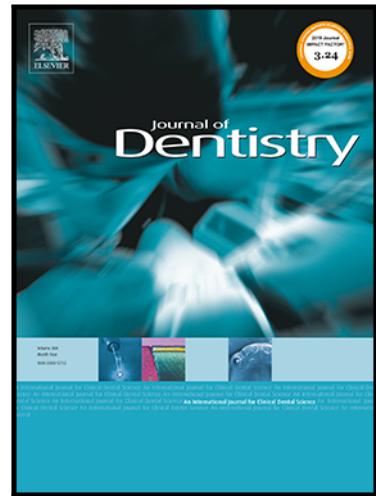


## Journal Pre-proof

Layered Deep Learning for automatic mandibular segmentation in cone-beam computed tomography.

Pieter-Jan Verhelst , Andreas Smolders , Thomas Beznik , Jeroen Meewis , Arne Vandemeulebroucke , Eman Shaheen , Adriaan Van Gerven , Holger Willems , Constantinus Politis , Reinhilde Jacobs

PII: S0300-5712(21)00209-8  
DOI: <https://doi.org/10.1016/j.jdent.2021.103786>  
Reference: JJOD 103786



To appear in: *Journal of Dentistry*

Received date: 21 June 2021  
Revised date: 5 August 2021  
Accepted date: 16 August 2021

Please cite this article as: Pieter-Jan Verhelst , Andreas Smolders , Thomas Beznik , Jeroen Meewis , Arne Vandemeulebroucke , Eman Shaheen , Adriaan Van Gerven , Holger Willems , Constantinus Politis , Reinhilde Jacobs , Layered Deep Learning for automatic mandibular segmentation in cone-beam computed tomography., *Journal of Dentistry* (2021), doi: <https://doi.org/10.1016/j.jdent.2021.103786>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Ltd.

Layered Deep Learning for automatic mandibular segmentation in cone-beam computed tomography.

Deep Learning for creating 3D mandibular models

Pieter-Jan Verhelst<sup>a,b,\*</sup> [pieterjan.verhelst@kuleuven.be](mailto:pieterjan.verhelst@kuleuven.be), Andreas Smolders<sup>c</sup>, Thomas Beznik<sup>c</sup>, Jeroen Meewis<sup>a,b</sup>, Arne Vandemeulebroucke<sup>b</sup>, Eman Shaheen<sup>a,b</sup>, Adriaan Van Gerven<sup>c</sup>, Holger Willems<sup>c</sup>, Constantinus Politis<sup>a,b</sup>, Reinhilde Jacobs<sup>a,b,d</sup>

<sup>a</sup>Department of Oral and Maxillofacial Surgery, University Hospitals Leuven, Kapucijnenvoer 33, BE-3000 Leuven, Belgium

<sup>b</sup>OMFS IMPATH Research Group, Department of Imaging and Pathology, Faculty of Medicine, KU Leuven, Kapucijnenvoer 33, BE-3000 Leuven, Belgium

<sup>c</sup>Relu BV, Kapeldreef 60, BE-3001, Leuven, Belgium

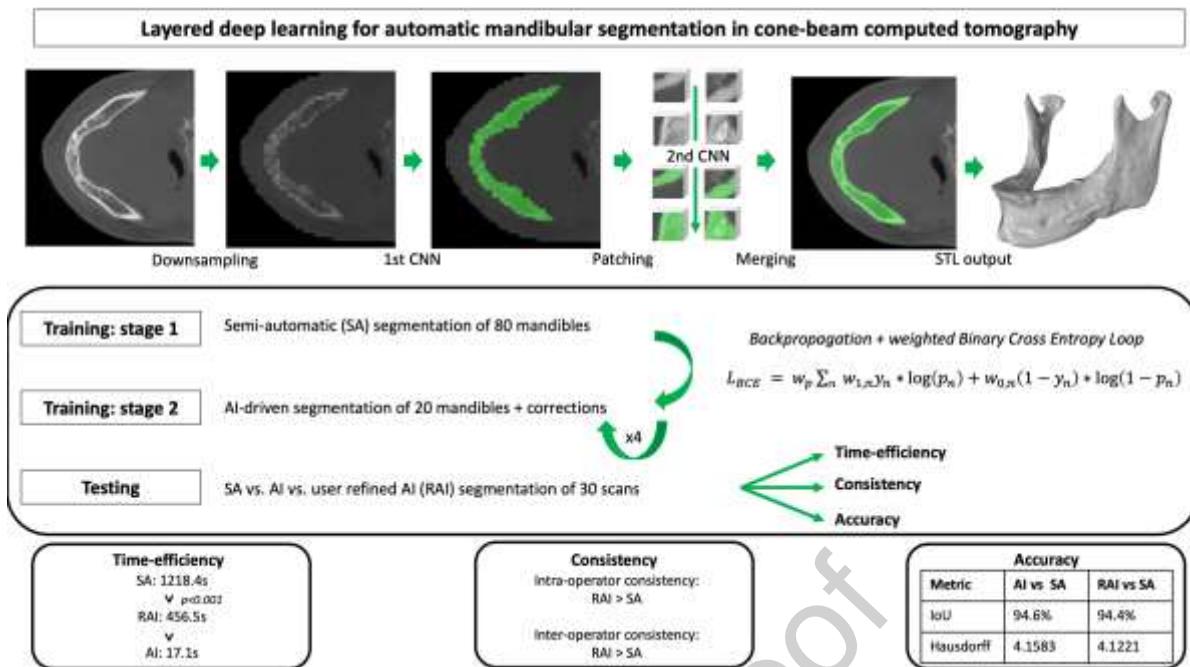
<sup>d</sup>Department of Dental Medicine, Karolinska Institutet, Box 4064, 141 04 Huddinge, Sweden

**\*Corresponding author.** Pieter-Jan Verhelst, MD, DDS, OMFS-IMPATH research group, Department of Imaging & Pathology, KU Leuven, Belgium, Department of Oral & Maxillofacial Surgery, University Hospitals of Leuven, Belgium, Kapucijnenvoer 33, BE-3000 Leuven, Belgium

## Keywords

Cone-beam computed tomography, Computer-generated 3D imaging, Artificial Intelligence, Neural Network Models, Mandible

## Graphical Abstract



## Abstract

### Objective

To develop and validate a layered deep learning algorithm which automatically creates 3D surface models of the human mandible out of CBCT imaging.

### Materials & methods

Two convolutional networks using a 3D U-Net architecture were combined and deployed in a cloud-based artificial intelligence (AI) model. The AI model was trained in two phases and iteratively improved to optimize the segmentation result using 160 anonymized full skull CBCT scans of orthognathic surgery patients (70 preoperative scans and 90 postoperative scans). Finally, the final AI model was tested by assessing timing, consistency, and accuracy on a separate testing dataset of 15 pre- and 15 postoperative full skull CBCT scans. The AI model was compared to user refined AI segmentations (RAI) and to semi-automatic segmentation (SA), which is the current clinical standard. The time needed for segmentation was measured in seconds. Intra- and inter-operator consistency were assessed to check if the segmentation protocols delivered reproducible results. The following consistency metrics were used: intersection over union (IoU), dice similarity coefficient (DSC), Hausdorff distance (HD), absolute volume difference and root mean square (RMS) distance. To

evaluate the match of the AI and RAI results to those of the SA method, their accuracy was measured using IoU, DSC, HD, absolute volume difference and RMS distance.

## Results

On average, SA took 1218,4s. RAI showed a significant drop ( $p<0.0001$ ) in timing to 456,5s (2.7-fold decrease). The AI method only took 17s (71.3-fold decrease). The average intra-operator IoU for RAI was 99,5% compared to 96,9% for SA. For inter-operator consistency, RAI scored an IoU of 99,6% compared to 94,6% for SA. The AI method was always consistent by default. In both the intra- and inter-operator consistency assessments, RAI outperformed SA on all metrics indicative of better consistency. With SA as the ground truth, AI and RAI scored an IoU of 94.6% and 94.4%, respectively. All accuracy metrics were similar for AI and RAI, meaning that both methods produce 3D models that closely match those produced by SA.

## Conclusion

A layered 3D U-Net architecture deep learning algorithm, with and without additional user refinements, improves time-efficiency, reduces operator error, and provides excellent accuracy when benchmarked against the clinical standard.

## Clinical significance

Semi-automatic segmentation in CBCT imaging is time-consuming and allows user-induced errors. Layered convolutional neural networks using a 3D U-Net architecture allow direct segmentation of high-resolution CBCT images. This approach creates 3D mandibular models in a more time-efficient and consistent way. It is accurate when benchmarked to semi-automatic segmentation.

## Keywords

- Cone-beam computed tomography
- Computer-generated 3D imaging
- Artificial Intelligence

- Neural Network Models
- Mandible

## List of abbreviations

CBCT	Cone-Beam Computed Tomography
CMF	Craniomaxillofacial
3D	Three-dimensional
DICOM	Digital Imaging and Communications in Medicine
MSCT	Multi-slice Computed Tomography
SA	Semi-automatic
AI	Artificial Intelligence
DL	Deep Learning
CNN	Convolutional Neural Network
GPU	Graphic Processing Unit
FOV	Field of view
STL	Standard Tessellation Language
RAI	Refined Artificial Intelligence
IoU	Intersection over Union
DSC	Dice Similarity Coefficient
HD	Hausdorff distance
RMS	Root Mean Square

## 1. Introduction

Cone-Beam Computed Tomography (CBCT) is a well-established imaging modality of the head- and neck region [1]. With lower radiation dose, increased spatial resolution, seated position of the patient and lower machine-investment cost, CBCT became a prominent tool in imaging of the craniomaxillofacial (CMF) bones and the dental apparatus [1–3]. It has also initiated an era of virtual treatment planning [4]. This planning relies on three-dimensional (3D) surface models, acquired through segmentation of the exported Digital Imaging and Communications in Medicine (DICOM) data. Such models are imported in virtual treatment planning software suites and surgical treatments are simulated. To transfer the simulation

to the patient, the surgeon can choose to produce surgical guides or patient specific implants through computer-aided manufacturing methods. Surgical guides aid in transferring the surgical plan to the actual patient by indicating an osteotomy site or the correct position and angulation of a drilling sequence. They are used in oncological surgery, dental implant surgery and orthognathic surgery [5]. 3D surface models of CMF structures and the dental apparatus are also used to produce patient specific implants. These patient specific implants have seen a spike in their use as they offer a good fit to the existing anatomy of a patient without the need for extensive sculpting of the acceptor site. Furthermore, they are tailor-made to deliver a specific and predictable result, that is sometimes hard to achieve with stock implants. Patient specific implants have been introduced in osteosynthesis plates[6], temporomandibular joint replacement[6], custom-made meshes for bone regeneration and bone augmentation[7,8], root analogue dental implants [9,10] and subperiosteal implants[11]. These applications require highly accurate surface models. For multi-slice CT (MSCT), this is usually achieved with thresholding and region growing. In CBCT data however, the intrinsic low image contrast, the lack of Hounsfield Units and increased noise and artifacts make this difficult, requiring substantial manual edits [12,13]. This current semi-automatic (SA) clinical standard is characterized by a high processing time and high risk of user induced error.

Artificial intelligence (AI) models in image segmentation, more specifically deep learning (DL) algorithms, promise to overcome the aforementioned caveats [14,15]. AI has been defined by the American National Standard Dictionary of Information Technology as ‘the capability of a device to perform functions normally associated with human intelligence such as reasoning, learning and self-improvement’ [16]. Machine learning is a type of AI technology frequently used in medical image analysis. It allows computers to learn the inherent statistical patterns of pairs of data (f.e. DICOM data) and annotated labels (f.e. Image segmentation). This type of supervised learning allows the computer to eventually make predictions on how a specific anatomical structure should be segmented on new cases[17]. DL is a subtype of machine learning which takes the computer’s autonomy even further by layering its algorithms in artificial neural networks, mimicking the human brain. By this layering, the DL networks become even more powerful[14]. Convolutional Neural Networks (CNN), a type of DL algorithms most commonly applied to analyze images, have

shown promising results in image segmentation in recent years [18]. These networks were already introduced in the 1980s [19,20]. However, CNNs rely on large amounts of data and training requires extensive computational resources, which impeded their practical application. Starting with AlexNet in 2012 [21], the increase in available data and surge in computing power has drawn much attention back to this field. In many computer vision applications, CNNs now outperform more classical approaches. Specifically for biomedical image segmentation, the U-Net architecture has beaten state-of-the-art performance in many applications [22]. This led to the development of many DL algorithms based on U-Net. For CMF CBCT scans, U-Nets have been successfully applied for the segmentation of different structures [23–25]. Current limitations in graphics processing unit (GPU) memory however limit the size of the image, making accurate direct segmentation of high-resolution skull images impossible, which creates difficulties for segmenting large structures, such as the mandible. This work proposes a novel two-step approach, in which one U-Net on a full-size low-resolution image is combined with another U-Net segmenting high-resolution region of interests.

The aim of this study was to develop and validate a layered deep learning algorithm that automatically creates 3D surface models of the human mandible from a CBCT scan. The hypothesis is that such an AI model could provide accurate 3D surface models of the mandible in a more reliable and time efficient way than the current clinical standard, being SA segmentation.

## 2. Materials and methods

### 2.1 Study Design

The study design is illustrated in **figure 1**. As the aim was to develop and validate an AI for automatic mandibular segmentation, this study consisted out of a development, training, and testing phase. This study was performed in accordance with the Artificial intelligence in dental research checklist by Schwendicke et al (Supplementary material S1)[17].

The development of the AI model is described in section 2.3 *Model Architecture*. Training of the model was subdivided into two stages. In the first stage, an initial training set of 30

preoperative scans and 50 postoperative scans was selected, the latter having higher variety due to the variable nature of postoperative bone healing and scattering by the osteosynthesis plates. These 80 DICOM-sets were imported into Mimics Innovation Suite (Version 21.0, Materialise NV, Leuven, Belgium) for SA segmentation. Mandibles in these scans were segmented using a segmentation workflow combining thresholding, region-growing and multi-sliced edits with automatic interpolation, resulting in a standard tessellation language (STL) file. As the goal was to develop an AI model for mandibular segmentation, we chose to omit the crowns of the mandibular teeth. Each scan was segmented by a maxillofacial imaging expert and was checked and corrected by a second expert, who both were calibrated.

The AI model was trained using this initial dataset. This model was afterwards deployed to the cloud-based platform “Virtual Patient Creator” (Relu BV, Leuven, Belgium), where users could upload scans and obtain an initial binary AI driven segmentation result (AI method) of the desired structure. Next, users could correct potential mistakes, delivering a refined AI driven segmentation result (RAI method). Using this workflow, 20 new scans (10 preoperative and 10 postoperative scans) were segmented by the AI, corrected by the two experts, and used to retrain the AI model. The process of adding 20 new scans and refining the segmentation was repeated until the operators applied virtually no refinements anymore, in this case 4 times, leading to a total dataset of 160 scans applied iteratively to train the AI model. Finally, the AI model was tested on a new dataset of 15 pre- and 15 postoperative scans.

## 2.2 Data acquisition

CBCT scans in this study were randomly selected out of a curated database of anonymized orthognathic surgery patients. This database consists out of high-quality full skull CBCT images taken for preoperative planning and postoperative follow-up of orthognathic surgery patients. All CBCT scans were taken with a NewTom VGi evo device (NewTom) with the scanning parameters: field of view (FOV) 24x19cm, voxel size 0.3mm<sup>3</sup>, 110 kV and 4.3mA. Scans were exported in DICOM format. No missing data was present. Ethical approval for

the development and use of the registry was given by the on-site ethical committee (B322201526790)

### 2.3 Model Architecture

In this work, the 3D U-Net architecture was used [22]. However, due to GPU memory limitations, direct segmentation of full resolution (629x816x816) scans was not possible. Therefore, a two-step approach was used (**figure 2**). A first 3D U-Net was used to provide a segmentation of a down-sampled image (96x96x96) of the FOV. Next, the full resolution image was broken down into a series of patches (96x96x96) with an overlap of 20%. Based on the rough segmentation, only patches containing a part of the mandible were retained and other patches not belonging to the mandible were filtered out. Each patch was subsequently segmented by another 3D U-Net. The resulting segmented patches were merged by summing up the logits (output of the network before the final activation) into a single segmentation, which, for voxels in parts with overlapping patches, represented weighted voting based on the confidence each patch had about that specific voxel. This two-step approach also eliminated the problem of class imbalance, as the distribution of mandible and background was balanced across the patches. The image then was binarized, flood-filled and only the largest connected component was retained. Finally, a marching cubes algorithm was applied on the binary image and the resulting mesh was decimated and smoothed to generate a 3D surface model relevant for clinical applications.

Both 3D U-Nets consisted of 4 encoder and 3 decoder blocks, each made up of 2 convolutions followed by ReLU activation and group normalization with 8 feature maps[26]. The number of features after the first encoder is 64 and is doubled in each of the following encoders. All convolutions had a kernel size of 3x3x3, stride 1 and dilation 1. Max pooling was applied after each encoder with kernel size 2x2x2 and stride 2, reducing the resolution with a factor 2 in all dimensions.

### 2.4 Training

Both 3D U-nets were trained using backpropagation and weighted Binary Cross Entropy Loss

$$L_{BCE} = w_p \sum_n w_{1,n} y_n * \log(p_n) + w_{0,n} (1-y_n) * \log(1-p_n)$$

with  $w_p$  = the weight of the patch, n= the voxel,  $w_{c,n}$  = the weight of voxel n and class c (0 or 1),  $y_n$ = the ground truth and  $p_n$ the predicted probability of the network. The weights of the patches are initially 1 and each class obtained the same weight. However, visual inspection by the operators in the improvement step of the data acquisition showed that the region around the mandibular condyle was sometimes inaccurately segmented because of the low contrast and the proximity to the glenoid fossa. Therefore, the weight of voxels in these regions was increased. Both models were implemented in PyTorch and trained using Nvidia RTX 2080 GPUs with 11 GB VRAM. The model parameters were optimized using the Adam optimizer with initial learning rate 1.25e-4 which was halved 7 times during 300 epochs. Data augmentations included random rotations, crops, scaling and elastic deformation, as well as input mixup [27]. A validation set containing 10% of the scans was used for early stopping.

In the improvement stage, each time a relatively small number of scans (20 scans) was added to the existing set (80-140 scans). However, these 20 scans had to teach the model to overcome its initial bias and correct for the errors it used to make. To leverage the new information without losing the advantage of using the complete dataset for training, the weight of the patches ( $w_p$ ) coming from the new scans was increased compared to the old patches. Furthermore, the parameters of the previous version of the AI model were taken as initial parameters for the training, after which the model was trained as described above. To avoid overfitting, we used batch normalization, early stopping as mentioned earlier and data augmentations.

## 2.5 Testing of the AI model

After development and training of the CNN, it was tested on the Virtual Patient Creator platform where it will be used for clinical purposes. Because of the lack of a gold standard, timing, consistency and accuracy of the AI and RAI method for creating 3D surface models of the mandible were compared to the SA method, based on a test set including 30 new scans. To be able to compare the outcomes of the three methods, the cutting plane to remove the teeth of each SA model was used to create a shared cranial plane of the dentoalveolar base

in all 3D surface models of a specific scan. All statistical analysis were performed in SPSS statistics v27.0.

### 2.5.1 Timing

The SA method was timed from when the DICOM data was opened in Mimics until the STL file was produced. For the AI method, the algorithm automatically timed the seconds needed by the AI model to produce the full-resolution binary segmentation result. Finally, for the RAI method, the time required for user refinements and subsequent STL generation was added to the AI method. All three methods were performed by two operators on 10 scans of the test set.

### 2.5.2 Consistency

Inter- and intra-operator consistency were measured for the SA and RAI methods. This was not evaluated for the AI method since it is by default consistent. To check inter-operator consistency, 10 STL files produced by the two operators were compared. To check the intra-operator consistency, one operator performed the SA and RAI method on two time points with a two-week interval. The following metrics were calculated (**table 1**):

- Intersection over Union (IoU)
- Dice similarity coefficient (DSC)
- Hausdorff distance (HD)
- Absolute volume difference between two STL models
- Root mean square (RMS) distance between surfaces of two STL-files

### 2.5.3 Accuracy

To compare the accuracy of the methods, one operator performed all methods on all 30 scans in the test set. The STL files generated by the SA method were set as the ground truth. The STL files of the AI and RAI methods were compared to it using the following metrics (**table 1**):

- IoU
- DSC
- HD

- Precision
- Recall
- Absolute volume difference between two STL files
- RMS distance between surfaces of two STL files

## 2.6 Statistical analysis

For all metrics, descriptive statistics were calculated. A two-way repeated measure ANOVA was performed to test for interaction effects between the operator and the method of segmentation on timing of the procedure. In this last analysis, only the SA and RAI methods were compared since only these had operator input and STL file generation. To check if the inter- and intra-operator consistency were significantly different between the SA and RAI method, a paired-sample t-test for all metrics was performed with a significance level set at  $p=0.05$ .

## 3. Results

### 3.1 Timing

**Table 2** provides an overview of required timing. The SA method took on average 20 minutes and 18 seconds. The AI model required 17 seconds on average to produce a binary segmentation result, a 71.3-fold decrease when compared to the SA method. When the AI model was combined with user refinements and STL generation (RAI method), mean timing increased to 456.16 seconds or 7 minutes and 36 seconds, a 2.7-fold decrease when compared to the SA method. Analysis of the studentized residuals showed that there was normality, as assessed by the Shapiro-Wilk test of normality and there were no outliers, as assessed by no studentized residuals greater than  $\pm 3$  standard deviations. The two-way repeated measures ANOVA showed no statistically significant interaction between method and operator on timing,  $F(1, 9) = 0.474$ ,  $p=0.509$ , allowing the evaluation of the main effect of method (SA vs RAI) independently from operator. This showed a statistically significant difference in timing between methods,  $F(1,9)=214.872$ ,  $p<0.001$ . On average, RAI segmentation is 761,9 seconds (12.7 minutes) faster than SA segmentation (95% CI: 644.6s-879.9s).

### 3.2 Consistency

**Table 3** provides the summarized intra- and inter-operator consistency metrics for SA and RAI methods. The RAI method scored an average IoU of 99,53% and 99,61% for intra-operator and inter-operator respectively. Overall, both methods had excellent intra-operator consistency with the RAI method outperforming the SA method on all metrics, indicative of better consistency. The difference in performance was significant for IoU, DSC, and RMS distance. Examination of the HD and absolute volumetric differences showed that these two metrics had a larger spread in data. Inter-operator consistency metrics were also excellent for both methods and the RAI method significantly outperformed the SA method on all metrics.

### 3.3 Accuracy

**Table 4** shows an overview of accuracy metrics comparing the AI and RAI STL files to those produced by the SA method (ground truth). On average, the IoU of the AI method was 94,6% and the RAI method scored 94,4%. Taking all metrics into account, both AI and RAI methods scored excellent, meaning 3D surface models produced by these methods closely match those produced by the SA method. The difference between AI vs SA and RAI vs SA is minimal. When looking at absolute volume differences (SA – AI or RAI), a bigger volume difference is seen in the RAI vs SA group implying that corrections mostly implied trimming the AI segmentation.

## 4. Discussion

3D models of CMF structures play a crucial role in diagnostics, treatment planning and patient communication. Historically, two main approaches for creating these models have been used: volume rendering and surface rendering [28]. If 3D visualization is the goal, volume rendering is the method of choice due to its higher accuracy when it comes to mixed tissue interfaces and time-efficiency [29]. However, when interaction with the 3D model is required, as is the case in 3D printing or virtual treatment planning, a surface mesh is necessary for which accurate segmentation is essential. The results of this study confirm a positive impact of using an AI model for automated segmentation. This model, with and without additional user refinements, provides an important gain in time-efficiency, reduces operator error, and provides excellent accuracy when it is benchmarked against SA

segmentation. As errors in the AI model are still possible, the Virtual Patient Creator platform allows corrections. If corrections where necessary, they mostly implied slight trimming the AI segmentation. Although this again introduces a risk of lowering consistency, this study showed that this risk is limited as excellent consistency metrics were observed in the RAI method. Regarding accuracy, both the AI as RAI method scored excellent.

No papers were retrieved evaluating accuracy of DL segmentation of mandibular bone from CBCT data. Therefore, we compared our findings to studies focusing on the performance DL segmentation of the mandible from regular CT data. Egger et al. used a fully convolutional network (FCN-8s) for automated mandibular segmentation out of CT data and achieved an average DSC of 89.64% [30]. Yan et al developed a symmetric CNN, enforcing convolution and deconvolution computation to be symmetric to increase segmentation performance. They report a mean DSC of 92.02%[31]. Qiu et al. evaluated the application of CNN's for mandibular segmentation, including teeth, in regular CT data [32]. They reported a mean DSC of 88.1% and a mean RMS distance of 0.5791mm. Tong et al. also used fully convolutional networks with and without a shape representation model and achieved a DSC of 92.07% and 93.6% respectively [33]. One recent study of Wang et al. evaluated multiclass CBCT image segmentation using deep learning in which simultaneous segmentation of teeth and jaw bones (mandible and maxilla) is combined. They reported ad DSC of 93,4% for jaw bones and a total segmentation time (jaw bones + teeth) of 25s[34]. The layered CNN approach in the current study outperforms these findings in a challenging full skull CBCT dataset. When compared to SA segmentation, AI method metrics were almost identical to those of the RAI method. Since the differences are minimal, this indicates that few corrections are needed in the RAI method. The increase in absolute volume difference in the RAI vs SA group suggests that when corrections are needed, they mostly concern trimming the initial AI segmentation. Finally, an increase in time-efficiency was noted. Compared to the SA method, the AI model showed a 71.3-fold decrease in time-consumption. When factoring in inspection of the results, making corrections, and producing the STL, a 2.7-fold decrease in time consumption was seen. Increasing the time-efficiency is one of the main concerns if we want these techniques to become routinely used in clinical practice.

Some limitations were present in the current study. As the main objective was to produce 3D models that will be used for 3D printing and virtual treatment planning, the AI method does not discriminate between cortical and medullary bone. It identifies the mandibular shape and contour and fills up the volume. This means, apart from its orifices, the mandibular canal is currently not identified in the segmentation and is ‘filled out’ as are the other internal structures of the mandible. A next step forward would be to segment the mandibular canal which identifies the course of the inferior alveolar nerve. Also, the current AI algorithm only focused on the mandibular bone and the crowns of the teeth were not included. However, AI models for individual tooth segmentation in CBCT imaging are being developed and tested[25]. There were two main reasons for this approach. First, by limiting the anatomical region of the AI model, manual labeling required for model training could focus on a specific region, possibly reducing errors in the training data. Secondly, by combining different AI models (mandibular canal, mandibular bone and teeth), one can always have a choice in which structures need to be segmented.

Another limitation in this study was the use of a single CBCT-machine and standardized scanning parameters. The data was exported out of a registry designed for orthognathic surgery follow-up for which standardization is great. However, this standardization introduces a risk for bias in training of the AI-model and the model should be strengthened by introducing scans from different CBCT machines with different field of views. Finally, the results of the accuracy assessments should be interpreted with some nuance due to the lack of a real golden standard, which is often the case in medical image analysis AI studies. We regarded the SA method as clinical standard, but on itself, it is also just an approximation of the mandibular anatomy. It is important to interpret the accuracy assessment as to which extent the AI and RAI methods produce 3D models like those derived from the SA method. Furthermore, 80 SA segmentations were initially used as training data for the AI model, carrying a certain risk of biased training data and increased accuracy scores. This was countered by a second training phase in which the training scans were segmented using an initial version of the AI model and allowing the operator to perform manual corrections.

This study demonstrated a layered 3D U-Net architecture AI model which automatically creates a 3D surface model of the human mandible from CBCT images. Our results confirm a positive impact of using an AI model for automated segmentation. The AI model, with and

without additional user refinements, provides an important gain in time-efficiency, reduces operator error, and provides excellent accuracy when it is benchmarked against the current standard semi-automatic segmentation. As the AI model was developed and tested in a clinically oriented cloud-based platform, it has the advantage of being clinician-friendly. The AI model can be accessed and used through a web-browser and does not rely on computational power on the user's side. This will facilitate its use in clinical practice. It's true power lies in the creation of a complete virtual patient by combining multiple AI models that allow segmentation of teeth, airways, soft tissues, the mandibular canal, and other CMF structures. Future research should focus on this multiclass approach. When this multiclass automatic segmentation approach becomes available in clinical practice, it will provide a solid base for further improvements in virtual treatment planning and designing surgical guides and patient specific implants.

### **CrediT authorship contribution statement**

**Pieter-Jan Verhelst:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Project administration. **Andreas Smolders:** Methodology, Software, Validation, Writing – original draft. **Thomas Beznik:** Methodology, Software, Validation, Formal analysis, Visualization, Writing – original draft. **Jeroen Meewis:** Data Curation, Validation, Investigation, Writing – review & editing. **Arne Vandemeulebroucke:** Data curation, Validation, Writing: review & editing. **Eman Shaheen:** Conceptualization, Data curation, Resources, Writing: review & editing. **Adriaan Van Gerven:** Conceptualization, Software, Methodology, Data curation, Writing: review & editing. **Holger Willems:** Conceptualization, Resources, Writing: review & editing, Supervision. **Constantinus Politis:** Conceptualization, Resources, Writing: review & editing, Supervision. **Reinhilde Jacobs:** Conceptualization, Resources, Methodology, Writing: review & editing, Supervision

### **Declaration of Competing Interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **Acknowledgements**

The data used for this study was retrieved from the LORTHOG Register, Department of Oral & Maxillofacial Surgery, University Hospital Leuven, Belgium, B322201526790. The development of the Virtual Patient Creator was partly made possible with the support of a Development Project of VLAIO (Flanders Innovation & Entrepreneurship).

## 5. References

- [1] J.B. Carter, J.D. Stone, R.S. Clark, J.E. Mercer, Applications of Cone-Beam Computed Tomography in Oral and Maxillofacial Surgery: An Overview of Published Indications and Clinical Usage in United States Academic Centers and Oral and Maxillofacial Surgery Practices, *J. Oral Maxillofac. Surg.* 74 (2016) 668–679.  
<https://doi.org/10.1016/j.joms.2015.10.018>.
- [2] A. Katsumata, A. Hirukawa, S. Okumura, M. Naitoh, M. Fujishita, E. Ariji, R.P. Langlais, Effects of image artifacts on gray-value density in limited-volume cone-beam computerized tomography, *Oral Surgery, Oral Med. Oral Pathol. Oral Radiol. Endodontology.* 104 (2007) 829–836. <https://doi.org/10.1016/j.tripleo.2006.12.005>.
- [3] B. Schlueter, K.B. Kim, D. Oliver, G. Sortiropoulos, Cone Beam Computed Tomography 3D Reconstruction of the Mandibular Condyle, *Angle Orthod.* 78 (2008) 880–888.  
<https://doi.org/10.2319/072007-339.1>.
- [4] W. De Vos, J. Casselman, G.R.J. Swennen, Cone-beam computerized tomography (CBCT) imaging of the oral and maxillofacial region: A systematic review of the literature, *Int. J. Oral Maxillofac. Surg.* 38 (2009) 609–625.  
<https://doi.org/10.1016/j.ijom.2009.02.028>.
- [5] D. Khorsandi, A. Fahimipour, P. Abasian, S.S. Saber, M. Seyedi, S. Ghanavati, A. Ahmad, A.A. De Stephanis, F. Taghavinezhaddilami, A. Leonova, R. Mohammadinejad, M. Shabani, B. Mazzolai, V. Mattoli, F.R. Tay, P. Makvandi, 3D and 4D printing in

- dentistry and maxillofacial surgery: Printing techniques, materials, and applications, Acta Biomater. 122 (2021) 26–49. <https://doi.org/10.1016/j.actbio.2020.12.044>.
- [6] M.F. Huang, D. Alfi, J. Alfi, A.T. Huang, The Use of Patient-Specific Implants in Oral and Maxillofacial Surgery, Oral Maxillofac. Surg. Clin. North Am. 31 (2019) 593–600. <https://doi.org/10.1016/j.coms.2019.07.010>.
- [7] A. Cucchi, A. Bianchi, P. Calamai, L. Rinaldi, F. Mangano, E. Vignudelli, G. Corinaldesi, Clinical and volumetric outcomes after vertical ridge augmentation using computer-aided-design/computer-aided manufacturing (CAD/CAM) customized titanium meshes: a pilot study, BMC Oral Health. 20 (2020) 219. <https://doi.org/10.1186/s12903-020-01205-4>.
- [8] F. Luongo, F.G. Mangano, A. Macchi, G. Luongo, C. Mangano, Custom-made synthetic scaffolds for bone reconstruction: A retrospective, multicenter clinical study on 15 patients, Biomed Res. Int. 2016 (2016). <https://doi.org/10.1155/2016/5862586>.
- [9] T. Dantas, S. Madeira, M. Gasik, P. Vaz, F. Silva, Customized Root-Analogue Implants: A Review on Outcomes from Clinical Trials and Case Reports, Materials (Basel). 14 (2021) 2296. <https://doi.org/10.3390/ma14092296>.
- [10] F.G. Mangano, M. De Franco, A. Caprioglio, A. MacChi, A. Piattelli, C. Mangano, Immediate, non-submerged, root-analogue direct laser metal sintering (DLMS) implants: A 1-year prospective study on 15 patients, Lasers Med. Sci. 29 (2014) 1321–1328. <https://doi.org/10.1007/s10103-013-1299-0>.
- [11] C. Mangano, A. Bianchi, F.G. Mangano, J. Dana, M. Colombo, I. Solop, O. Admakin, Custom-made 3D printed subperiosteal titanium implants for the prosthetic restoration of the atrophic posterior mandible of elderly patients: a case series, 3D Print. Med. 6 (2020) 1–14. <https://doi.org/10.1186/s41205-019-0055-x>.

- [12] Y. Fan, R. Beare, H. Matthews, P. Schneider, N. Kilpatrick, J. Clement, P. Claes, A. Penington, C. Adamson, Marker-based watershed transform method for fully automatic mandibular segmentation from cBct images, Dentomaxillofacial Radiol. 48 (2019). <https://doi.org/10.1259/dmfr.20180261>.
- [13] M. Vaitiekūnas, D. Jegelevičius, A. Sakalauskas, S. Grybauskas, Automatic Method for Bone Segmentation in Cone Beam Computed Tomography Data Set, Appl. Sci. 10 (2019) 236. <https://doi.org/10.3390/app10010236>.
- [14] H.-P. Chan, R.K. Samala, L.M. Hadjiiski, C. Zhou, Deep Learning in Medical Image Analysis, in: Adv. Exp. Med. Biol., 2020: pp. 3–21. [https://doi.org/10.1007/978-3-030-33128-3\\_1](https://doi.org/10.1007/978-3-030-33128-3_1).
- [15] G. Wang, W. Li, M.A. Zuluaga, R. Pratt, P.A. Patel, M. Aertsen, T. Doel, A.L. David, J. Deprest, S. Ourselin, T. Vercauteren, Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning, IEEE Trans. Med. Imaging. 37 (2018) 1562–1573. <https://doi.org/10.1109/TMI.2018.2791721>.
- [16] D. Longley, M. Shain, Dictionary of information technology, Macmillan International Higher Education, 1988.
- [17] F. Schwendicke, T. Singh, J.-H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, S. Uribe, J. Krois, Artificial intelligence in dental research: Checklist for authors, reviewers, readers, J. Dent. 107 (2021) 103610. <https://doi.org/10.1016/j.jdent.2021.103610>.
- [18] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image Segmentation Using Deep Learning: A Survey, IEEE Trans. Pattern Anal. Mach. Intell. 8828 (2021) 1–20. <https://doi.org/10.1109/TPAMI.2021.3059968>.
- [19] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36

- (1980) 193–202. <https://doi.org/10.1007/BF00344251>.
- [20] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Comput.* 1 (1989) 541–551. <https://doi.org/10.1162/neco.1989.1.4.541>.
- [21] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM.* 60 (2017) 84–90. <https://doi.org/10.1145/3065386>.
- [22] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 9901 LNCS (2016) 424–432. [https://doi.org/10.1007/978-3-319-46723-8\\_49](https://doi.org/10.1007/978-3-319-46723-8_49).
- [23] Z. Cui, C. Li, W. Wang, ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images, in: 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., IEEE, 2019: pp. 6361–6370. <https://doi.org/10.1109/CVPR.2019.00653>.
- [24] G.H. Kwak, E.-J. Kwak, J.M. Song, H.R. Park, Y.-H. Jung, B.-H. Cho, P. Hui, J.J. Hwang, Automatic mandibular canal detection using a deep convolutional neural network, *Sci. Rep.* 10 (2020) 5711. <https://doi.org/10.1038/s41598-020-62586-8>.
- [25] P. Lahoud, M. EzEldeen, T. Beznik, H. Willems, A. Leite, A. Van Gerven, R. Jacobs, Artificial intelligence for fast and accurate 3D tooth segmentation on CBCT, *J. Endod.* (2021). <https://doi.org/10.1016/j.joen.2020.12.020>.
- [26] Y. Wu, K. He, Group Normalization, *Int. J. Comput. Vis.* 128 (2020) 742–755. <https://doi.org/10.1007/s11263-019-01198-w>.
- [27] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk

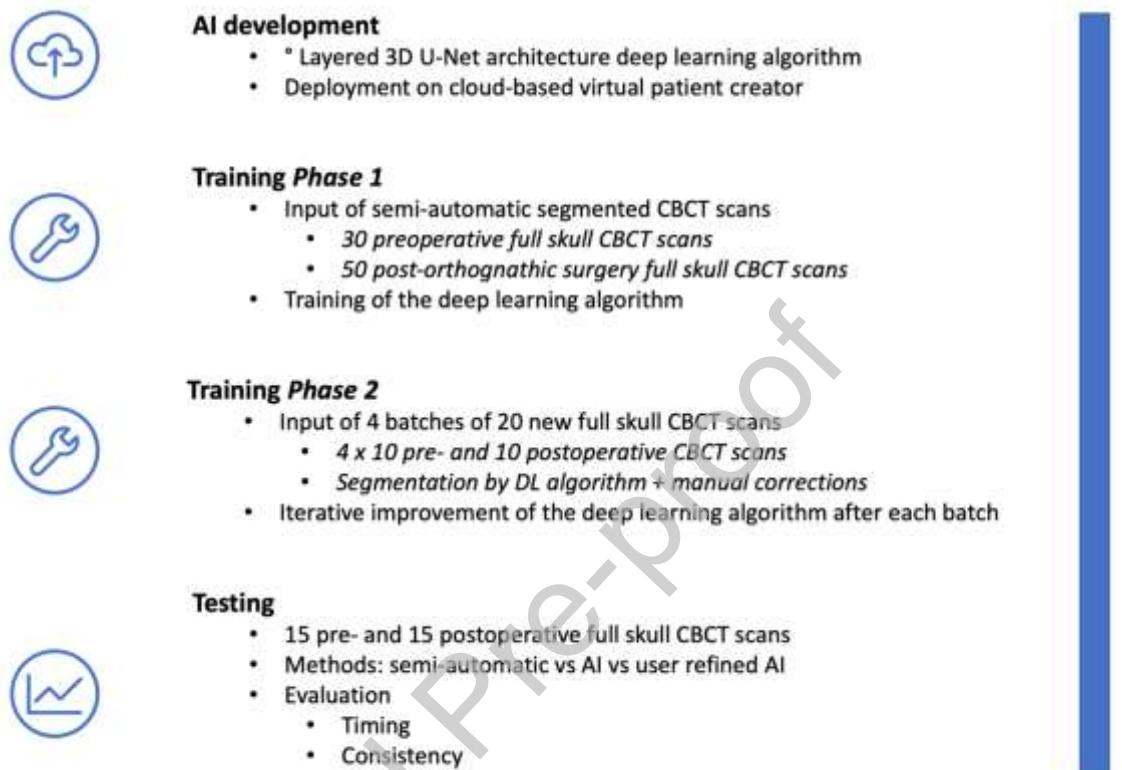
- Minimization, ArXiv. (2017) 1–13. <http://arxiv.org/abs/1710.09412>.
- [28] J.K. Udupa, H.M. Hung, K.S. Chuang, Surface and volume rendering in three-dimensional imaging: A comparison, *J. Digit. Imaging.* 4 (1991) 159–168.  
<https://doi.org/10.1007/BF03168161>.
- [29] E.K. Fishman, D.R. Ney, D.G. Heath, F.M. Corl, K.M. Horton, P.T. Johnson, Volume rendering versus maximum intensity projection in CT angiography: What works best, when, and why, *Radiographics*. 26 (2006) 905–922.  
<https://doi.org/10.1148/rg.263055186>.
- [30] J. Egger, B. Pfarrkirchner, C. Gsaxner, L. Lindner, D. Schmalstieg, J. Wallner, Fully Convolutional Mandible Segmentation on a valid Ground- Truth Dataset, in: 2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., IEEE, 2018: pp. 656–660.  
<https://doi.org/10.1109/EMBC.2018.8512458>.
- [31] M. Yan, J. Guo, W. Tian, Z. Yi, Symmetric convolutional neural network for mandible segmentation, *Knowledge-Based Syst.* 159 (2018) 63–71.  
<https://doi.org/10.1016/j.knosys.2018.06.003>.
- [32] B. Qiu, J. Guo, J. Kraeima, H.H. Glas, R.J.H. Borra, M.J.H. Witjes, P.M.A. van Ooijen, Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network, *Phys. Med. Biol.* 64 (2019) 175020. <https://doi.org/10.1088/1361-6560/ab2c95>.
- [33] N. Tong, S. Gou, S. Yang, D. Ruan, K. Sheng, Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks, *Med. Phys.* 45 (2018) 4558–4567.  
<https://doi.org/10.1002/mp.13147>.
- [34] H. Wang, J. Minnema, K.J. Batenburg, T. Forouzanfar, F.J. Hu, G. Wu, Multiclass CBCT

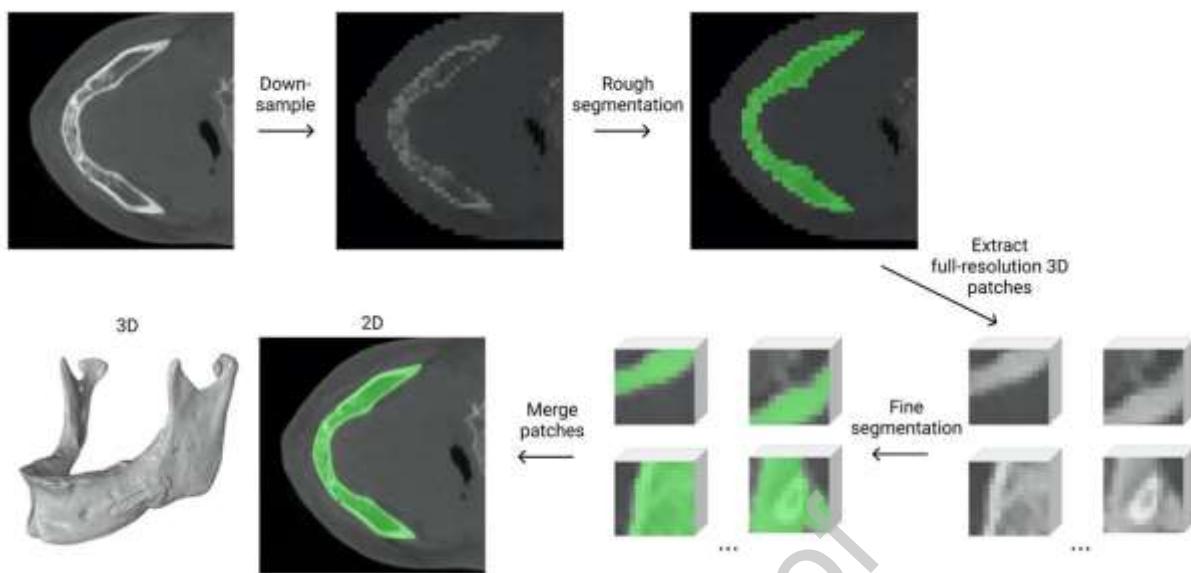
Image Segmentation for Orthodontics with Deep Learning, J. Dent. Res. 100 (2021)  
943–949. <https://doi.org/10.1177/00220345211005338>.

Journal Pre-proof

## Figure and tables

**Figure 1: overview of the study design**



**Figure 2: Illustration of the layered deep learning algorithm**

In the first step, the full scan is downsampled to a fixed size because of GPU memory limitations and given as input to a first CNN to perform a low-resolution segmentation of the mandible. Using this segmentation, 3D patches of the full-resolution scan are extracted. In the second step, the extracted patches are fed to another CNN, and the segmentations are then combined to create the full-resolution segmentation map.

**Table 1:** overview of metrics used for consistency and accuracy assessment

Metric	Formula	Legend
<b>Intersection over union (IoU)</b>	$IoU(A, B) = \frac{ A \cap B }{ A \cup B } = \frac{TP}{TP + FP + FN}$	This ratio represents the number of overlapping voxels between the volume of interest and the ground truth divided by their union. An IoU of 1 is a perfect match.
<b>Dice similarity coefficient (DSC)</b>	$DSC(A, B) = \frac{2  A \cap B }{ A  +  B } = \frac{2 TP}{2 TP + FP + FN}$	This ratio also represents the overlap of voxels between the volume of interest and the ground truth over their union. A DSC of 1 is a perfect match.
<b>Hausdorff distance (HD)</b>	$d_{Hausdorff}(A, B) = \{sup_{a \in A} inf_{b \in B} d(a, b), sup_{b \in B} inf_{a \in A} d(a, b)\}$	This is the maximum difference measured in mm between the ground truth and the volume of interest. A Hausdorff distance of 0 mm is a perfect match.
<b>Absolute volume difference between two STL models (absvoldif)</b>	$ABS(A_{STL1} - B_{STL2})$	This is the absolute volumetric difference, measured in $mm^3$ between two STL-files in which A the volume of STL1 and B the volume STL2.
<b>Root mean square distance between surfaces of two STL-files (RMS)</b>	$RMS\ distance(d) = \sqrt{\frac{1}{n}(d_1^2 + d_2^2 + \dots + d_n^2)}$ $d = distance\ (mm)\ between\ two\ closest\ points\ of\ the\ two\ surfaces$	This is a measure of the imperfection of the fit of a surface of interest to the ground truth in mm. An RMS distance of 0mm is a perfect match.
<b>Precision</b>	$Precision(A, B) = \frac{TP}{TP + FP}$	Fraction of voxels that the algorithm predicted to belong to the class of interest that belonged to that class.
<b>Recall</b>	$Recall(A, B) = \frac{TP}{TP + FN}$	Fraction of voxels that belonged to the class of interest, that the algorithm managed to predict.

**A** = Volumetric data of operator 1, time point 1 or SA method; **B** = Volumetric data of operator 2, time point 2 or AI/RAI method; **TP** = true positives; **TN** = true negatives; **FP** = false positives; **FN** = false negatives

**Table 2: Timing of segmentation methods**

Method	Mean (s)	SD (s)	Min (s)	Max (s)
SA	1218.4 (100%)	187.7	939.9	1556.3
AI	17.1 (37.5%)	1.6	13.5	19.5
RAI	456.5 (1.4%)	158.4	177.5	766.4

*SA: semi-automatic method, AI: artificial intelligence method, RAI: user refined artificial intelligence method. Percentages next to the mean indicate the relative number compared to the timing of the SA method.*

*s: seconds, SD: standard deviation, min: minimal value, max: maximal value*

**Table 3: Consistency assessment of SA and RAI methods****Intra-operator consistency**

	SA		RAI		Paired difference (RAI-SA)	
	Mean +- SD	[min - max]	Mean +- SD	[min - max]	Mean	p-value
IoU	0.9692 ± 0.0054	[0.9580 - 0.9750]	0.9953 ± 0.0027	[0.9909 - 0.9989]	0.0261	<0.00001*
DSC	0.9844 ± 0.0028	[0.9785 - 0.9873]	0.9976 ± 0.0013	[0.9954 - 0.9994]	0.0133	<0.00001*
HD (mm)	3.0934 ± 0.9314	[1.9209 - 4.5497]	2.5118 ± 1.1181	[1.5 - 5.1351]	-0.5817	0.0914
AbsVolDif (mm <sup>3</sup> )	224.6330 ± 204.1721	[36.5006 - 603.17]	100.9557 ± 129.0458	[8.507 - 436.7018]	-123.6772	0.2045
RMS (mm)	0.2003 ± 0.0467	[0.1500 - 0.3063]	0.1176 ± 0.0544	[0.0480 - 0.2464]	-0.0826	0.0015*

**Inter-operator consistency**

	SA		RAI		Paired difference (RAI-SA)	
	Mean +- SD	[min - max]	Mean +- SD	[min - max]	Mean	p-value
IoU	0.9464 +- 0.0121	[0.9279 - 0.9603]	0.9961 +- 0.0021	[0.9929 - 0.9999]	0.0497	<0.001*
DSC	0.9725 +- 0.0064	[0.9626 - 0.9797]	0.9981 +- 0.0011	[0.9964 - 0.9996]	0.0256	<0.001*
HD (mm)	4.1337 +- 0.8209	[3 - 5.4415]	2.1533 +- 0.8173	[0.9 - 3.6125]	-1.9804	<0.001*
AbsVolDif (mm <sup>3</sup> )	1183.288 +- 1044.6551	[17.115 - 3325.9499]	106.3806 +- 89.24	[11.6179 - 309.7144]	-1076.9075	0.011*
RMS (mm)	0.4333 +- 0.1559	[0.2760 - 0.7647]	0.0382 +- 0.1493	[0.0382 - 0.1493]	-0.3485	<0.001*

SA: semi-automatic method, AI: artificial intelligence method, RAI: user refined artificial intelligence method.

SD: standard deviation, min: minimal value, max: maximal value

IoU: Intersection over union, DSC: Dice coefficient score, HD: Hausdorff distance, AbsVolDif: absolute volume difference, RMS: root mean square distance

\* Statistically significant difference between SA and RAI, p&lt;0.05

**Table 4: accuracy assessment of AI and RAI vs SA methods**

Metric		AI vs SA	RAI vs SA
IoU	Mean	0.9459	0.9442
	SD	0.0117	0.0113
	St e mean	0.0021	0.0021
	Median	0.9477	0.9457
	5% percentile	0.9190	0.9178
	95% percentile	0.9621	0.9595
DSC	Mean	0.9722	0.9713
	SD	0.0062	0.0060
	St e Mean	0.0011	0.0011
	Median	0.9731	0.9721
	5% percentile	0.9577	0.9571
	95% percentile	0.9807	0.9793
HD (mm)	Mean	4.1583	4.1221
	SD	2.8549	2.8776
	St e mean	0.5212	0.5254
	Median	3.0075	3.0075
	5% percentile	1.6604	1.6604
	95% Percentile	12.3783	
Precision	Mean	0.9519	0.9498
	SD	0.0104	0.0100
	St e mean	0.0019	0.0018
	Median	0.9530	0.9514
	5% percentile	0.9284	0.9268
	95% percentile	0.9661	0.9642
Recall	Mean	0.9934	0.9937
	SD	0.0031	0.0030
	St e mean	0.0006	0.0006
	Median	0.9936	0.9995
	5% percentile	0.9871	0.9878
	95% percentile	0.9977	0.9978
AbsVolDif (mm <sup>3</sup> )	Mean	2118.8582	2250.1563
	SD	489.5242	468.8839
	St e mean	89.3745	85.6061
	Median	1985.9879	2138.8447
	5% percentile	1389.1034	1497.4535
	95% percentile	3180.3445	3248.3787
RMS (mm)	Mean	0.2634	0.2690
	SD	0.0594	0.0562
	St e mean	0.0108	0.0103
	Median	0.2478	0.2617
	5% percentile	0.1940	0.2003

95% percentile	0.3909	0.3938
----------------	--------	--------

SA: semi-automatic method. AI: artificial intelligence method. RAI: user refined artificial intelligence method.

SD: standard deviation. St e Mean: standard error of the Mean

IoU: Intersection over union. DSC: Dice coefficient score. HD: Hausdorff distance. AbsVolDif: absolute volume difference. RMS: root mean square distance

Journal Pre-proof