Beyond Benchmarks: Towards Robust Artificial Intelligence Bone Segmentation in Socio-Technical Systems

Von der Medizinischen Fakultät
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines Doktors der Zahnmedizin genehmigte
Dissertation

vorgelegt von

Kunpeng Xie

aus Changchun (China)

Berichter: Universitätsprofessor Dr. med. Dr. med. dent. Frank Hölzle
Universitätsprofessor Dr. med. Dipl.-Phys. Daniel Truhn, M.Sc

Tag der mündlichen Prüfung: 29.06.2026

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

This dissertation is based on the following published article:

Xie K, Gruber LJ, Crampen M, et al. Beyond benchmarks: Towards robust artificial intelligence bone segmentation in socio-technical systems. Expert Systems With Applications. 2026;299:130031. doi:10.1016/j.eswa.2025.130031

Dedication

In loving memory of my mother.

What I could not give in time, I will honor through steadfast work and honesty.

List of Publications

Publication of the Dissertation:

Xie K, Gruber LJ, Crampen M, Li Y, Ferreira A, Tappeiner E, Gillot M, Schepers J, Xu J, Pankert T, Beyer M, Shahamiri N, ten Brink R, Dot G, Weschke C, van Nistelrooij N, Verhelst PJ, Guo Y, Xu Z, Bienzeisler J, Rashad A, Flügge T, Cotton R, Vinayahalingam S, Ilesan R, Raith S, Madsen D, Seibold C, Xi T, Bergé S, Nebelung S, Kodym O, Sundqvist O, Thieringer F, Lamecker H, Coppens A, Potrusil T, Kraeima J, Witjes M, Wu G, Chen X, Lambrechts A, Soares Cevidanes LH, Zachow S, Hermans A, Truhn D, Alves V, Egger J, Röhrig R, Hölzle F, Hinrichs-Puladi B. Beyond Benchmarks: Towards Robust Artificial Intelligence Bone Segmentation in Socio-Technical Systems. Expert Systems with Applications. 2025. DOI: 10.1016/j.eswa.2025.130031

Other publications:

Motmaen I, **Xie K***, Schönbrunn L, Berens J, Grunert K, Plum AM, Raufeisen J, Ferreira A, Hermans A, Egger J, Hölzle F. Insights into predicting tooth extraction from panoramic dental images: Artificial intelligence vs. dentists. Clinical Oral Investigations. 2024 Jun 18;28(7):381. ***(Co-first)**

Barajas Ordoñez F, **Xie K**, Ferreira A, Siepmann R, Chargi N, Nebelung S, Truhn D, Bergé S, Bruners P, Egger J, Hölzle F, Wirth M, Kuhl C, Hinrichs-Puladi B. Skeletal Muscle Radiation Attenuation at C3 Predicts Survival in Head and Neck Cancer. Current Oncology. 2025 Oct 10, *32*(10), 587.

Li Y, Luijten G, Gsaxner C, Grunert K, Bader A, Hölzle F, Röhrig R, de la Fuente M, Egger J, **Xie K**, Hinrichs-Puladi B. Usability Study of Augmented Reality Visualization Modalities on Localization Accuracy in the Head and Neck: Randomized Crossover Trial. JMIR Serious Games. 2025 Oct 21. **(In print)**

Raufeisen J, **Xie K**, Hörst F, Braunschweig T, Li J, Kleesiek J, Röhrig R, Egger J, Leibe B, Hölzle F, Hermans A. Cyto R-CNN and CytoNuke Dataset: Towards reliable whole-cell segmentation in bright-field histological images. Computer Methods and Programs in Biomedicine. 2024 Jul 1;252:108215.

Li Y, Drobinsky S, Becker P, **Xie K**, Lipprandt M, Mueller CA, Egger J, Hölzle F, Röhrig R, Radermacher K, de la Fuente M. Accuracy and efficiency of drilling trajectories with augmented reality versus conventional navigation randomized crossover trial. npj Digital Medicine. 2024 Nov 10;7(1):316.

Becker P, Li Y, Drobinsky S, Egger J, **Xie K**, Rashad A, Radermacher K, Röhrig R, de la Fuente M, Hölzle F, Puladi B. Development and validation of collaborative robot-assisted cutting method for iliac crest flap raising: Randomized crossover trial. Scientific Reports. 2025 May 15;15(1):16909.

Moors JJ, Xu Z, **Xie K**, Rashad A, Egger J, Röhrig R, Hölzle F, Puladi B. Full-thickness skin graft versus split-thickness skin graft for radial forearm free flap donor site closure: protocol for a systematic review and meta-analysis. Systematic Reviews. 2024 Feb 26;13(1):74.

Becker AK, Puladi B, **Xie K**, Cassataro A, Götzl R, Hölzle F, Beier JP, Knüchel-Clarke R, Braunschweig T. HER3 (ERBB3) amplification in liposarcoma-a putative new therapeutic target?. World Journal of Surgical Oncology. 2024 May 17;22(1):131.

Expert Systems With Applications 299 (2026) 130031

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Beyond benchmarks: Towards robust artificial intelligence bone segmentation in socio-technical systems

Kunpeng Xie [a,b], Lennart Johannes Gruber [c], Martin Crampen [a,b], Yao Li [a,b], André Ferreira [a,b,d,j], Elias Tappeiner [g], Maxime Gillot [ae,af], Jan Schepers [e], Jiangchang Xu [f], Tobias Pankert [h,a], Michel Beyer [i,ak], Negar Shahamiri [j], Reinier ten Brink [k,ah], Gauthier Dot [l], Charlotte Weschke [m], Niels van Nistelrooij [n,r,ag], Pieter-Jan Verhelst [o,ai], Yan Guo [f,a], Zhibin Xu [a], Jonas Bienzeisler [b], Ashkan Rashad [a], Tabea Flügge [r,ag], Ross Cotton [q], Shankeeth Vinayahalingam [n], Robert Ilesan [aa], Stefan Raith [a,h], Dennis Madsen [s], Constantin Seibold [j], Tong Xi [n], Stefaan Bergé [n], Sven Nebelung [y], Oldřich Kodym [t], Osku Sundqvist [x], Florian M. Thieringer [i,ak], Hans Lamecker [m], Antoine Coppens [p], Thomas Potrusil [w,aj], Joep Kraeima [k,ah], Max Witjes [k], Guomin Wu [v], Xiaojun Chen [f], Adriaan Lambrechts [e], Lucia H Soares Cevidanes [ae], Stefan Zachow [u,r], Alexander Hermans [z,y], Daniel Truhn [y], Victor Alves [d], Jan Egger [j,ab,ac,ad], Rainer Röhrig [b], Frank Hölzle [a], Behrus Hinrichs-Puladi [a,b,*]

[a] Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen 52074 Aachen, Germany
[b] Institute of Medical Informatics, University Hospital RWTH Aachen 52074 Aachen, Germany
[c] Department of Oral and Maxillofacial Surgery, University Medical Center Goettingen 37075 Goettingen, Germany
[d] Center Algoritmi/LASI, University of Minho, 4710-057 Braga, Portugal
[e] Materialise NV, 3001 Leuven, Belgium
[f] School of Mechanical Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China
[g] UMIT TIROL - Private University For Health Sciences and Health Technology, 6060 Hall in Tyrol, Austria
[h] Inzipio GmbH, 52070 Aachen, Germany
[i] Department of Oral and Cranio-Maxillofacial Surgery, University Hospital Basel 4031 Basel, Switzerland
[j] Institute of Artificial Intelligence in Medicine (IKIM), University Hospital Essen (AöR), 45131 Essen, Germany
[k] Department of Oral & Maxillofacial Surgery, University Medical Center Groningen 9713 GZ Groningen, the Netherlands
[l] Université Paris Cité, Université Sorbonne Paris Nord, Inserm UMR1333 Oral Heath, F-92120 Montrouge, France
[m] 1000shapes GmbH, 12247 Berlin, Germany
[n] Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, 6525 GA Nijmegen, the Netherlands
[o] Department of Oral and Maxillofacial Surgery, University Hospitals Leuven 3000 Leuven, Belgium
[p] Relu BV, 3001 Leuven, Belgium
[q] Synopsys Northern Europe Ltd., Exeter EX4 3PL, United Kingdom
[r] Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Department of Oral and Maxillofacial Surgery, 12203 Berlin, Germany
[s] University of Zürich CH-8006 Zürich, Switzerland
[t] TESCAN 3DIM, 61700 Brno, Czech Republic
[u] Department of Visual- and Data-centric Computing, Zuse Institute Berlin (ZIB), 14195 Berlin, Germany
[v] Craniomaxillofacial Plastic and Cosmetic Center, Hospital of Stomatology, Jilin University, 130021 Changchun, China
[w] CADS GmbH, 4320 Perg, Austria
[x] Planmeca Oy, FIN-00880 Helsinki, Finland
[y] Department of Diagnostic and Interventional Radiology, RWTH Aachen University, 52074 Aachen, Germany
[z] Visual Computing Institute, Computer Science and Natural Sciences, RWTH Aachen University, 52074 Aachen, Germany
[aa] Department of Oral and Maxillofacial Surgery, Lucerne Cantonal Hospital, 6000 Lucerne, Switzerland
[ab] Center for Virtual and Extended Reality in Medicine (ZvRM), Essen University Hospital (AöR), 45147 Essen, Germany

* Corresponding author at: Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, 52074 Aachen, Germany.
E-mail address: bpuladi@ukaachen.de (B. Hinrichs-Puladi).

https://doi.org/10.1016/j.eswa.2025.130031
Received 8 June 2025; Received in revised form 3 October 2025; Accepted 12 October 2025
Available online 17 October 2025
0957-4174/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

[ac] Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), 45147 Essen, Germany
[ad] University of Duisburg-Essen, Faculty of Computer Science, 45127 Essen, Germany
[ae] University of Michigan, Ann Arbor, 48109-1079 Michigan, United States
[af] CPE Lyon, 69100 Villeurbanne, Lyon, France
[ag] Einstein Center Digital Future (ECDF), 10117 Berlin, Germany
[ah] 3D Lab University Medical Center Groningen, University of Groningen 9713 GZ Groningen, the Netherlands
[ai] OMFS-IMPATH Research Group, Department of Imaging and Pathology, Faculty of Medicine, KU Leuven 3000 Leuven, Belgium
[aj] KLS Martin Group, 78532 Tuttlingen, Germany
[ak] Medical Additive Manufacturing Research Group (Swiss MAM), Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland

ARTICLE INFO

Keywords:
Artificial Intelligence
Image Processing, Computer-Assisted
Imaging, Three-Dimensional
Tomography, X-Ray Computed
Practice Guideline
Mandible Segmentation

ABSTRACT

Despite the advances in automated medical image segmentation, AI models still underperform in various clinical settings, posing challenges for integration into real-world workflows. In this pre-registered prospective multicenter evaluation, we analyzed 20 state-of-the-art mandibular segmentation models across 19,218 segmentations of 1,000 clinically resampled CT/CBCT scans. Our results suggest that for a given model, segmentation accuracy can vary by up to 25% in Dice score as socio-technical factors such as voxel size, bone orientation, and patient conditions (e.g., osteosynthesis or pathology) shift from favorable to adverse. Higher sharpness, isotropic smaller voxels, and neutral orientation significantly improved results, while metallic osteosynthesis and anatomical complexity led to significant degradation. Our findings challenge the common view of AI models as "plug-and-play" tools and suggest evidence-based optimization recommendations for both clinicians and developers. This will in turn boost the integration of AI segmentation tools in routine healthcare.

1. Introduction

With the ongoing digital transformation of healthcare, segmentation-based acquisition of anatomical and pathological structures has become an essential step in both clinical practice and research. Applications scenarios span over a wide field including diagnostic, image-guided radiotherapy and virtual surgical planning (Al-Hassani, 2025; Apostolakis et al., 2022; Fu et al., 2021; Van Baar et al., 2018). However, manual segmentation is still labor intensive and time-consuming. To address this issue, a large number of automatic segmentation methods for different structures have emerged in the last decades, and among them artificial intelligence (AI) models utilizing deep learning methods are the most promising ones (Liu et al., 2023; Qiu et al., 2021); Van Eijnatten et al., 2018). In the segmentation of mandible for example, AI models have progressed beyond research settings and have begun to translate to clinical use as certified medical software in practice (Ileșan et al., 2023; Pankert et al., 2023; Verhelst et al., 2021; Yang and Su, 2021; Zachow, 2015). However, despite decades of algorithmic advancements, there remains no standardized clinical integration protocol for AI segmentation models, leaving clinical integration a major challenge (Antonelli et al., 2022; Qiu et al., 2021; Rajamani et al., 2023).

This may be due to the technocentric paradigm that has been in place for decades of comparing and developing algorithms in different challenges to push the limits of performance and ultimately surpass human capabilities (Maier-Hein et al., 2018). While this technocentric perspective has brought us powerful models and refreshed leaderboards, it often overlooks the complex socio-technical systems in which AI models are applied. In terms of clinicians, recent work shows that their adoption of AI generated results hinge on transparency, robustness, and real-world applicability rather than benchmark metrics alone (Benlalia et al., 2025; Reddy, 2022; Shin, 2024; Van de Sande et al., 2024). Additionally, in real-world situations, medical imaging data is often acquired prospectively based on specific clinical needs, including a wide range of possible imaging protocols as well as different patient factors. In this respect, a shift in perspective from a techno-centric preoccupation to a socio-technical perspective (Whetton and Georgiou, 2010), which explicitly considers clinical contexts such as diverse imaging protocols, patient demographics, and practical workflow integration, would be highly beneficial in facilitating the effective translation of AI segmentation models into clinical routines and research settings (Al-Mistarehi et al., 2023).

Consequently, we need to understand how socio-technical factors affect the performance of AI segmentation models in general. A previous study found that factors such as the imaging modalities, scanning devices, and the reconstruction protocols may impact segmentation outcomes (Gruber et al., 2024). While some studies have begun to explore these factors, previous studies have either focused on limited factors or used only a single AI model, leaving a comprehensive understanding of these interactions largely unveiled (Huang et al., 2021; Minnema et al., 2019).

To address this issue, instead of simply comparing models' performance, we evaluated the impact of socio-technical factors on the overall performance of multiple AI models in this study. For this purpose, we chose the mandible, which is morphologically complex and a representative in bone segmentation, as the segmentation target and created a benchmark dataset that balanced both patient and imaging features. Notably, our study recruited the largest number of AI models for mandible segmentation evaluated to date, including systems developed by multiple independent research groups and companies. This sourcing of models ensures that our evaluation captures a broad spectrum of network designs, training data, and intended use cases. By systematically resampling the original data, we could experimentally control the impact of different factors as they would be controllable during medical image acquisition. We then evaluated the segmentation results to explore the general impact of imaging, patient, and anatomical region factors on the model performance. Based on the results, we further suggest best practice recommendations for clinicians in applying AI segmentation models. In addition, we put forward requirements for AI developers, who are expected to create next-generation models that are informed by the clinical challenges encountered in AI models. Our study provides a reliable evidence base for future clinical integration guidelines of AI segmentation models, helping bridge the gap between technical performance and practical deployment.

2. Methods

In this multicenter study we evaluated state-of-the-art AI models from 20 different centers and companies around the world (Table 1). The study protocol was registered prospectively in the German clinical trial registry (DRKS) under registration ID DRKS00032736. All technical details can be found in this study protocol. The ethics application of the study was approved by the ethics committee at University Hospital RWTH Aachen (EK 23–312). No informed consent was needed due to the use of anonymized retrospective patient data.

Table 1
AI models Summary. *Basic infomation of the recruited AI models.*

Name	Institute/ Company	Location	Architecture
AC-Seg (Pankert et al., 2023)	Inzipio GmbH	Aachen, Germany	3D-UNet (Çiçek et al., 2016)
SegCBCT	University Zurich	Zurich, Switzerland	3D-UNet
SKUBA	CADS	Perg, Austria	nnUNet (Isensee et al., 2021)
MandibleSegNet	Charité University Medicine/ZIB	Berlin, Germany	nnUNet
3D-JMax (Ileşan et al., 2023)	University Hospital Basel	Basel, Switzerland	3D-UNet
Mandible	1000shapes	Berlin. Germany	nnUNet
JawFracNet (Van Nistelrooij et al., 2024)	Radboudumc	Nijmegen, The Netherlands	3D-UNet
JLU-Mandible	Jilin University	Changchun, China	U-Mamba (Ma et al., 2024)
Relu Creator (Verhelst et al., 2021)	Relu BV	Leuven, Belgium	3D-UNet
MandiSeg-Swin	IKIM Essen	Essen, Germany	SwinUNETR (Hatamizadeh et al., 2022)
DentalSegmentator (Dot et al., 2024)	Arts et Métiers Institute of Technology	Paris, France	nnUNet
nnHaN-Net (Tappeiner et al., 2022)	UMIT TIROL	Tirol, Austria	nnUNet
Planmeca Romexis Smart Lite	Planmeca	Helsinki, Finland	DynUNet (Ranzini et al., 2021)
FastJaw	TESCAN	Czech Republic	Cascaded U-nets
Simpleware CMF	Synopsys	California, USA	Confidential
Materialise CMF segmentation model	Materialise NV	Leuven, Belgium	Confidential
KAAK	UMCG	Groningen, The Netherlands	nnUNet
Edge Supervison Segmentation (Xu et al., 2021)	SJTU	Shanghai, China	3D-VNet (Milletari et al., 2016)
AMASSS-CBCT (Gillot et al., 2022)	University of Michigan (Public)	Michigan, USA	3D-UNETR
MedLSAM (Lei et al., 2025)	OpenMedLab (Public)	Shanghai, China	MedSAM (Ma et al., 2024) MedLAM

2.1. Dataset preparation

To build a balanced benchmark dataset with respect to patient-related features, we applied a targeted stratified selection strategy. From a single center we selected 50 computed tomography (CT) and 50 cone-beam CT (CBCT) scans from 100 patients. Cases were selected within feature strata like artifacts, bone pathology, osteosynthesis and dentition, to meet predefined quotas (Supplementary Table S1). In terms of patient characteristics, the sex ratio is adjusted to 1:1 and the average age of patients was 48.47 years (range 19 – 91 years). All selected scans were de-identified by cropping out the region above the inferior border of the orbital rim. Cases were excluded if cropping was not possible without affecting the condyle region. We systematically resampled the original 100 selected cases to create an additional 900 volumes, for a total of 1,000 volumes. This method, instead of selecting 1,000 cases directly, gave us full control over the voxel size, slice thickness, sharpness, noise and rotation of the mandible.

To obtain a balanced dataset, the features of the original CT/CBCT volumes were profiled prior to resampling. These features were quantified and measured in five aspects: a) voxel size (XY); b) slice thickness; c) sharpness; d) noise; e) rotation of the mandible. Voxel size a) and slice thickness b) were read directly from the DICOM tags. Sharpness c) was measured by resampling each scan to 1 mm^3 voxels, applying a Sobel operator within tissue mask ($-120 \sim 3000$ HU), and averaging the absolute gradient values. Noise d) was measured by subtracting a median-filtered copy of the same resampled volume from the original and taking the standard deviation of the residual within the same tissue mask. Mandible rotations e) were calculated using bone landmarks. Detailed measurement methods and formulas are provided in Supplementary Methods. Based on these measurements, we chose five types of resampling methods namely: a) increase the slice thickness; b) expand the voxel size (XY); c) sharpening / smoothing; d) Gaussian-noise / denoise; e) rotation in axial, coronal and sagittal plane. A set of factors were tested and used in resampling these features respectively (Supplementary Table S4). By adjusting these factors, we have managed to approximate the distributions of features on the resampled dataset to the reference distribution from public datasets (Aerts et al., 2014; Cipriano et al., 2022; Cptac, 2018; Grossberg et al., 2020; Vallières et al., 2017a, 2017b; Wee and Dekker, 2019) or normal distribution. Factor settings are listed in Supplementary Methods. A total of 3,727,360 resampling combinations of imaging features were generated, from which 900 were randomly selected and resampled volumes were generated accordingly. These down-sampled volumes, together with the initial 100 scans, resulted in a balanced final dataset of 1,000 volumes. The final distribution of patient and imaging features can be found in Supplementary Fig. S1.

2.2. Ground truths

Mandible segmentations of the original scans were performed by two surgeons experienced in segmentation (K.X. and L.G.) independently in different software, K.X. in Mimics (Version 21.0) and L.G. in 3D Slicer (Version 5.6.2). The quality of segmentations was checked and approved by a third surgeon (B.HP.). The principle of the segmentation was to preserve the anatomical bone structure of the mandible. In this case, all teeth, including dental implants, crowns and bridges, were segmented along with the mandible. Osteosynthesis materials (e.g. reconstruction plates, fixation screws/plates) were excluded in the segmentation, except for the part inside the mandible. The cancellous bone and the mandibular nerve canal were filled in so that the final segmentation result is free of internal cavities. Since resampling is not changing the anatomy of the bone, we applied the same resampling protocol in voxel scaling and rotation to the original ground truths to obtain corresponding segmentation results for the resampled 900 cases.

2.3. Model Recruitment

The segmentation models included in this study need to meet the following criteria: a) deep learning based fully automatic segmentation tool; b) developed within the last five years; c) the output of the model is the mesh model or label map of the whole mandible; d) already trained and ready to use. Based on the literature study of a systematic review, we listed a group of models available in publications and searched further in online databases for other models published after the systematic review (Qiu et al., 2021). We contacted 35 corresponding authors and ten of them agreed to participate in the study. In addition, ten companies that offer mandible segmentation tools as a service were contacted. Eight of them joined our study. Furthermore, we have searched public repositories for available models and applied two trained models. With a data transfer agreement (DTA), the final dataset was shared with the collaborators, and segmentation results were returned to RWTH Aachen for evaluation. If a DTA was not feasible or the model was publicly available, inference was conducted locally at RWTH Aachen University.

2.4. Evaluation

To further evaluate the segmentation quality in different anatomical regions of the mandible, we delineated nine ROIs by K.X and controlled by B.HP.: condyle L/R, inferior alveolar nerve (IAN) entrance L/R, IAN exits L/R, dentition, inferior border. The last ROI, mandible body, was defined as the rest of mandible excluding the ROIs. All of the above ROIs were created based on reference points manually labelled on the volume by K.X. Segmentation results were aligned to the corresponding volume and compared to both manual ground truths. The mean value of the two evaluations was taken as the final result. We chose four metrics for evaluation: Dice Similarity Coefficient (DSC, i.e. overlap measurement), Normalized Surface Dice (NSD, i.e. boundary agreement), 95 percentile Hausdorff Distance (HD95, i.e. worst-case boundary error), and Mean Average Surface Distance (MASD, i.e. average boundary deviation). All metrics were calculated on the whole mandible and on all ROIs respectively using the python package from Nikolov et al. (Nikolov et al., 2021; Google DeepMind, 2022). No evaluations in the dentition region were conducted if the AI model cannot segment the teeth. All evaluations were conducted anonymously to secure the interests of all researchers and companies. It is important to note that participating companies and institutions only provided model information and conducted inference on the benchmark dataset, without involvement in ground truth annotation, data analysis or evaluation results.

2.5. Statistical analysis

The statistical analysis was conducted with the R programming language (Version 4.4.2). For descriptive statistics on data with non-normal distribution, we applied the non-parametric Mann-Whitney *U* test to evaluate statistical significance, followed by a bootstrap procedure with 5,000 replicates to obtain the 95 % CI for the median difference. Factors listed in the above section were set as fixed effect in the Linear Mixed-effect Models (LMMs) while the difference of the AI models was considered as random effect. We checked the collinearity of selected fixed effects and found that sharpness and noise were highly corelated with a variable inflation factor of 10.943. In this case, noise was removed from the list of factors. We scaled the factors and tested multiple combinations of settings and selected one optimal LMM for each ROI and the whole mandible on each metric. LMM results on DSC are displayed in Fig. 7. LMMs on other metrics and details of fitted models are described in Supplementary Figs. S2, S3 and S4. From models described above we establish the evidence base for our recommendations.

3. Results

3.1. Recruited AI models and overall segmentation results

A total of 20 commercial and research AI models for mandible segmentation from different countries across the world were recruited in this study, with the workflow shown in Fig. 1. All models were developed over the last 5 years and listed in Table 1. The evaluation was performed on ground truths of two investigators with an interrater correlation of 95.7 % in DSC. From the 1,000 volumes to be segmented, on average 942 volumes were successfully segmented and 19,218 segmentations were evaluated. The model designations are listed in descending order according to the number of volumes with DSC greater than 90 % in their segmentation results (Fig. 2a). Only one model (S) was unable to segment any CBCT volume.

Table 2 presents the overall performance of the models on our benchmark dataset, including the CT and CBCT subsets. The metrics used were: DSC as primary metric, NSD, HD95, and MASD (Maier-Hein et al., 2024). The mean values of DSC and NSD for all models are both 81.7 %. While the mean values of HD95 and MASD are 14.89 mm and 2.73 mm, respectively. Model *A* demonstrates the best performance across almost all metrics. We explored the effect of the type of training data on the segmentation results (Figs. 2b, c). It is interesting to note that the models trained with only CBCT data show better results than the models trained with only CT data (Mann-Whitney *U* test, $p < 0.001$), and the median difference was estimated as 5.10 % with a 95 % bootstrap confidence interval (CI) of [4.71 %, 5.51 %]. Yet the difference is not significant between CBCT and combination of both data modalities (Mann-Whitney *U* test, $p = 0.733$). Commercial models demonstrate

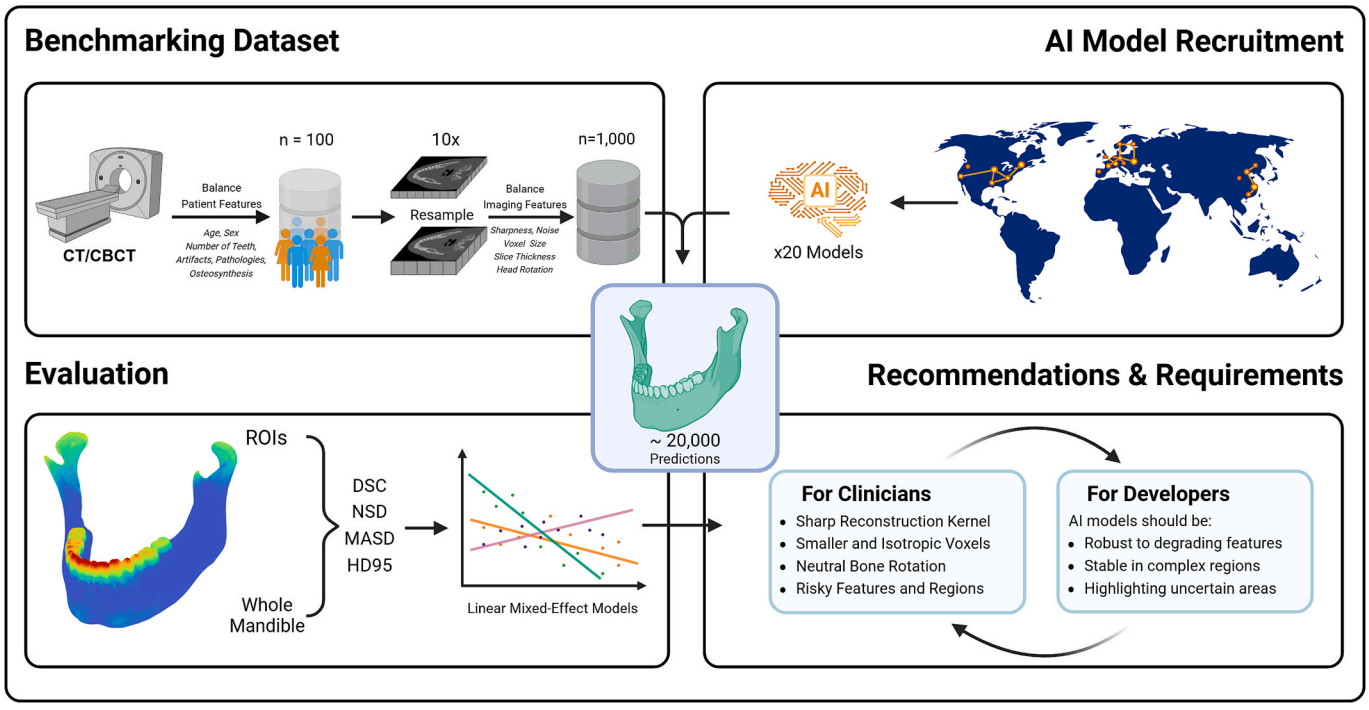


Fig. 1. Workflow of the study. Created in BioRender. Xie, K. (2025) https://BioRender.com/4vhvewq.

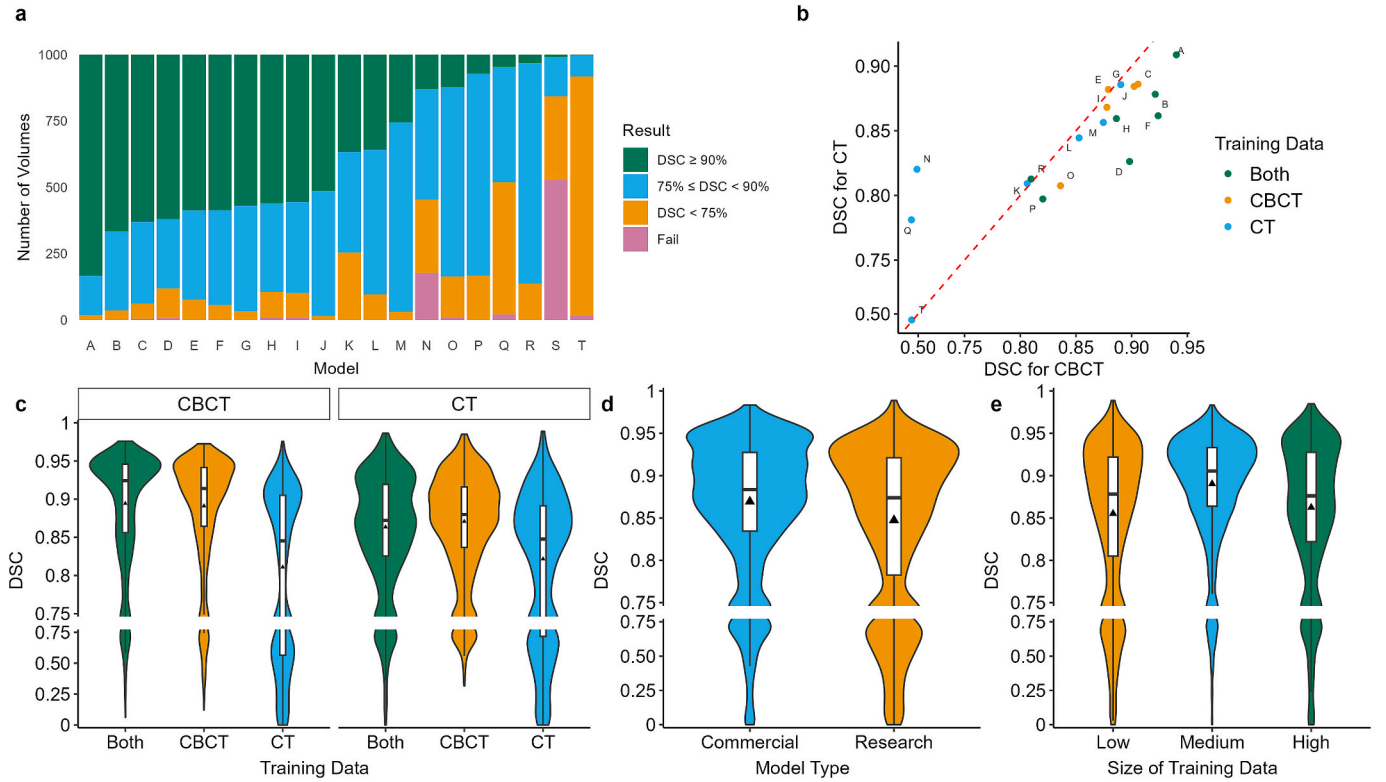


Fig. 2. Model related factors and segmentation performance (**a**) Ranking of models based on segmentation quality. Decrease by number of good cases (DSC ≥ 0.9) (**b**) Distribution of model performance in CT and CBCT subsets based on mean DSC (**c**) Impact of training data on overall segmentation performance (**d**) Impact of model type (**e**) Impact of the size of training dataset. Low: 0–150 cases; Medium: 150–300 cases; High: 300 + cases.

better performance compared to research models (Mann-Whitney *U* test, $p < 0.001$), with a median difference of 1.03 % [95 % CI: 0.75 %, 1.34 %]. Regarding the amount of training data, the models trained on a moderate number of scans (150–300 cases) exhibited the optimal segmentation performance among all groups (Kruskal-Wallis test, $p < 0.001$; Figs. 2d,e). The median DSC difference between the medium and low groups was 2.70 % [95 % CI: 2.39 %, 2.97 %], and between the medium and high groups was 2.87 % [95 % CI: 2.55 %, 3.16 %]. We further explored the model performance on the 100 raw volumes and the resampled subset (900 volumes) of our dataset and summarized it in Supplementary Table S5. Compared to the results of the entire benchmark dataset and its resampled subset, all metrics show an overall improvement on the raw volumes. There are 16 out of 20 models performing better on the raw volumes than resampled ones.

3.2. Imaging factors

Fig. 3 shows the effect of imaging factors on segmentation performance of AI models. Higher sharpness level generally leads to better segmentation results (Fig. 3c). Further analysis in LMMs shows a 0.50 % increase in DSC per 500 Hounsfield Unit (HU)/mm increase in sharpness (LMM, $\beta = 0.001$ %, $p < 0.001$). However, the DSC improvements reached a plateau beyond a certain sharpness level (approximately 5000 HU/mm). This pattern was also observed regarding noise, where a moderate noise level led to the best segmentation performance. Larger voxel sizes in the XY plane significantly reduced segmentation performance, with a 0.16 % decrease in DSC for every 0.1 mm increase in in-plane voxel size (LMM, $\beta = -1.62$ %, $p < 0.001$). Increasing slice thickness also had a negative impact, with DSC declining by 0.10 % for every 0.1 mm increase in slice thickness (LMM, $\beta = -0.955$ %, $p < 0.001$). Rotation of the mandible in all three planes resulted in negative effects on segmentation performance, with axial and sagittal rotations reducing DSC by 0.51 % and 0.69 % per 5-degree increase, respectively (LMM, $\beta_{axial} = -0.102$ %, $\beta_{sagittal} = -0.138$ %, $p < 0.001$), while coronal rotation had no significance ($p = 0.520$).

In univariable descriptive statistics, the AI models showed better performance on CBCT data than that of CT data (Mann-Whitney *U* test, $p < 0.001$; Fig. 3a), with a median DSC difference of 3.20 % [95 % CI: 2.96 %, 3.45 %]. For the use of different CBCT devices, no significant difference was found (Mann-Whitney *U* test, $p = 0.198$; Fig. 3b). Yet a marginal decline of 1.43 % in median DSC [95 % CI: 1.02 %, 1.78 %] is found in CT device C (Mann-Whitney *U* test, $p < 0.001$). However, in multivariable analysis the segmentation performance of the AI model on CT data is improved by 4.13 % compared to CBCT data, (LMM, $\beta = 4.129$ %, $p < 0.001$).

3.3. Patient-related factors

Fig. 4 displays the relationship between patient-related factors and segmentation performance. Male patients showed slightly better segmentation results than female patients, with a 1.0 % higher DSC for males (LMM, $\beta = 0.989$ %, $p < 0.001$). Older patients showed a decrease in DSC, but this effect was not significant (LMM, $\beta = -0.011$ %, $p = 0.126$). We used the mean value of HU across the mandibular region to assess bone density and found that lower bone density reduced segmentation performance (Fig. 4c). The number of teeth in lower dentition positively influenced segmentation performance, with each additional tooth increasing DSC by 0.38 % (LMM, $\beta = 0.378$ %, $p < 0.001$). On the other hand, the presence of bone pathology (e.g. fractures, major cysts) reduced DSC by 0.71 % (LMM, $\beta = -0.708$ %, $p < 0.05$). Osteosynthesis material had the most significant negative effect, decreasing DSC by 7.90 % (LMM, $\beta = -7.90$ %, $p < 0.001$). Artifacts (e.g. metal, shadow) also negatively impacted segmentation, but showed no significant effect on DSC (LMM, $\beta = -0.212$ %, $p = 0.331$).

Table 2
Performance Summary. Segmentation performance (mean ± sd) of AI models on the whole mandible on our benchmark dataset. Best performances were marked in bold. Model S failed to segment CBCT volumes. Models anonymized by descending order of number of segmentations with DSC > 90 %.

Model	DSC (%)			NSD (%)			HD 95 (mm)			MASD (mm)		
	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall
A	**90.87± 9.19**	**94.01± 4.86**	**92.44± 7.51**	**93.77± 9.80**	**94.37± 5.55**	**94.07± 7.97**	4.05± 20.70	**2.26± 5.05**	**3.16± 15.09**	**0.60± 3.02**	**0.29± 0.46**	**0.44± 2.16**
B	87.82± 7.03	92.11± 5.48	89.97± 6.66	89.56± 9.84	91.93± 7.02	90.75± 8.62	4.71± 11.64	3.90± 10.74	4.31± 11.20	0.64± 1.17	0.60± 1.78	0.62± 1.50
C	88.60± 7.22	90.57± 8.87	89.59± 8.15	91.34± 9.60	91.90± 9.60	91.62± 9.60	7.29± 28.00	5.08± 12.31	6.18± 21.64	1.71± 11.48	0.72± 1.99	1.21± 8.25
D	82.62± 18.74	89.81± 11.81	86.24± 16.03	83.40± 20.21	89.50± 12.03	86.48± 16.87	19.90± 56.52	4.68± 9.96	12.23± 41.12	3.15± 11.08	0.70± 1.76	1.91± 7.99
E	88.19± 6.92	87.91± 12.35	88.05± 10.01	89.18± 9.48	85.71± 13.66	87.44± 11.88	10.23± 33.46	11.32± 23.69	10.77± 28.98	1.45± 5.52	1.77± 4.24	1.61± 4.92
F	86.16± 8.05	92.37± 6.80	89.27± 8.07	88.60± 10.81	93.62± 8.52	91.11± 10.04	5.98± 13.37	3.08± 9.73	4.53± 11.78	0.78± 1.82	0.48± 1.51	0.63± 1.68
G	88.57± 4.81	89.02± 10.82	88.79± 8.37	90.04± 7.00	87.57± 10.76	88.80± 9.16	6.70± 14.40	8.13± 11.86	7.41± 13.20	0.62± 0.95	0.95± 2.41	0.78± 1.84
H	85.94± 12.05	88.63± 9.26	87.29± 10.81	87.80± 13.99	87.76± 10.85	87.78± 12.50	12.65± 29.67	9.47± 17.75	11.06± 24.49	2.81± 11.15	1.15± 2.27	1.98± 8.09
I	86.81± 8.61	87.77± 13.56	87.29± 11.35	88.66± 12.07	88.65± 15.19	88.65± 13.70	4.68± 11.40	9.26± 18.64	6.96± 15.59	0.61± 0.91	1.32± 3.11	0.96± 2.31
J	88.43± 3.98	90.22± 4.61	89.32± 4.39	90.45± 6.78	89.95± 6.30	90.20± 6.55	6.92± 22.98	5.92± 7.42	6.42± 17.07	0.74± 2.20	0.61± 1.04	0.68± 1.72
K	80.92± 18.31	80.63± 13.45	80.78± 16.06	83.16± 18.25	80.64± 12.93	81.90± 15.86	16.74± 26.51	10.19± 9.70	13.47± 20.22	2.15± 4.49	1.33± 1.35	1.74± 3.34
L	84.45± 10.07	85.28± 13.85	84.86± 12.11	86.58± 12.82	83.54± 13.65	85.06± 13.32	9.17± 21.49	7.70± 13.45	8.44± 17.93	1.19± 2.62	1.20± 2.34	1.20± 2.48
M	85.64± 4.10	87.45± 7.56	86.55± 6.15	87.12± 10.02	86.13± 9.80	86.63± 9.92	**3.87± 9.59**	5.18± 12.65	4.52± 11.24	0.61± 0.69	1.02± 3.22	0.82± 2.34
N	82.02± 15.89	49.38± 34.36	69.13± 29.55	83.12± 16.03	47.03± 32.13	68.87± 29.56	17.48± 61.10	52.16± 37.07	33.29± 54.35	2.90± 13.85	16.60± 16.09	9.14± 16.40
O	80.74± 8.61	83.61± 9.50	82.17± 9.17	80.76± 12.62	82.72± 10.81	81.74± 11.79	37.07± 47.93	10.63± 16.60	23.88± 38.24	5.12± 7.57	1.41± 2.83	3.27± 6.01
P	79.72± 11.64	82.02± 10.83	80.87± 11.30	83.31± 10.90	78.50± 9.30	80.91± 10.41	6.48± 15.68	7.38± 15.00	6.93± 15.34	0.86± 1.95	1.09± 2.49	0.98± 2.24
Q	78.11± 16.30	46.39± 27.64	62.59± 27.58	79.74± 15.64	47.31± 24.90	63.87± 26.28	15.75± 26.39	32.67± 24.16	24.05± 26.69	2.09± 4.19	6.43± 6.69	4.22± 5.96
R	81.26± 8.63	80.97± 10.36	81.12± 9.53	84.55± 8.06	79.54± 8.05	82.05± 8.43	5.83± 16.56	4.53± 4.46	5.18± 12.14	0.75± 2.69	0.64± 0.66	0.69± 1.96
S	50.71± 30.98	**NA**	50.71± 30.98	50.78± 30.43	**NA**	50.78± 30.43	105.38± 130.33	**NA**	105.38± 130.33	21.47± 29.62	**NA**	21.47± 29.62
T	47.23± 25.31	46.52± 20.26	46.87± 22.87	36.49± 21.60	34.73± 13.85	35.60± 18.11	64.49± 23.82	34.85± 5.88	49.53± 22.76	13.58± 9.00	7.99± 4.13	10.76± 7.52
Overall	81.38± 17.57	81.97± 20.02	81.66± 18.80	82.57± 19.64	80.74± 20.96	81.69± 20.31	17.97± 46.97	11.63± 19.88	14.89± 36.57	3.13± 10.34	2.31± 5.70	2.73± 8.42

3.4. Anatomical regions

Figs. 5 and 6 visualize the case-wise segmentation using heatmaps. Most errors can be observed in the condyle, dentition, and part of the mandibular body. The segmentation performance of the AI model is significantly degraded in regions of impaired mandibular continuity (Case 21 & 65), bone pathology (Case 16 & 61), and osteosynthesis material (Case 17 & 86) (Supplementary Table S3). The segmentation results in Table 3 further demonstrate the differences in segmentation performance across Regions Of Interest (ROIs). The mandibular body performed the worst in terms of HD 95 and MASD. In terms of DSC, the condyle in CBCT had the lowest score of 78.07 %. In addition, the dentition also had the lowest NSD value of 84.16 %, indicating a lack of accurate boundary segmentation in this region. In summary, the mandibular body has the highest segmentation error in the distance-based metrics, whereas the condylar and dentition regions exhibit the lowest DSC and NSD, respectively.

4. Discussion

Although AI models have proven their performance, there are many open questions regarding the integration and limitations of current AI models in clinical routine as well as research. Recent qualitative research confirms that clinicians demand concrete insights into when and why AI fails in clinical settings, suggesting the need for comprehensive socio-technical evaluations (Scott et al., 2024). Based on an experimental study with 20 current state-of-the-art AI models and the analysis of imaging features, patient characteristics, and anatomical regions on segmentation results, we were able to obtain new insights and provide recommendations for optimized social-technical setting, including clinical data acquisition and the requirements for future development of AI-based segmentation. To begin, our study required the creation of a benchmark dataset, as directly using public datasets or random sampling of private cases would not have been appropriate. Public datasets may overlap with the training data of the models under evaluation, and random sampling of private cases could not ensure a balance of imaging and patient features necessary for statistical analysis. Accordingly, we built a benchmark dataset that is balanced across imaging and patient factors, using the distribution of 1,788 cases from public datasets as reference (Aerts et al., 2014; Cipriano et al., 2022; Cptac, 2018; Grossberg et al., 2020; Vallières et al., 2017a, 2017b; Wee and Dekker, 2019). The required number of volumes was determined through a prior sample-size calculation in our study protocol. Previous studies have shown that resampling could simulate multiple CBCT/CT scans from the same patient in a different image reconstruction settings (Naziroglu et al., 2017). Rotational movements of the patient's head could also be simulated using resampling methods (Gardner et al., 2023). Hence, we have created a quasi-experiment setting by resampling

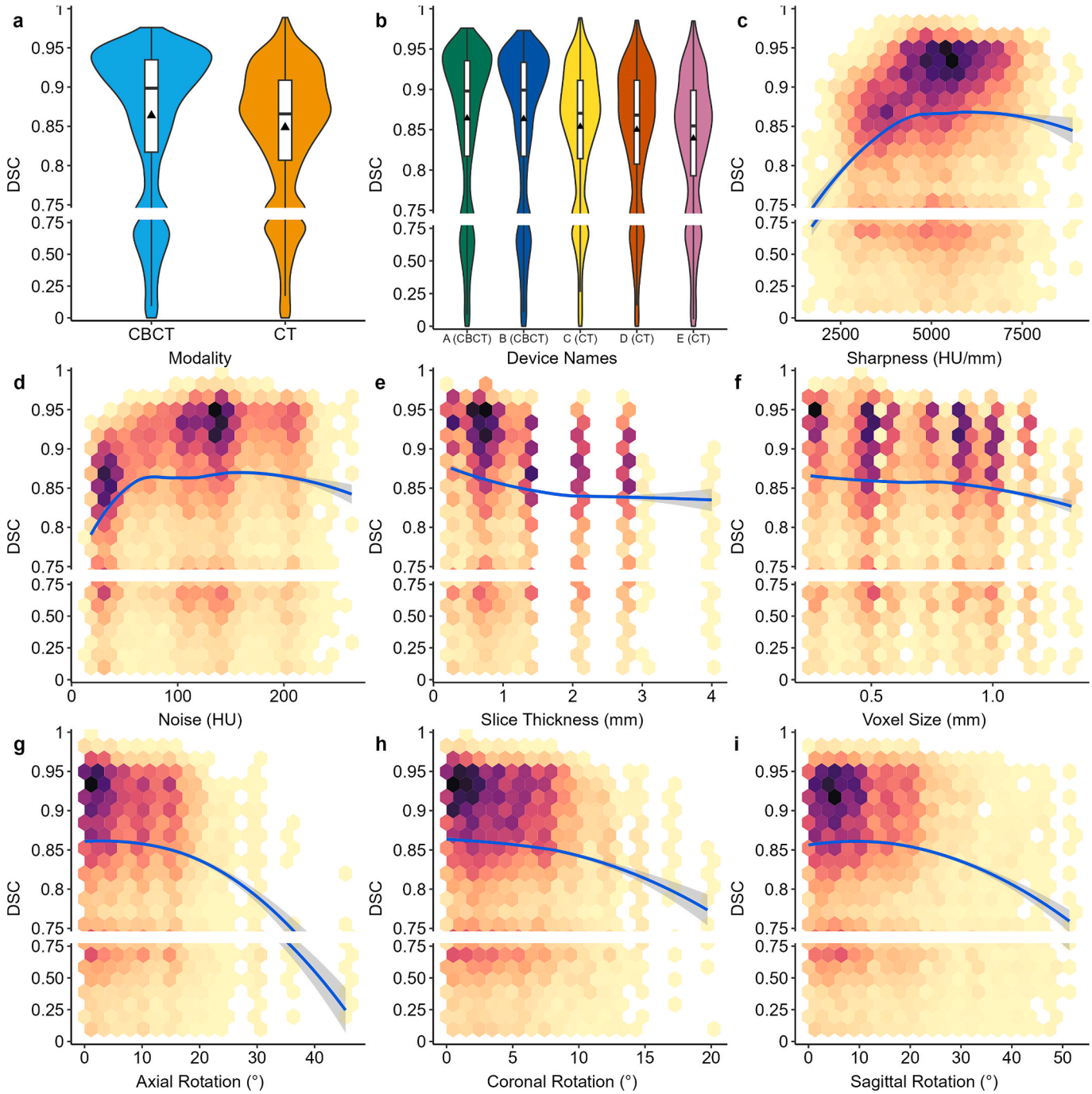


Fig. 3. Image quality related factors and segmentation performance measured in DSC (**a**) distribution of segmentation performance in CBCT and CT scans (**b**) segmentation performance in five devices used in the study (**c**) relationship between image sharpness and segmentation performance (**d**) the effect of image noise image noise and segmentation performance (**e**) relationship between slice thickness and segmentation performance (**f**) the impact of voxel size on segmentation performance (**g**) ~ (**i**) the effect of bone rotation on segmentation performance. Colored hexagonal bins represent the distribution of data points. Darker colors indicate higher data density, while brighter colors indicate lower data density.

original CT/CBCT scans and manual screening of patient characteristics. This method provides enough data for the LMM to reveal the underlying factors influencing the performance of AI models.

4.1. Regulations on AI models

Among the 20 models selected for this study, the overall segmentation performance of the commercial models that had received MDR/FDA approval was higher than that of the research models (Fig. 2d). This suggests a positive impact of regulatory policies on the commercial model development and deployment process. However, the costs associated with certifying software as a medical device could be substantial. Regardless of the type of model, monitoring post-deployment performance is a critical step in improving safety as well as the effectiveness of AI models in clinical practice (Pesapane et al., 2024). This is also a key feature of the overall product lifecycle approach used by the FDA (FDA, 2024). As our study demonstrates, end-users should expect degradation in the performance of current static AI models as a result of changes in imaging protocols or changes in patient populations. One possible solution is dynamic fine-tuning of deployed models. However, the changes

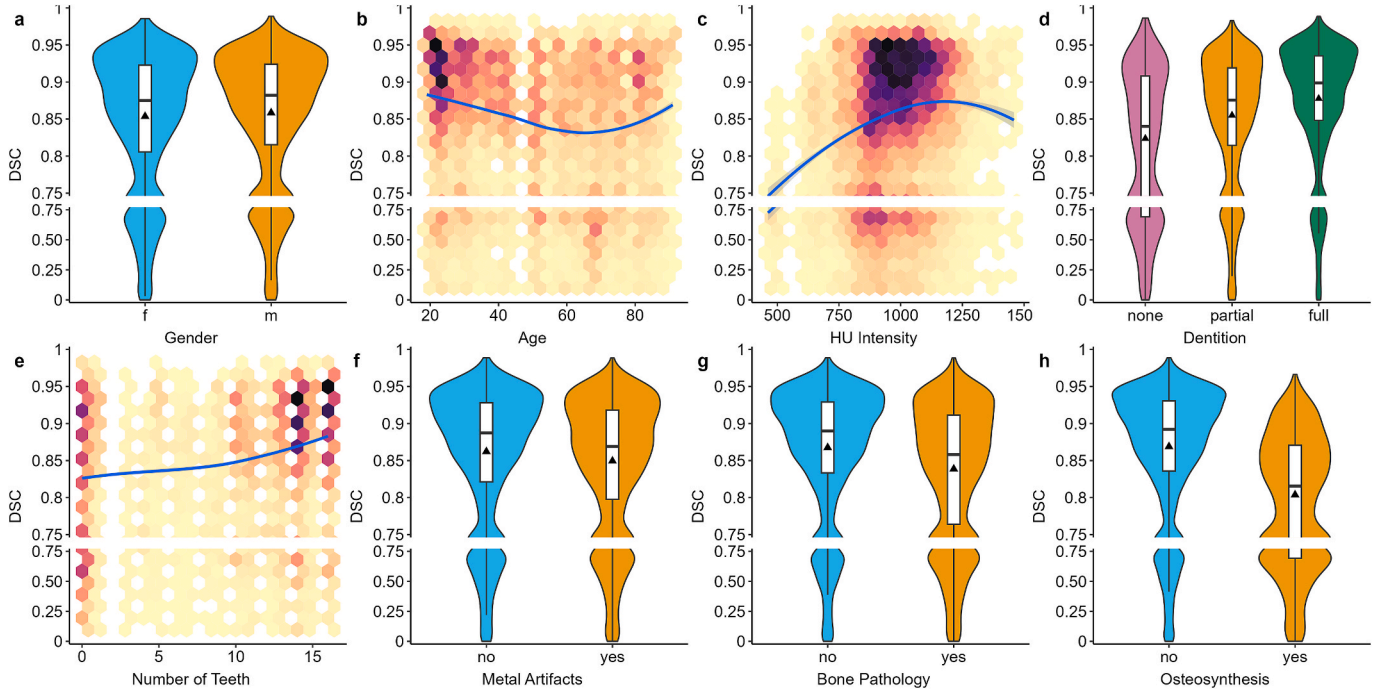


Fig. 4. Patient related factors and segmentation performance (**a**)Comparison of between female and male patients (**b**) relationship between age and segmentation performance (**c**)The effect of Hounsfield Unit (HU) intensity on segmentation performance (**d**)(**e**) The impact of dentition status and teeth count on segmentation performance **(f)** Comparison of segmentation performance between cases with and without metal artifacts **(g)** Influence of bone pathology on segmentation performance **(h)** The effect of osteosynthesis on segmentation performance.

in performance as well as risk associated with this continuous learning may cause the product's metrics to differ from those at the time of initial certification, which would pose a significant regulatory challenge (Pianykh et al., 2020). While regulators are actively developing guidance policies for dynamic tuning models, all approved AI tools have been static up to this date (European Commission, 2024; Fda, 2024). Therefore, the optimization of image acquisition protocols may be a viable alternative solution on static models. Furthermore, the identification of patient characteristics and anatomical regions that cause performance declines could lead to a strategy for intervening, both in the development of AI models as well as in their application.

4.2. *Imaging factors and modality*

The first questions arise in the optimal reconstruction protocol during the acquisition of medical imaging. Our investigation of one of the most versatile human bones, the mandible, suggests several key factors affecting the quality of AI-based bone segmentation. Elevated sharpness, decreased voxel size, and ensuring standardized patient positioning can all improve AI-based segmentation to a certain degree (Fig. 3). The results are in accordance with findings from traditional segmentation algorithms. Puggelli et al. reconstructed CT scans of porcine tibiae with different kernels and evaluated the segmentation accuracy compared to laser scanning (Puggelli et al., 2019). The results demonstrated that sharp reconstruction kernel accuracy was higher than that of the soft kernel. The reason for that may be because the bone-soft tissue boundary is better defined in these images. Similarly, another study based on the segmentation results on CBCT scans of an AI model of 11 dry mandibles with different voxel sizes, revealed that larger voxels (0.45 mm) resulted in significant segmentation errors compared to smaller voxels (0.15 mm) (surface scans as reference) (Alrashed et al., 2024). In contrast, Huang et al. concluded when applying one single AI model onto 183 CT scans of 11 patients with different voxel sizes, slice thickness and simulated doses, that there is no need for a strict image resolution (Huang et al., 2021). Our comprehensive analysis with 20 models, however, underlined that lower sharpness (increased blurriness) as well as larger voxel size may have a negative impact on segmentation performance. This should be considered in the reconstruction protocols when incorporating AI models.

Another important factor is bone rotation during scanning (in our case the mandible). El Bachaoui et al. collected a total of 20 CBCT scans from 5 fresh cadavers at four different positions (El Bachaoui et al., 2023). They concluded that the effect of sagittal rotation of the head on segmentation accuracy is clinically negligible (manual segmentation as reference). However, this study investigated a limited range of rotations in the sagittal plane only. In contrast, our study included a wide range of combined rotations in all three reference planes. Our results show that bone rotation in the axial and sagittal planes negatively affects the segmentation results (Fig. 3). This finding is probably due to the underlying distribution of the training data used by AI models. Attention should be paid to the standard positioning of the mandible, especially during CT scanning, as there is more freedom of movement for mandible on supine CT scans that lack chin fixation compared to CBCT. If a proper bone positioning cannot be achieved, post processing into a normalized bone position should be considered.

Regarding the imaging modality, most of the models trained with single modal data (CBCT or CT) were also able to segment scans of the other modality. Only one model, which trained solely on CT data, was unable to do so, as it successfully extracted the skull but was unable to separate the mandible from it. Such results indicate that CBCT and CT are interchangeable in this task, likely due to their similar fundamental imaging principles. Nevertheless, AI segmentation on CBCT demonstrated higher accuracy in descriptive statistics, but the AI model was even better at segmenting the CT data in LMM analysis which took multiple factors into account. The main reason for this may be that the original voxel size of CBCT (0.268 mm in average) is smaller than that of CT (0.442 mm in average), and smaller voxels size leads to better segmentation (Fig. 7). Another reason could be the anisotropy of CT voxels, i.e., slice thickness is generally not equal to in-plane voxel size. In previous studies, this negative effect was predominantly observed in the

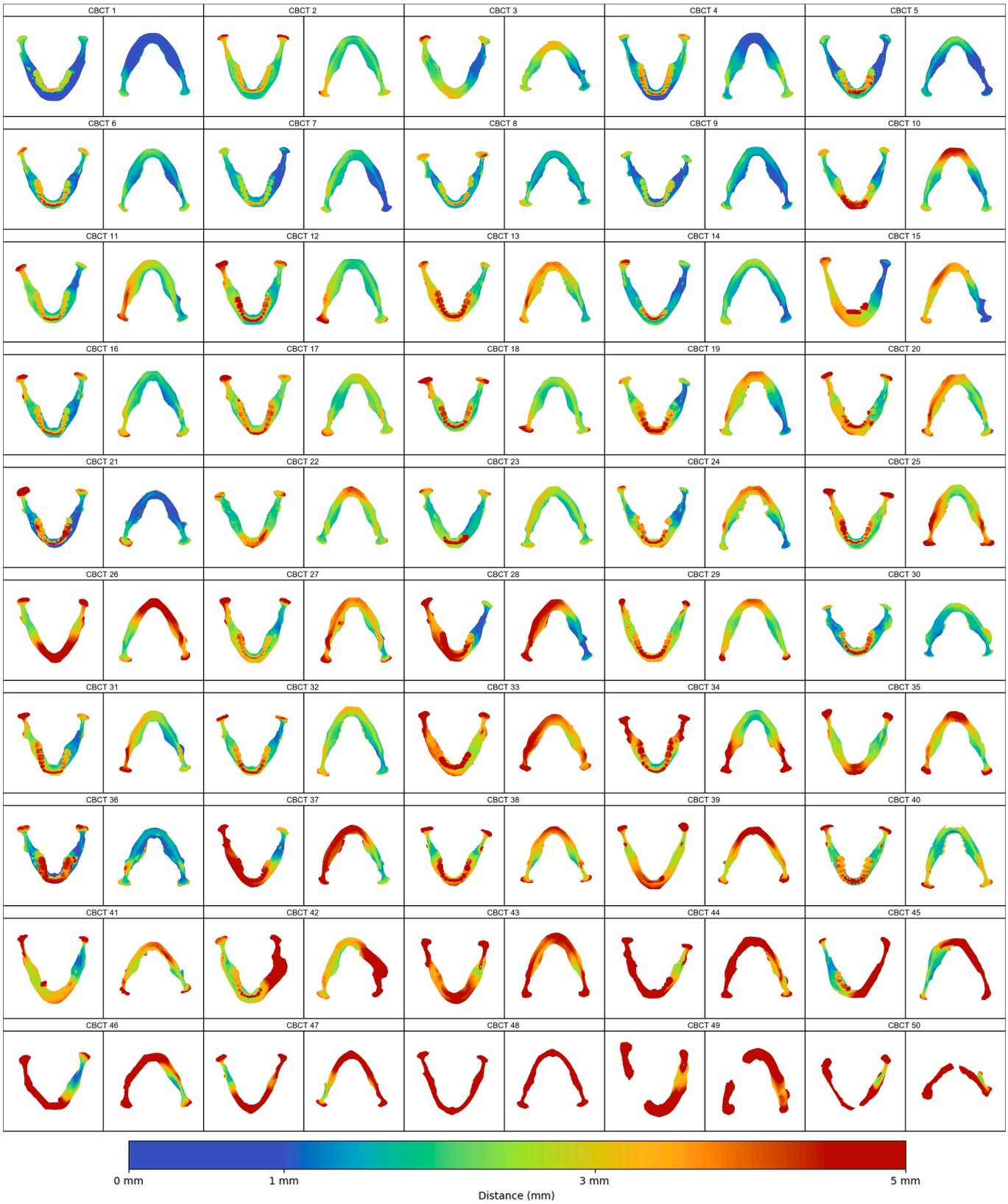


Fig. 5. Heatmaps showing the average surface distance between AI segmentation results and the ground truths of CBCT scans. These segmentations were performed on the original scan and the 9 resample variants by 19 models (failed in model S), resulting in around 190 segmentations per case. Cases arranged in descending order of overall mean DSC.

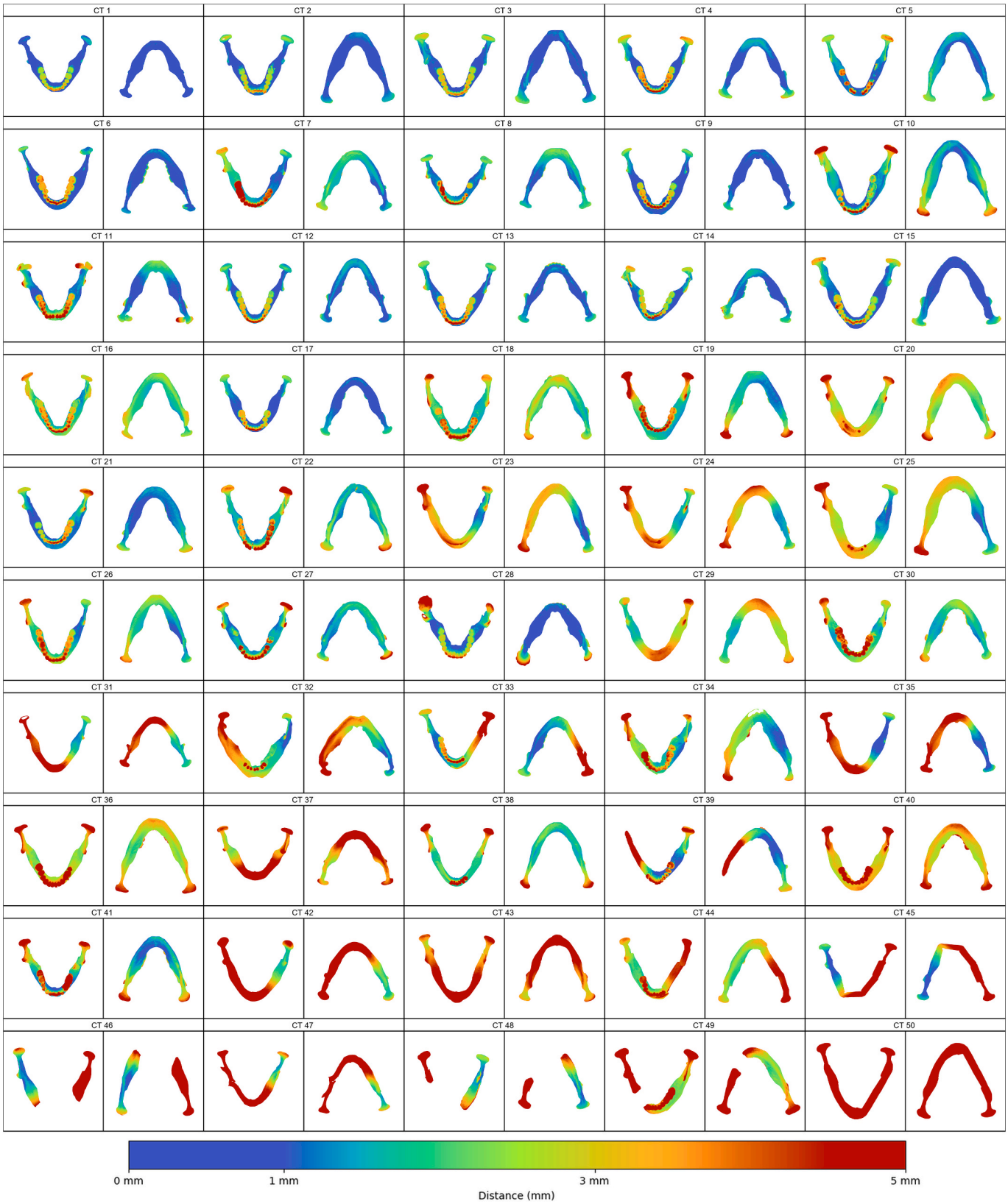


Fig. 6. Heatmaps showing the average surface distance between AI segmentation results and the ground truths of CBCT scans. These segmentations were performed on the original scan and the 9 resample variants by 20 models, resulting in around 200 segmentations per case. Cases arranged in descending order of overall mean DSC.

Table 3
Anatomical Region Summary. *Performance of AI models (mean ± sd) on 5 anatomical regions and the whole mandible. Worst performances were marked in bold.*

ROI	DSC (%)			NSD (%)			HD 95 (mm)			MASD (mm)		
	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall
Condyle	82.26± 19.19	**78.07± 23.54**	**80.26± 21.48**	90.46± 17.37	**82.24± 22.10**	86.53± 20.19	2.24± 4.31	2.71± 2.92	2.47± 3.72	0.43± 0.89	0.62± 0.84	0.52± 0.87
Dentition	**80.15± 19.31**	83.13± 18.10	81.60± 18.79	85.41± 18.94	82.83± 18.58	**84.16± 18.81**	7.99± 26.86	5.42± 9.56	6.74± 20.41	1.51± 6.01	1.11± 2.81	1.32± 4.74
IAN Foramen	82.19± 14.45	85.58± 13.28	83.82± 14.00	95.94± 9.19	96.46± 8.77	96.19± 8.99	1.09± 1.02	0.93± 0.52	1.01± 0.82	0.12± 0.23	0.12± 0.15	0.12± 0.19
Inferior Border	84.64± 16.84	85.31± 18.81	84.97± 17.83	87.19± 17.44	84.08± 19.48	85.66± 18.53	10.11± 35.85	8.24± 15.90	9.19± 27.88	1.80± 7.74	1.54± 4.02	1.67± 6.19
Mandible Body	80.51± 18.36	82.37± 21.13	81.41± 19.78	**85.02± 20.49**	85.39± 21.93	85.20± 21.20	**21.67± 60.00**	**10.23± 20.96**	**16.10± 45.76**	**4.30± 14.22**	**2.41± 6.67**	**3.38± 11.24**
Whole Mandible	81.38± 17.57	81.97± 20.02	81.66± 18.80	82.57± 19.64	80.74± 20.96	81.69± 20.31	17.97± 46.97	11.63± 19.88	14.89± 36.57	3.13± 10.34	2.31± 5.70	2.73± 8.42

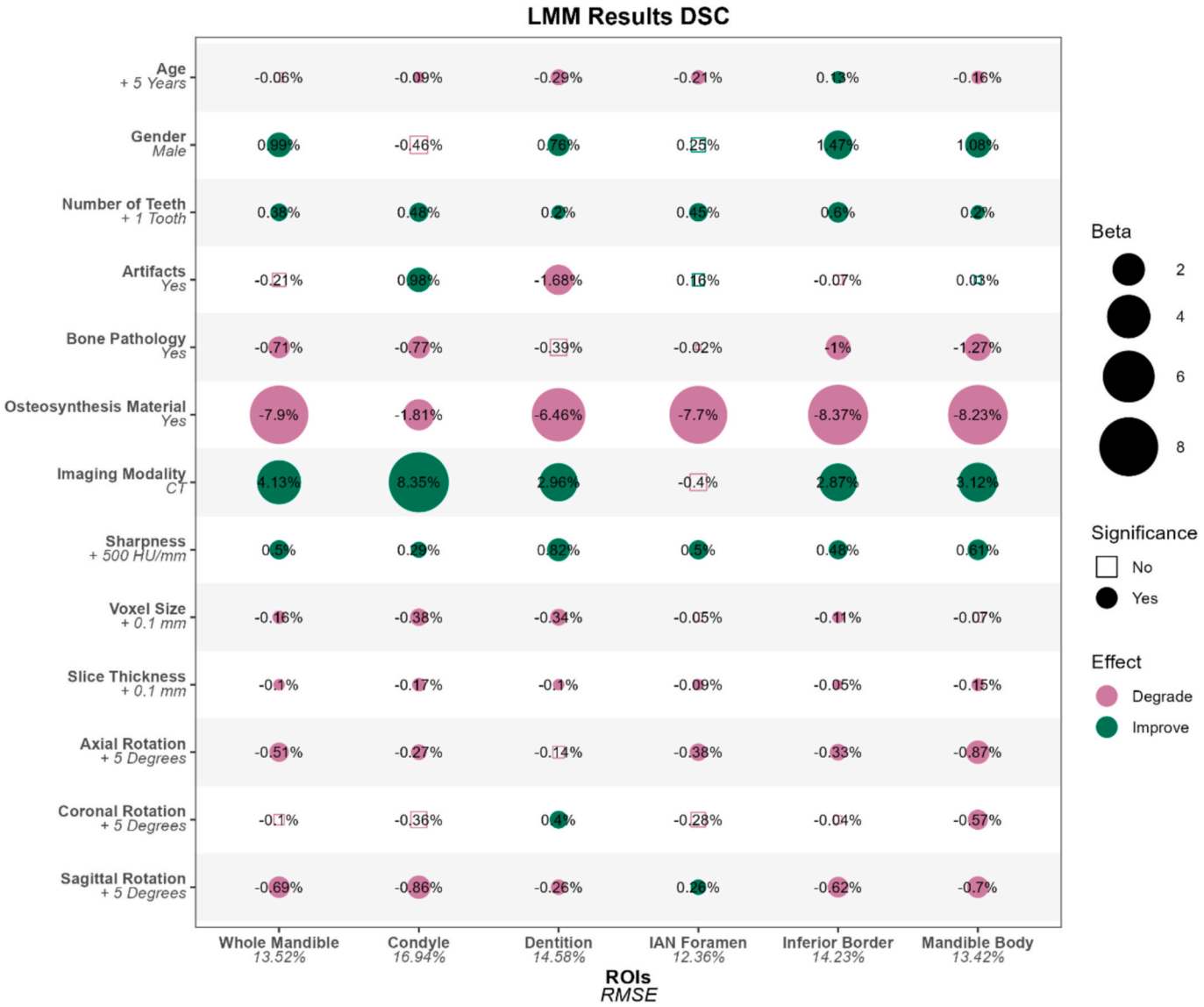


Fig. 7. LMMs fitted on evaluation results in DSC% of five ROIs and the whole mandible. Factor considered significant when $p < 0.05$.

inter-slice direction, with the main areas affected including the cranial side of the condyle, the inferior border of the mandible, and the alveolar ridge, which is also observed in a previous study (Gruber et al., 2024). In contrast, LMM considers voxel size and slice thickness as independent factors, avoiding the interference of voxel morphology on modality. In conclusion, the use of high-resolution CT scans with isotropic voxels may further improve bone segmentation results of AI models.

4.3. Patient-related factors and regions of interests

Beside image-related factors, patient-related factors may also affect segmentation accuracy. Our results showed slightly better segmentation performance in males (LMM, $\beta = 0.99$ %, $p < 0.001$) (Fig. 4). Yet this difference is marginal, it suggests that the AI models can be readily applied to both sexes. Interestingly, the presence of teeth improved segmentation results (for each additional tooth, LMM, $\beta = 0.38$ %, $p < 0.001$). A possible explanation is that teeth act as extra anatomical landmarks for the AI models. Lacking teethless training data could also be a reason. Although restorations and implants are typically the source of artifacts, LMM analysis considered artifacts an individual factor, allowing our study to identify the impact of teeth on segmentation outcomes. However, bone pathology and osteosynthesis materials significantly reduced accuracy. This result aligns to that from the study of Cui et al. of one single AI model, where evaluated on an external dataset of 407 CBCT scans, missing teeth (DSC, −0.8 %), malocclusion (DSC, −0.9 %), and metal artifacts (DSC, −2.0 %) negatively affected segmentation results (Cui et al., 2022).

The accuracy of mandible segmentation varies in different anatomical regions (Table 3). The condyle exhibits lower accuracy, primarily due to its thin cortical bone and low density of cancellous bone, as well as the surrounding high-density cranial base structures. This results in lower contrast in the condylar region, especially in CBCT images (Cuadros Linares et al., 2019). This was confirmed by our LMM analysis across anatomical regions, where the segmentation performance of the condylar region in CT images is improved by 8.59 % in DSC (LMM, $\beta = 8.35$ %, $p < 0.001$) compared to CBCT images, while the improvement of the whole mandible segmentation is merely 4.1 % (LMM, $\beta = 4.13$ %, $p < 0.001$,Fig. 7). The mandible body also exhibits a higher degree of error in segmentation, which may partially be attributed to the presence of artifacts from the crowns and brackets (Hirschinger et al., 2015). Another reason for the drop in the performance on the mandibular body is the discontinuity of the mandible, often accompanied by large osteosynthesis reconstruction plates (Figs. 5 & 6). This could lead to a partial or total segmentation failure, which in turn severely affects the overall segmentation performance of the mandibular body and the whole mandible.

4.4. Clinical applicability

Ideally, AI segmentation models should not be sensitive to reconstruction protocols, patient factors, and anatomical regions, which are highly variable in a socio-technical system. However, due to the limitations in architecture and training data, the current models have not yet reached this goal. Nevertheless, according to our findings, the segmentation performance of the model can be improved by optimizing the imaging protocol. Simulated calculation with results from LMMs suggested that with a recommended protocol (CT scan, sharpness of about 5000 HU/mm, voxel size of 0.5 mm, and neutral bone position), an increase of 9.02 % in DSC for AI segmentation can be expected, comparing to the worst combination. In terms of patient characteristics, AI segmentation on a young male with complete dentition, without artifacts, pathology, or osteosynthesis, the DSC would increase by 16.59 % compared to the worst combination of features. With these two aspects into account, the difference in DSC between the cases adapted most to fitting predicted requirements of AI models in general and those least adapted would be 25 %. A real pair of examples can be found in our dataset (Case 21 and Case 78, Supplementary Table S2), where the mean DSC for AI segmentation of the original volume was 71.82 % and 91.49 %, respectively, with a difference of 19.67 %. This 20 % difference in DSC is substantial in terms of workload since cases with DSC above 90 % require minor adjustment and those below 75 % need intensive manual involvement (Sarria et al., 2024).

Nevertheless, studies have shown that manual refinement on AI segmentation results with low DSC is still faster than full manual segmentation (Jindanil et al., 2023; Walker et al., 2014). However, the segmentation failure caused by osteosynthesis or continuity loss mentioned in the previous section requires manual thorough coverage, and this extreme condition is also a considerable challenge for clinicians. Furthermore, medical products like surgical guide and patient-specific implants are highly dependent on accurate surface tomography. A poor surface accuracy can introduce substantial errors in downstream surgical planning, potentially affecting clinical outcomes.

4.5. Recommendations and requirements

To narrow this performance gap in clinical practice, collaboration between clinicians and AI developers must focus on mutual adjustments informed by real-world needs. Clinicians can optimize imaging protocols to align with current AI capabilities, while developers should prioritize the requirements that address recurring clinical challenges.

For clinicians, understanding the technical limits of AI models is critical. To improve bone segmentation outcomes, we recommend using CT scans with small, isotropic voxels (0.5 mm or smaller) and high-sharpness protocols when possible. In terms of modality, clinicians should be aware of the potential performance drop in susceptible regions like condyles in CBCT. Also, ensure target bones are positioned neutrally during scans, if not possible (e.g. trauma), adjust the images to a standard orientation before segmentation. In cases with edentulous mandible, large implants, or bone pathologies, clinicians should expect lower accuracy and prepare for manual corrections.

For AI developers, the next-gen models should be stable in performance even when faced with non-ideal clinical conditions. This includes robustness to patient features like bone pathology and osteosynthesis. Considering the sparsity of specific patient group, synthetics data can be a viable option. Segmentation performance in complex anatomical regions (e.g. condyles) should be prioritized, which could be achieved through regionally weighted loss functions or adversarial training for specific structures. In addition, models should explicitly flag uncertain or low-confidence segmentation regions by heatmaps or scores to guide clinician review, particularly in high-risk cases involving bone pathologies or surgical planning.

Our recommendations above are already backed by existing imaging standards. The EU Radiation-Protection Guideline for dental and maxillofacial CBCT endorses sub-0.4-mm voxels and routine quality checks (European Commission: Directorate-General for Energy, 2012). Similarly, the 2024 Swiss consensus guidelines (Lübbers et al., 2024) specify high-resolution protocols and mandatory expert review when metal hardware or bone defects are present, validating our call for extra caution in osteosynthesis and pathology cases. The 2024 European Commission report on CT justification also urges adoption of objective, pre-scan acceptance criteria and regular audits, a role that our recommendations can play within PACS-integrated quality controls (European Commission: Directorate-General for Energy et al., 2024). Because current professional guidelines already point in the same direction, adopting our empirically derived thresholds becomes not an extra burden but an evidence-strengthened extension of best practice.

4.6. Limitation

Our study recruited the largest number of AI models to date and comprehensively analyzed the socio-technical factors including patient factors and imaging factors on segmentation performance. The results showed that in general models performed worse on downsampled volumes compared to the original volume in all metrics (Supplementary Table S5). Yet four models (F, N, P, Q) exhibited better performance on downsampled subset instead, which appear to be more adaptable to these volumes than original ones. However, due to confidentiality concerns of the models, we are unable to further discuss the source of this adaptability. Another limitation is that our dataset did not include cases under the age of 18 years because they are not common cases for

mandibular bone segmentation. This prevented us from fully capturing anatomical variability in all clinical situations, especially in patients who grow and develop during childhood and adolescence.

4.7. Future work

On our benchmark dataset, the current models still have a certain number of unsatisfying segmentation results, and clinicians need to refine them manually using various tools (Fig. 2a). Further refinement using interactive automated segmentation algorithms, such MedSAM (Ma et al., 2024) and nnInteractive (Isensee et al., 2025); could streamline this "last mile" by allowing clinicians to correct errors via intuitive prompts. Additionally, this study only briefly summarized the basic architecture used by the models, and due to confidentiality reasons, we were not able to examine in detail the configuration of the training parameters of each model. A meta-analysis on tooth segmentation models classified and evaluated various models based on backbone networks, demonstrating the evolution of algorithms and their performances over time (Kot et al., 2025). Future research could explore these technical factors, potentially by collaborating to configure models and data in a controlled environment for further experiments.

5. Conclusion

This multi-center study shows that the performance of AI mandible segmentation is dynamically shaped by socio-technical factors, including imaging protocols, patient-specific factors and anatomical complexity. Two pillars are essential to the success of clinical translation of AI models: clinicians should adapt their workflows to the current limitations of AI, and developers must tackle the upcoming requirements that address persistent clinical challenges. For clinical teams, this means choosing high-resolution CT protocols when possible, ensuring standardized patient positioning and rechecking AI output in cases involving bone pathology or osteosynthesis. For AI developers, the requirements for the next-gen AI segmentation models are summarized from clinical failures. Models must remain robust to common clinical variabilities like rotation. Models should further improve the accuracy of error-prone anatomical regions (e.g., condyles) and provide intuitive uncertainty feedback to guide clinical reviews. Beyond these technical recommendations, our quantified operating suggestions could also serve as objective criteria in pre-market submissions for regulatory review. These are not standalone checklists but interconnected obligations. Only through this dual commitment can AI progress from a static algorithm and technocentric preoccupation to a trustworthy clinical ally in a socio-technical system.

Institutional Review Board Statement

The ethics application of the study was approved by the ethics committee at the RWTH Aachen University (approval number EK 23–312, 26th October 2023, Prof. Dr. Ralf Hausmann).

Informed Consent Statement

No informed consent was needed due to the use of anonymized retrospective patient data.

Code Availability

The code for dataset preparation and model evaluation were implemented in Python (Version 3.11.0). The source code and R code for statistical analysis is available on GitHub (https://github.com/OMFSdigital/AI_Mandible_Benchmarking).

CRediT authorship contribution statement

Kunpeng Xie: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Writing – review & editing. **Lennart Johannes Gruber:** Formal analysis, Investigation, Data curation, Writing – review & editing. **Martin Crampen:** Software, Formal analysis, Investigation, Writing – review & editing. **Yao Li:** Validation, Formal analysis, Writing – review & editing. **André Ferreira:** Validation, Formal analysis, Writing – review & editing. **Elias Tappeiner:** Formal analysis, Resources, Writing – review & editing. **Maxime Gillot:** Formal analysis, Resources, Writing – review & editing. **Jan Schepers:** Formal analysis, Resources, Writing – review & editing. **Jiangchang Xu:** Formal analysis, Resources, Writing – review & editing. **Tobias Pankert:** Formal analysis, Resources, Writing – review & editing. **Michel Beyer:** Formal analysis, Resources, Writing – review & editing. **Negar Shahamiri:** Formal analysis, Resources, Writing – review & editing. **Reinier ten Brink:** Formal analysis, Resources, Writing – review & editing. **Gauthier Dot:** Formal analysis, Resources, Writing – review & editing. **Charlotte Weschke:** Formal analysis, Resources, Writing – review & editing. **Niels van Nistelrooij:** Formal analysis, Resources, Writing – review & editing. **Pieter-Jan Verhelst:** Formal analysis, Resources, Writing – review & editing. **Yan Guo:** Formal analysis, Writing – review & editing. **Zhibin Xu:** Formal analysis, Writing – review & editing. **Jonas Bienzeisler:** Formal analysis, Writing – review & editing. **Ashkan Rashad:** Formal analysis, Writing – review & editing. **Tabea Flügge:** Formal analysis, Resources, Writing – review & editing. **Ross Cotton:** Formal analysis, Resources, Writing – review & editing. **Shankeeth Vinayahalingam:** Formal analysis, Resources, Writing – review & editing. **Robert Ilesan:** Formal analysis, Resources, Writing – review & editing. **Stefan Raith:** Formal analysis, Resources, Writing – review & editing. **Dennis Madsen:** Formal analysis, Resources, Writing – review & editing. **Constantin Seibold:** Formal analysis, Resources, Writing – review & editing. **Tong Xi:** Formal analysis, Resources, Writing – review & editing. **Stefaan Bergé:** Formal analysis, Resources, Writing – review & editing. **Sven Nebelung:** Formal analysis, Resources, Writing – review & editing. **Oldřich Kodym:** Formal analysis, Resources, Writing – review & editing. **Osku Sundqvist:** Formal analysis, Resources, Writing – review & editing. **Florian M. Thieringer:** Formal analysis, Resources, Writing – review & editing. **Hans Lamecker:** Formal analysis, Resources, Writing – review & editing. **Antoine Coppens:** Formal analysis, Resources, Writing – review & editing. **Thomas Potrusil:** Formal analysis, Resources, Writing – review & editing. **Joep Kraeima:** Formal analysis, Writing – review & editing. **Max Witjes:** Formal analysis, Resources, Writing – review & editing. **Guomin Wu:** Formal analysis, Resources, Writing – review & editing. **Xiaojun Chen:** Formal analysis, Resources, Writing – review & editing. **Adriaan Lambrechts:** Formal analysis, Resources, Writing – review & editing. **Lucia H Soares Cevidanes:** Formal analysis, Writing – review & editing. **Stefan Zachow:** Formal analysis, Resources, Writing – review & editing. **Alexander Hermans:** Formal analysis, Writing – review & editing. **Daniel Truhn:** Formal analysis, Writing – review & editing. **Victor Alves:** Formal analysis, Writing – review & editing. **Jan Egger:** Formal analysis, Resources, Writing – review & editing. **Rainer Röhrig:** Formal analysis, Resources, Writing – review & editing. **Frank Hölzle:** Formal analysis, Resources, Writing – review & editing. **Behrus Hinrichs-Puladi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Funding

This research received no external funding.

Declaration of competing interest

This research employs eight commercial AI models from companies. Some of the co-authors are employed by or have financial ties with these companies. Jan Schepers and Adriaan Lambrechts are employed by Materialise NV. Tobias Pankert and Stefan Raith are employed by Inzipio GmbH. Charlotte Weschke and Hans Lamecker are employed by 1000shapes. Ross Cotton is employed by Synopsys Northern Europe Ltd. Oldřich Kodym is employed by TESCAN 3DIM. Antoine Coppens is employed by Relu BV. Thomas Potrusil is employed by CADS GmbH and

KLS Martin Group. Osku Sundqvist is employed by Planmeca Oy. It is important to note that the companies and institutions only provided model information and conducted inference on the benchmark dataset, without involvement in ground truth annotation, investigation or raw data analysis. In addition, model performance data have been anonymized for all authors (except for Kunpeng Xie and Behrus Hinrichs-Puladi) using model designation codes. Despite these relationships, all necessary measures were taken during the study's design, data collection, and analysis to ensure the objectivity and integrity of the research findings. The author Behrus Hinrichs-Puladi and Jan Egger are members of the editorial board of the International Journal "Expert Systems with Applications". All other authors declare no conflicts of interest.

Acknowledgement

We thank the anonymous patients whose CT and CBCT scans formed the basis of this study. Open access funding provided by the Open Access Publishing Fund of RWTH Aachen University. Behrus Hinrichs-Puladi was funded by the Medical Faculty of RWTH Aachen University as part of the Clinician Scientist Program. During the preparation of this work the authors used DeepL in order to polish the English language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2025.130031.

Data availability

Due to the model anonymity nature of the study, only the evaluation result with code names of the model is made available in our repository. Benchmarking dataset and the model predictions are available on request from the corresponding author.

References

Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., & Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications, 5*, 4006. https://doi.org/10.1038/ncomms5006

Al-Hassani, R. T. (2025). Artificial Intelligence for Medical Diagnostics in IoT-Based Healthcare Networks: Foundations and Future Trends. *Babylonian Journal of Networking, 2025*, 70–79. https://doi.org/10.58496/BJN/2025/006.

Al-Mistarehi, A.-H., M. Mijwil, M., Filali, Y., Bounabi, M., Ali, G., & Abotaleb, M. (2023). Artificial Intelligence Solutions for Health 4.0: Overcoming Challenges and Surveying Applications. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, *2023*, 15–20. https://doi.org/10.58496/MJAIH/2023/003.

Alrashed, S., Dutra, V., Chu, T.-M.-G., Yang, C.-C., & Lin, W.-S. (2024). Influence of exposure protocol, voxel size, and artifact removal algorithm on the trueness of segmentation utilizing an artificial-intelligence-based system. *Journal of Prosthodontics: Official Journal of the American College of Prosthodontists, 33*(6), 574–583. https://doi.org/10.1111/jopr.13827

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., & Cardoso, M. J. (2022). The Medical Segmentation Decathlon. *Nature Communications, 13*(1), 4128. https://doi.org/10.1038/s41467-022-30695-9

Apostolakis, D., Michelinakis, G., Kamposiora, P., & Papavasiliou, G. (2022). The current state of computer assisted orthognathic surgery: A narrative review. *Journal of Dentistry, 119*, Article 104052. https://doi.org/10.1016/j.jdent.2022.104052

Benlalia, Z., Al-Kateb, G., Mzili, M., & Diamanti, K. (2025). Reimagining Intelligence: A Comprehensive Review of Human-Centric AI Systems and their Societal and Healthcare Integration. *Mesopotamian Journal of Artificial Intelligence in Healthcare, 2025*, 173–186. https://doi.org/10.58496/MJAIH/2025/017.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3d U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Lecture Notes in Computer Science: Vol. 9901, Medical Image Computing and Computer-Assisted Intervention {u2013} MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II* (pp. 424–432). Cham: Springer International Publishing; Imprint: Springer. https://doi.org/10.1007/978-3-319-46723-8_49.

Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., & Grana, C. (2022). Deep Segmentation of the Mandibular Canal: A New 3D Annotated Dataset of CBCT Volumes. *IEEE Access, 10*, 11500–11510. https://doi.org/10.1109/ACCESS.2022.3144840

Cptac. (2018). *Head and Neck Squamous Cell Carcinoma Collection (CPTAC-HNSCC)*. *The Cancer Imaging Archive*. https://doi.org/10.7937/K9/TCIA.2018.UW45NH81

Cuadros Linares, O., Bianchi, J., Raveli, D., Batista Neto, J., & Hamann, B. (2019). Mandible and skull segmentation in cone beam computed tomography using super-voxels and graph clustering. *The Visual Computer, 35*(10), 1461–1474. https://doi.org/10.1007/s00371-018-1511-0

Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J. [Jiameng], . . . Shen, D. (2022). A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nature Communications*, *13*(1), 2096. https://doi.org/10.1038/s41467-022-29637-2.

Dot, G., Chaurasia, A., Dubois, G., Savoldelli, C., Haghighat, S., Azimian, S., & Gajny, L. (2024). Dentalsegmentator: Robust open source deep learning-based CT and CBCT image segmentation. *Journal of Dentistry, 147*, Article 105130. https://doi.org/10.1016/j.jdent.2024.105130

El Bachaoui, S., Verhelst, P.-J., de Faria Vasconcelos, K., Shaheen, E., Coucke, W., Swennen, G., & Politis, C. (2023). The impact of CBCT-head tilting on 3D condylar segmentation reproducibility. *Dento Maxillo Facial Radiology, 52*(6), Article 20230072. https://doi.org/10.1259/dmfr.20230072

European Commission (2024, December 10). Article 96: Guidelines from the Commission on the Implementation of this Regulation | EU Artificial Intelligence Act. Retrieved from https://artificialintelligenceact.eu/article/96/.

European Commission: Directorate-General for Energy (2012). *Cone beam CT for dental and maxillofacial radiology: evidence-based guidelines*. Luxembourg. Retrieved from Publications Office of the European Union website: https://data.europa.eu/doi/10.2768/21874 https://doi.org/10.2768/21874.

European Commission: Directorate-General for Energy, Brkljačić, B., Karoussou-Schreiner, A., Sosna, J., Foley, S., Bly, R., . . . Szucsich, M. (2024). *European co-ordinated action on improving justification of computed tomography: results and recommendations from a first-time multi-national study on CT justification in the EU*. Luxembourg. Retrieved from Publications Office of the European Union website: https://data.europa.eu/doi/10.2833/80267 https://doi.org/10.2833/80267.

FDA (2024a). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. Retrieved from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence.

FDA (2024b). US FDA Artificial Intelligence and Machine Learning Discussion Paper. Retrieved from https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2021). A review of deep learning based methods for medical image multi-organ segmentation. *1120-1797*, *85*, 107–122. https://doi.org/10.1016/j.ejmp.2021.05.003.

Gardner, M., Bouchta, Y. B., Sykes, J., & Keall, P. J. (2023). A kinematics-based method for creating deformed patient-derived head and neck CT scans. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2023*, 1–4. https://doi.org/10.1109/EMBC40787.2023.10340930

Gillot, M., Baquero, B., Le, C., Deleat-Besson, R., Bianchi, J., Ruellas, A., & Prieto, J. C. (2022). Automatic multi-anatomical skull structure segmentation of cone-beam computed tomography scans using 3D UNETR. *PLoS One1, 17*(10), Article e0275033. https://doi.org/10.1371/journal.pone.0275033

Google DeepMind (2022). Surface distance metrics [Computer software]. Github. Retrieved from [Web] https://github.com/google-deepmind/surface-distance.

Grossberg, A., Elhalawani, H., Mohamed, A., Mulder, S., Williams, B., White, A. L., . . . M.D. Anderson Cancer Center Head And Neck Quantitative Imaging Working Group (2020). *HNSCC*. The Cancer Imaging Archive. https://doi.org/10.7937/k9/tcia.2020.a8sh-7363.

Gruber, L. J., Egger, J., Bönsch, A., Kraeima, J., Ulbrich, M., van den Bosch, V., & Puladi, B. (2024). Accuracy and Precision of Mandible Segmentation and its Clinical Implications: Virtual reality, Desktop Screen and Artificial Intelligence. *Expert Systems with Applications, 239*, Article 122275. https://doi.org/10.1016/j.eswa.2023.122275

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022, January 4). *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. Retrieved from http://arxiv.org/pdf/2201.01266.

Hirschinger, V., Hanke, S., Hirschfelder, U., & Hofmann, E. (2015). Artifacts in orthodontic bracket systems in cone-beam computed tomography and multislice computed tomography. *Journal of Orofacial Orthopedics = Fortschritte Der Kieferorthopadie : Organ/official Journal Deutsche Gesellschaft Fur Kieferorthopadie, 76 (2), 152–60*, 162–163. https://doi.org/10.1007/s00056-014-0278-9

Huang, K., Rhee, D. J., Ger, R., Layman, R., Yang, J., Cardenas, C. E., & Court, L. E. (2021). Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. *Journal of Applied Clinical Medical Physics, 22*(5), 168–174. https://doi.org/10.1002/acm2.13207

Ileşan, R. R., Beyer, M., Kunz, C., & Thieringer, F. M. (2023). Comparison of Artificial Intelligence-based applications for Mandible Segmentation: From established Platforms to In-House-developed Software. *Bioengineering (Basel, Switzerland), 10*(5). https://doi.org/10.3390/bioengineering10050604

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., . . . Maier-Hein, K. (2025). *nnInteractive: Redefining 3D Promptable Segmentation*. arXiv. https://doi.org/10.48550/arXiv.2503.08373.

Jindanil, T., Marinho-Vieira, L. E., de-Azevedo-Vaz, S. L., & Jacobs, R. (2023). A unique artificial intelligence-based tool for automated CBCT segmentation of mandibular incisive canal. *Dento Maxillo Facial Radiology*, *52*(8), 20230321. https://doi.org/10.1259/dmfr.20230321.

Kot, W. Y., Au Yeung, S. Y., Leung, Y. Y., Leung, P. H., & Yang, W.-F. (2025). Evolution of deep learning tooth segmentation from CT/CBCT images: A systematic review and meta-analysis. *BMC Oral Health, 25*(1), 800. https://doi.org/10.1186/s12903-025-05984-6

Lei, W., Xu, W., Li, K., Zhang, X., & Zhang, S. (2025). Medlsam: Localize and segment anything model for 3D CT images. *Medical Image Analysis, 99*, Article 103370. https://doi.org/10.1016/j.media.2024.103370

Liu, P., Sun, Y. [Ying], Zhao, X., & Yan, Y. (2023). Deep learning algorithm performance in contouring head and neck organs at risk: A systematic review and single-arm meta-analysis. *Biomedical Engineering Online*, *22*(1), 104. https://doi.org/10.1186/s12938-023-01159-y.

Lübbers, H.-T., Bornstein, M. M., Dagassan-Berndt, D., Filippi, A., Suter, V. G. A., & Dula, K. (2024). Revised consensus guidelines for the use of cone-beam computed tomography/ digital volume tomography. *Swiss Dental Journal, 134*(4), 86–115. https://doi.org/10.61872/sdj-2024-04-07.

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications, 15*(1), 654. https://doi.org/10.1038/s41467-024-44824-z

Ma, J., Li, F., & Wang, B. (2024, January 9). *U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation*. Retrieved from http://arxiv.org/pdf/2401.04722.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., & Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications, 9*(1), 5217. https://doi.org/10.1038/s41467-018-07619-7

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., & Jäger, P. F. (2024). Metrics reloaded: Recommendations for image analysis validation. *Nature Methods, 21*(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). In *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation* (pp. 565–571). IEEE. https://doi.org/10.1109/3DV.2016.79.

Minnema, J., van Eijnatten, M., Hendriksen, A. A., Liberton, N., Pelt, D. M., Batenburg, K. J., & Wolff, J. (2019). Segmentation of dental cone-beam CT scans affected by metal artifacts using a mixed-scale dense convolutional neural network. *Medical Physics, 46*(11), 5027–5035. https://doi.org/10.1002/mp.13793

Naziroglu, R. E., van Ravesteijn, V. F., van Vliet, L. J., Streekstra, G. J., & Vos, F. M. (2017). Simulation of scanner- and patient-specific low-dose CT imaging from existing CT images. *Physica Medica : PM : An International Journal Devoted to the Applications of Physics to Medicine and Biology : Official Journal of the Italian Association of Biomedical Physics (AIFB)*, *36*, 12–23. https://doi.org/10.1016/j.ejmp.2017.02.009.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., de Fauw, J., & Ronneberger, O. (2021). Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research, 23*(7), Article e26151. https://doi.org/10.2196/26151

Pankert, T., Lee, H., Peters, F., Hölzle, F., Modabber, A., & Raith, S. (2023). Mandible segmentation from CT data for virtual surgical planning using an augmented two-stepped convolutional neural network. *International Journal of Computer Assisted Radiology and Surgery, 18*(8), 1479–1488. https://doi.org/10.1007/s11548-022-02830-w

Pesapane, F., Hauglid, M. K., Fumagalli, M., Petersson, L., Parkar, A. P., Cassano, E., & Horgan, D. (2024). The translation of in-house imaging AI research into a medical device ensuring ethical and regulatory integrity. *European Journal of Radiology, 182*, Article 111852. https://doi.org/10.1016/j.ejrad.2024.111852

Pianykh, O. S., Langs, G., Dewey, M., Enzmann, D. R., Herold, C. J., Schoenberg, S. O., & Brink, J. A. (2020). Continuous Learning AI in Radiology: Implementation Principles and Early applications. *Radiology, 297*(1), 6–14. https://doi.org/10.1148/radiol.2020200038

Puggelli, L., Uccheddu, F., Volpe, Y., Furferi, R., & Di Feo, D. (2019). Accuracy Assessment of CT-Based 3D Bone Surface Reconstruction. In F. Cavas-Martínez, B. Eynard, F. J. Fernández Cañavate, D. G. Fernández-Pacheco, P. Morer, & V. Nigrelli (Eds.), *Lecture Notes in Mechanical Engineering. Advances on mechanics, design engineering and manufacturing* (pp. 487–496). New York NY: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-030-12346-8_47.

Qiu, B., van der Wel, H., Kraeima, J., Glas, H. H., Guo, J., Borra, R. J. H., & van Ooijen, P. M. A. (2021). Automatic Segmentation of Mandible from conventional Methods to Deep Learning-a Review. *J. Pers. Med., 11*(7). https://doi.org/10.3390/jpm11070629

Rajamani, S. T., Rajamani, K., Venkateshvaran, A., Triantafyllopoulos, A., Kathan, A., & Schuller, B. W. (2023). Toward Detecting and addressing Corner cases in Deep Learning based Medical image Segmentation. *IEEE Access, 11*, 95334–95345. https://doi.org/10.1109/ACCESS.2023.3311134

Ranzini, M. B. M., Fidon, L., Ourselin, S. [Sébastien], Modat, M., & Vercauteren, T. (2021, March 21). *MONAIfbs: MONAI-based fetal brain MRI deep learning segmentation*. Retrieved from http://arxiv.org/pdf/2103.13314.

Reddy, S. (2022). Explainability and artificial intelligence in medicine. *The Lancet. Digital Health, 4*(4), e214–e215. https://doi.org/10.1016/S2589-7500(22)00029-2

Sarria, G. R., Kugel, F., Roehner, F., Layer, J., Dejonckheere, C., Scafa, D., & Schmeel, L. C. (2024). Artificial Intelligence-based Autosegmentation: Advantages in Delineation, absorbed Dose-distribution, and Logistics. *Advances in Radiation Oncology, 9*(3), Article 101394. https://doi.org/10.1016/j.adro.2023.101394

Scott, I. A., van der Vegt;, A., Lane, P., McPhail, S., & Magrabi, F. (2024). Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ Health & Care Informatics, 31*(1). https://doi.org/10.1136/bmjhci-2023-100971

Shin, J. (2024). Revolutionizing Medical Imaging with Artificial intelligence Real-Time Segmentation for Enhanced Diagnostics. *EDRAAK, 2024*, 18–25. https://doi.org/10.70470/edraak/2024/003.

Tappeiner, E., Welk, M., & Schubert, R. (2022). Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. *International Journal of Computer Assisted Radiology and Surgery, 17*(11), 2103–2111. https://doi.org/10.1007/s11548-022-02649-5

Vallières, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J. W. L., & El Naqa, I. (2017). Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports, 7*(1), 10117. https://doi.org/10.1038/s41598-017-10371-5

Vallières, M., Kay-Rivest, E., Perrin, L., Liem, X., Furstoss, C., Khaouam, N., & Sultanem, K. (2017). *Data from Head-Neck-PET-CT. The Cancer Imaging Archive*. https://doi.org/10.7937/K9/TCIA.2017.8oje5q00

Van Baar, G. J., Forouzanfar, T., Liberton, N. P., Winters, H. A., & Leusink, F. K. (2018). Accuracy of computer-assisted surgery in mandibular reconstruction: A systematic review. *Oral Oncology, 84*, 52–60. https://doi.org/10.1016/j.oraloncology.2018.07.004

Van de Sande, D., Chung, E. F. F., Oosterhoff, J., van Bommel, J., Gommers, D., & van Genderen, M. E. (2024). To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *npj Digital Medicine, 7*(1), 58. https://doi.org/10.1038/s41746-024-01064-1

Van Eijnatten, M., van Dijk, R., Dobbe, J., Streekstra, G., Koivisto, J., & Wolff, J. (2018). Ct image segmentation methods for bone used in medical additive manufacturing. *Medical Engineering & Physics, 51*, 6–16. https://doi.org/10.1016/j.medengphy.2017.10.008

Van Nistelrooij, N., Schitter, S., van Lierop, P., Ghoul, K. E., König, D., Hanisch, M., & Vinayahalingam, S. (2024). Detecting Mandible Fractures in CBCT Scans using a 3-Stage Neural Network. *Journal of Dental Research, 103*(13), 1384–1391. https://doi.org/10.1177/00220345241256618

Verhelst, P.-J., Smolders, A., Beznik, T., Meewis, J., Vandemeulebroucke, A., Shaheen, E., & Jacobs, R. (2021). Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography. *Journal of Dentistry, 114*, Article 103786. https://doi.org/10.1016/j.jdent.2021.103786

Walker, G. V., Awan, M., Tao, R., Koay, E. J., Boehling, N. S., Grant, J. D., & Fuller, C. D. (2014). Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiotherapy and Oncology : Journal of the European Society for Therapeutic Radiology and Oncology, 112*(3), 321–325. https://doi.org/10.1016/j.radonc.2014.08.028

Wee, L., & Dekker, A. (2019). *Data from HEAD-NECK-RADIOMICS-HN1. The Cancer Imaging Archive*. https://doi.org/10.7937/tcia.2019.8kap372n

Whetton, S., & Georgiou, A. (2010). Conceptual challenges for advancing the socio-technical underpinnings of health informatics. *The Open Medical Informatics Journal, 4*, 221–224. https://doi.org/10.2174/1874431101004010221

Xu, J., Liu, J. [Jiannan], Zhang, D., Zhou, Z., Zhang, C., & Chen, X. (2021). A 3D segmentation network of mandible from CT scan with combination of multiple convolutional modules and edge supervision in mandibular reconstruction. *Computers in Biology and Medicine*, *138*, 104925. https://doi.org/10.1016/j.compbiomed.2021.104925.

Yang, W.-F., & Su, Y.-X. (2021). Artificial intelligence-enabled automatic segmentation of skull CT facilitates computer-assisted craniomaxillofacial surgery. *Oral Oncology, 118*, Article 105360. https://doi.org/10.1016/j.oraloncology.2021.105360

Zachow, S. (2015). Computational Planning in Facial Surgery. *Facial Plastic Surgery: FPS, 31*(5), 446–462. https://doi.org/10.1055/s-0035-1564717

Acknowledgement

I am deeply grateful to Dr. Dr. Behrus Hinrichs-Puladi, my direct supervisor and senior author, for his generous support, clear structural advice, and commitment that enabled the successful implementation of this project.

I also thank Prof. Dr. Dr. Frank Hölzle, my doctoral supervisor, for the opportunity to complete a doctorate at the Department of Oral and Maxillofacial Surgery, and for the intensive support that ensured excellent working conditions. I further wish to thank Prof. Dr. Rainer Röhrig for the research infrastructure and teaching provided at the Institute of Medical Informatics, and for his support of this project.

My sincere appreciation goes to all co-authors for their essential contributions to this work.

Above all, I would like to thank my wife, Dr. Xiaoying Chen, for her unwavering support and encouragement over the past 13 years. Her presence during difficult times made them much easier to endure.

Affidavit according to § 5 (1) for Data Retention

Hiermit erkläre ich, dass die dieser Dissertation zu Grunde liegenden Originaldaten in der Institut für Medizinische Informatik, des Universitätsklinikums Aachen, hinterlegt sind.

Affidavit according to § 5 (1) and (2), and § 11 (3) 12 of the doctoral studies regulations

I, **Kunpeng Xie,** hereby declare on oath, that I have contributed a significant part, and thus majority, of the publication:
Xie K, Gruber LJ, Crampen M, Li Y, Ferreira A, Tappeiner E, Gillot M, Schepers J, Xu J, Pankert T, Beyer M, Shahamiri N, ten Brink R, Dot G, Weschke C, van Nistelrooij N, Verhelst PJ, Guo Y, Xu Z, Bienzeisler J, Rashad A, Flügge T, Cotton R, Vinayahalingam S, Ilesan R, Raith S, Madsen D, Seibold C, Xi T, Bergé S, Nebelung S, Kodym O, Sundqvist O, Thieringer F, Lamecker H, Coppens A, Potrusil T, Kraeima J, Witjes M, Wu G, Chen X, Lambrechts A, Soares Cevidanes LH, Zachow S, Hermans A, Truhn D, Alves V, Egger J, Röhrig R, Hölzle F, Hinrichs-Puladi B.: Beyond Benchmarks: Towards Robust Artificial Intelligence Bone Segmentation in Socio-Technical Systems; Expert Systems with Applications.; 2025, DOI: https://doi.org/10.1016/j.eswa.2025.130031
The contributions to the publication were as follows:

	Kunpeng Xie	Lennart Johannes Gruber	Marrtin Crampen	Yao Li	André Ferreira	Rest of the Authors	Jan Egger	Rainer Röhrig	Frank Hölzle	Behrus Hinrichs-Puladi	Sum (%)
Conceptualization	45									55	100
Methodology	70									30	100
Software	70		20							10	100
Validation				30	20					50	100
Formal analysis	55	1	1	1	1	13	1	1	1	25	100
Investigation	65	10	5							20	100
Resources						60	5	5	5	25	100
Data curation	70	5								25	100
Writing - original draft preparation	100										100
Writing - review and editing	20	1	1	1	1	13	1	1	1	60	100
Visualization	70									30	100
Supervision										100	100
Project administration										100	100
Funding										100	100

During the preparation of this work the authors used DeepL to polish the English language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.
The position as the first author obviously arises from this significant contribution.

_____
Signature of the doctoral candidate

As supervisor or corresponding author I confirm the statements of Kunpeng Xie and as a representative of the collaborative partners

_____ _____
Signature of the doctoral supervisor Signature of the corresp. Author

As co-author I endorse the statement of Univ.-Prof. Dr. Dr. Frank Hölzle

_____
V. Alves S. Bergé M. Beyer J. Bienzeisler R. ten Brink L. Soares Cevidanes X. Chen A. Coppens

_____
R. Cotton M. Crampen G. Dot J. Egger A. Ferreira T. Flügge M. Gillot L. Gruber

_____
Y. Guo A. Hermans R. Ilesan O. Kodym J. Kraeima H. Lamecker A. Lambrechts Y. Li

Affidavit according to § 5 (1) and (2), and § 11 (3) 12 of the doctoral studies regulations

I, **Kunpeng Xie,** hereby declare on oath, that I have contributed a significant part, and thus majority, of the publication:
Xie K, Gruber LJ, Crampen M, Li Y, Ferreira A, Tappeiner E, Gillot M, Schepers J, Xu J, Pankert T, Beyer M, Shahamiri N, ten Brink R, Dot G, Weschke C, van Nistelrooij N, Verhelst PJ, Guo Y, Xu Z, Bienzeisler J, Rashad A, Flügge T, Cotton R, Vinayahalingam S, Ilesan R, Raith S, Madsen D, Seibold C, Xi T, Bergé S, Nebelung S, Kodym O, Sundqvist O, Thieringer F, Lamecker H, Coppens A, Potrusil T, Kraeima J, Witjes M, Wu G, Chen X, Lambrechts A, Soares Cevidanes LH, Zachow S, Hermans A, Truhn D, Alves V, Egger J, Röhrig R, Hölzle F, Hinrichs-Puladi B.: Beyond Benchmarks: Towards Robust Artificial Intelligence Bone Segmentation in Socio-Technical Systems; Expert Systems with Applications.; 2025, DOI: https://doi.org/10.1016/j.eswa.2025.130031
The contributions to the publication were as follows:

	Kunpeng Xie	Lennart Johannes Gruber	Marrtin Crampen	Yao Li	André Ferreira	Rest of the Authors	Jan Egger	Rainer Röhrig	Frank Hölzle	Behrus Hinrichs-Puladi	Sum (%)
Conceptualization	45									55	100
Methodology	70									30	100
Software	70		20							10	100
Validation				30	20					50	100
Formal analysis	55	1	1	1	1	13	1	1	1	25	100
Investigation	65	10	5							20	100
Resources						60	5	5	5	25	100
Data curation	70	5								25	100
Writing - original draft preparation	100										100
Writing - review and editing	20	1	1	1	1	13	1	1	1	60	100
Visualization	70									30	100
Supervision										100	100
Project administration										100	100
Funding										100	100

During the preparation of this work the authors used DeepL to polish the English language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.
The position as the first author obviously arises from this significant contribution.

_____
Signature of the doctoral candidate

As supervisor or corresponding author I confirm the statements of Kunpeng Xie and as a representative of the collaborative partners

_____ _____
Signature of the doctoral supervisor Signature of the corresp. Author

As co-author I endorse the statement of Univ.-Prof. Dr. Dr. Frank Hölzle

_____
D. Madsen S. Nebelung N. van Nistelrooij T. Pankert T. Potrusil S. Raith A. Rashad R. Röhrig

_____
J. Schepers C. Seibold N. Shahamiri O. Sundqvist E. Tappeiner F. Thieringer D. Truhn P.-J. Verhelst

_____
S. Vinayahalingam C. Weschke M. Witjes G. Wu T. Xi J. Xu Z. Xu S. Zachow

CURRICULUM VITAE

Name: Kunpeng Xie

Date of Birth: 25.01.1995

Place of Birth: Changchun, China

Academic Background

2014 – 2019	Bachelor of Stomatology in Jilin University
2019 – 2022	Master of Stomatology in Jilin University (Department of Oral and Maxillofacial Surgery, Hospital of Stomatology, Jilin University)
Since 2022	Doctoral Student (Dr. med. dent) in RWTH Aachen University (Department of Oral and Maxillofacial Surgery & Institute of Medical Informatics)

Professional Background

2022 – 2025	Doctoral Student Full-time research in Department of Oral and Maxillofacial Surgery & Institute of Medical Informatics
2019 – 2022	Resident Surgeon Virtual Surgical Planning, 3D-Printing, Patient management and clinical research in Department of Oral and Maxillofacial Surgery, Hospital of Stomatology, Jilin University
2018 – 2019	Internship Clinical rotation in Hospital of Stomatology, Jilin University

Qualifications

2025	National Health Professional Technical Qualification (Intermediate) - Attending Surgeon of Oral and Maxillofacial Surgery, PR China
2020	Medical Practitioners Qualification of China

Aachen, 07.11.2025

Kunpeng Xie