The AAPM RT-MAC Grand Challenge 2019 dataset analyzed in this study is publicly available from The Cancer Imaging Archive at <https://doi.org/10.7937/tcia.2019.bcfjqfqb> (reference [33]). The interaction logs and questionnaire data generated during the user study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy restrictions concerning the study participants.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5; OpenAI, San Francisco, CA, USA) and Grammarly (version 14.1258.0; Grammarly, Inc., San Francisco, CA, USA) for the purposes of language editing, sentence restructuring, and LaTeX formatting support. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CBCT	Cone Beam Computed Tomography
CT	Computed Tomography
DSC	Dice Similarity Coefficient
MRI	Magnetic Resonance Imaging
MRTK	Mixed Reality Toolkit
NIfTI	Neuroimaging Informatics Technology Initiative
NLL	Negative Log-Likelihood

SUS System Usability Scale
VR Virtual Reality

References

- Black, W.C.; Welch, H.G. Advances in Diagnostic Imaging and Overestimations of Disease Prevalence and the Benefits of Therapy. *N. Engl. J. Med.* **1993**, *328*, 1237–1243.
- Meesters, A.M.L.; Duis, K.; Banierink, H.; Stirler, V.M.A.; Wouters, P.C.R.; Kraeima, J.; Vries, J.-P.P.M.; Witjes, M.J.H.; IJpma, F.F.A. What Are the Interobserver and Intraobserver Variability of Gap and Stepoff Measurements in Acetabular Fractures? *Clin. Orthop. Relat. Res.* **2020**, *478*, 2801–2808.
- Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; Xie, Y. From CNN to Transformer: A Review of Medical Image Segmentation Models. *J. Digit. Imaging* **2024**, *37*, 1529–1547.
- Recht, M.P.; Dewey, M.; Dreyer, K.; Langlotz, C.; Niessen, W.; Prainsack, B.; Smith, J.J. Integrating Artificial Intelligence into the Clinical Practice of Radiology: Challenges and Recommendations. *Eur. Radiol.* **2020**, *30*, 3576–3584.
- Nikolov, S.; Blackwell, S.; Zverovitch, A.; Mendes, R.; Livne, M.; De Fauw, J.; Patel, Y.; Meyer, C.; Askham, H.; Romera-Paredes, B.; et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *J. Med. Internet Res.* **2021**, *23*, e26151.
- Ong, C.S.; Deib, G.; Yesantharao, P.; Qiao, Y.; Pakpoor, J.; Hibino, N.; Hui, F.; Garcia, J.R. Virtual Reality in Neurointervention. *J. Vasc. Interv. Neurol.* **2018**, *10*, 17–22.
- Liu, X.; Huang, D.; Yao, J.; Dong, J.; Song, L.; Wang, H.; Yao, C.; Chu, W. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* **2025**, *6*, 285.
- Yang, G.; Ye, Q.; Xia, J. Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond. *Inf. Fusion* **2022**, *77*, 29–52.
- Neri, E.; Aghakhanyan, G.; Zerunian, M.; Gandolfo, N.; Grassi, R.; Miele, V.; Giovagnoni, A.; Laghi, A. Explainable AI in Radiology: A White Paper of the Italian Society of Medical and Interventional Radiology. *Radiol. Med.* **2023**, *128*, 755–764.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115.
- Choi, K.-S.; Chan, S.-T.; Leung, C.H.-M.; Chui, Y.-P. Stereoscopic Three-Dimensional Visualization for Immersive and Intuitive Anatomy Learning. In *Proceedings of the 2016 IEEE Eighth International Conference on Technology for Education (T4E), Mumbai, India, 2–4 December 2016*; IEEE: New York, NY, USA, 2016; pp. 184–187.
- Orsmaa, L.; Kangas, J.; Rasouli, N.; Jaskari, J.; Sahlsten, J.; Mehtonen, H.; Järnstedt, J.; Kaski, K.; Raisamo, R. Diving, Grabbing and Teleporting: Methods for Medical 3D Image Manipulation in VR. *Interact. Comput.* **2025**, *37*, 105–119.
- Rakkolainen, I.; Farooq, A.; Kangas, J.; Hakulinen, J.; Rantala, J.; Turunen, M.; Raisamo, R. Technologies for Multimodal Interaction in Extended Reality—A Scoping Review. *Multimodal Technol. Interact.* **2021**, *5*, 81.
- Spiegler, P.; Harirpoush, A.; Xiao, Y. Towards User-Centered Interactive Medical Image Segmentation in VR with an Assistive AI Agent. *Virtual Real.* **2025**, *30*, 20. <https://doi.org/10.1007/s10055-025-01284-0>.
- Orsmaa, L.; Saukkoriipi, M.; Kangas, J.; Rasouli, N.; Järnstedt, J.; Mehtonen, H.; Sahlsten, J.; Jaskari, J.; Kaski, K.; Raisamo, R. Interactive AI Annotation of Medical Images in a Virtual Reality Environment. *Int. J. Comput. Assist. Radiol. Surg.* **2025**, *21*, 49–58. <https://doi.org/10.1007/s11548-025-03497-9>.
- Jones, Z.; Drouin, S.; Kersten-Oertel, M. Exploring Interaction Paradigms for Segmenting Medical Images in Virtual Reality. *Int. J. Comput. Assist. Radiol. Surg.* **2025**, *20*, 1713–1721.
- Xu, Y.; Quan, R.; Xu, W.; Huang, Y.; Chen, X.; Liu, F. Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. *Bioengineering* **2024**, *11*, 1034.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Cham, Switzerland, 2015; pp. 234–241.
- Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Cham, Switzerland, 2016; pp. 424–432.
- Kerfoot, E.; Clough, J.; Oksuz, I.; Lee, J.; King, A.P.; Schnabel, J.A. Left-Ventricle Quantification Using Residual U-Net. In *Statistical Atlases and Computational Models of the Heart (STACOM 2018)*; Springer: Cham, Switzerland, 2019; pp. 371–380.
- Saukkoriipi, M.; Sahlsten, J.; Jaskari, J.; Al-Tahmeesschi, A.; Ruotsalainen, L.; Kaski, K. Head and Neck Tumor Segmentation Using Pre-RT MRI Scans and Cascaded DualUNet. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*; Springer: Cham, Switzerland, 2024; pp. 191–203.

22. Järnstedt, J.; Sahlsten, J.; Jaskari, J.; Kaski, K.; Mehtonen, H.; Lin, Z.; Hietanen, A.; Sundqvist, O.; Varjonen, V.; Mattila, V.; et al. Comparison of Deep Learning Segmentation and Multigrader-Annotated Mandibular Canals of Multicenter CBCT Scans. *Sci. Rep.* **2022**, *12*, 18598.
23. Wahid, K.A.; Dede, C.; El-Habashy, D.M.; Kamel, S.; Rooney, M.K.; Khamis, Y.; Abdelaal, M.R.; Ahmed, S.; Corrigan, K.L.; Chang, E.; et al. Overview of the Head and Neck Tumor Segmentation for Magnetic Resonance Guided Applications (HNTS-MRG) 2024 Challenge. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*; Springer: Cham, Switzerland, 2024; pp. 1–35.
24. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *European Conference on Computer Vision (ECCV) Workshops*; Springer: Cham, Switzerland, 2022; pp. 205–218.
25. Yu, A.C.; Mohajer, B.; Eng, J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiol. Artif. Intell.* **2022**, *4*, e210064.
26. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211.
27. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
28. Jiang, X.; Osl, M.; Kim, J.; Ohno-Machado, L. Calibrating Predictive Model Estimates to Support Personalized Medicine. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 263–274.
29. Gillmann, C.; Saur, D.; Wischgoll, T.; Scheuermann, G. Uncertainty-Aware Visualization in Medical Imaging—A Survey. *Comput. Graph. Forum* **2021**, *40*, 665–689.
30. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates: Red Hook, NY, USA, 2019; Volume 32, pp. 8024–8035.
31. Jadon, S. A Survey of Loss Functions for Semantic Segmentation. In *Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Via del Mar, Chile, 27–29 October 2020*; IEEE: New York, NY, USA, 2020; pp. 1–7.
32. OpenXR Toolkit. Available online: <http://mbucchia.github.io/OpenXR-Toolkit/> (accessed on 30 September 2023).
33. Cardenas, C.; Mohamed, A.; Sharp, G.; Gooding, M.; Veeraraghavan, H.; Yang, J. *Data from AAPM RT-MAC Grand Challenge 2019*; The Cancer Imaging Archive; National Cancer Institute (NCI): Rockville, MD, USA, 2019. <https://doi.org/10.7937/tcia.2019.bcfjqfqb>.
34. Lewis, J.R.; Utesch, B.S.; Maher, D.E. UMUX-LITE: When There’s No Time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), Paris, France, 27 April–2 May 2013*; ACM: New York, NY, USA, 2013; pp. 2099–2102.
35. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101.
36. Reyes, J.; Batmaz, A.U.; Kersten-Oertel, M. Trusting AI: Does Uncertainty Visualization Affect Decision-Making? *Front. Comput. Sci.* **2025**, *7*, 1464348.
37. Boukhelifa, N.; Johnson, C.R.; Potter, K. Visualization and Decision Making Design Under Uncertainty. *IEEE Comput. Graph. Appl.* **2023**, *43*, 23–25.
38. Ibrahim, M. Confidence-Based Trust Calibration in Human-AI Teams. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 1257.
39. Albarrak, A.M. Improving the Trustworthiness of Interactive Visualization Tools for Healthcare Data Through a Medical Fuzzy Expert System. *Diagnostics* **2023**, *13*, 1733.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.