Accuracy And Generalizability of an Open-Source Deep Learning Model For Facial Bone Segmentation on CT and CBCT Scans

Nikolaos Gkantidis[1]*, Mohammed Ghamri[1,2], Gauthier Dot[3,4]*

[1]Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, University of Bern, CH-3010 Bern, Switzerland;

[2] Jeddah Second Health Cluster, 23816 Jeddah, Saudi Arabia

[3] Université Paris Cité, Université Sorbonne Paris Nord, Inserm UMR 1333 Oral Heath, F-92120 Montrouge, France.

[4] Service Médecine Bucco-Dentaire, AP-HP, Hôpital Pitié-Salpêtrière, F-75013 Paris, France

* Corresponding authors : nikolaos.gkantidis@unibe.ch and gauthier.dot@aphp.fr

Abstract

Aim: To evaluate the accuracy and generalizability of DentalSegmentator, an open-source deep learning tool, for automated segmentation of skeletal facial surfaces from computed tomography (CT) scans acquired under different imaging conditions.

Materials and Methods: Ten human skulls were scanned using a CT scanner and three cone beam CT (CBCT) protocols (including an ultra-low-dose protocol) on two CBCT devices. High-accuracy reference surface models were acquired using an optical scanner. CBCT an CT scans were segmented automatically using DentalSegmentator. Three facial regions (forehead, zygomatic process, maxillary process) were defined on each model for quantitative assessment. Accuracy was measured as the mean absolute distance (MAD) and the standard deviation of absolute distances (SDAD) between segmented and reference models after best-fit superimposition.

Results: Repeated segmentations were identical, confirming perfect reproducibility. Across all acquisition settings and regions, DentalSegmentator produced highly accurate skeletal surface models, with an overall MAD of 0.088 mm (IQR 0.073) and SDAD of 0.061 mm (IQR 0.028). Significant but small differences were detected between imaging systems (MAD: $p < 0.001$; SDAD: $p = 0.003$), with CT scans showing slightly reduced trueness compared with CBCT images.

Conclusion: The open-source DentalSegmentator tool produced accurate skeletal facial surface segmentations across diverse CT and CBCT settings, demonstrating excellent generalizability, including under low-radiation conditions. Minor differences in trueness between imaging systems were small and unlikely to impact clinical or research use.

Clinical Significance: Deep learning offers a robust foundation for automated 3D craniofacial surface extraction, supporting broader adoption of AI-driven workflows in both clinical and research contexts.

Keywords: Cone-Beam Computed Tomography; Facial morphology; Three-Dimensional Imaging; Accuracy; Artificial Intelligence; Deep Learning; Automated segmentation.

Introduction

Digital dentistry and craniofacial surgery workflows now depend greatly on digital 3D models created from multi-slice computed tomography (CT) or cone-beam computed tomography (CBCT) scans. Models depicting skeletal facial surfaces are particularly important for computer-assisted surgical planning [1], virtual patient generation [2], or the superimposition of craniofacial structures over time [3]. After acquiring the scans, an important step in the workflow is accurately delineating anatomical structures through a process called segmentation. Over the past five years, there has been a significant shift in the segmentation methods used for dento-maxillo-facial (DMF) CT and CBCT scans. Although manual segmentation is still considered the reference standard, artificial intelligence (AI) - particularly deep learning–based methods - has demonstrated great potential for automating this process [4–7]. Notably, these automated methods are already available to clinicians through several open-source and commercial solutions for segmenting all anatomical parts of the DMF complex [2,8–10].

Given the surge in AI-based segmentation solutions, it is crucial to evaluate them with methodological rigor and adhere to guidelines such as DENTAL-AI for planning, conducting, and reporting studies [11]. A recent review of AI-based DMF segmentation summarized the situation: although research papers report very encouraging results that are on par with or even better than those of expert-based segmentation, methodological flaws in the included studies may affect the results [5]. High-quality studies addressing real clinical problems or advances in AI methodology are still needed [12].

The accuracy of AI-based segmentation refers to the trueness and the precision of the resulting 3D model [13]. It is typically measured by comparing automated segmentation (test) with expert-based segmentation (reference standard) [14]. This approach is justified by the fact that obtaining a hard ground truth, such as a skeletal facial surface, is impossible in a clinical environment. Therefore, reference standard annotation poses a significant challenge because it depends on manual annotations or manual corrections of automated annotations, which are susceptible to operator variability [11]. To the best of our knowledge, no previous study has evaluated the accuracy of AI-based DMF segmentation using hard ground truth. The only feasible methodological approach to accurately depict the actual anatomy of a subject is an ex vivo study design on dry cadaveric specimens [15]. The current best practice for obtaining ground truth data is to scan hydrated specimens with a high-accuracy optical surface scanner [16]. In order to obtain CT or CBCT imaging that is comparable to that of actual patients, the dry skeletal specimens must be embedded in soft tissue simulants during acquisition [17].

The generalizability of an AI-based segmentation model refers to its ability to perform accurately on new data, including data obtained using previously unseen equipment, acquisition protocols, or resolutions [18]. Ideally, it is evaluated using an external test set made of independent data acquired from other institutions [19]. Although test sets have increased in recently published studies, few studies have examined how equipment and acquisition parameters affect automated segmentation results [20]. The effect of low-radiation protocols on automated segmentation accuracy, in particular, has yet to be clarified [21].

The clinical acceptability of AI-based segmentation depends on the use of clinically relevant and clinically significant metrics to evaluate model performance [12]. The majority of studies on DMF segmentation report the Dice coefficient as the main evaluation metric [1,5]. Although it is useful for comparing models, this metric has limited meaning for clinicians and may not always be clinically relevant [14]. Though normalized surface distance has been proposed as a clinically relevant

alternative, it is still derived from engineering and may lack clinical significance [22]. Other methods are often used in clinical studies. For instance, the mean distance between 3D model surfaces or colour-coded distance maps are commonly used to evaluate the results of 3D superimposition [23,24]. The use of these methods in evaluating AI-based segmentation has yet to be investigated. We believe that a methodological study applying such clinically meaningful methodologies will strengthen the foundation of deep learning-based methods [12].

Therefore, the present ex vivo study aimed to evaluate the accuracy and generalizability of DentalSegmentator, an open-source, AI-based segmentation method [8], in reconstructing 3D skeletal surface models of the face. Accuracy was assessed by comparing segmentation results with direct surface scans that served as ground truth. The generalizability of the AI model was evaluated by including various CT and CBCT scanners and acquisition protocols, including an ultra-low-dose protocol. Clinically significant measures and distance maps were used to ensure clinical acceptability of the results. It was assumed that there were no statistically significant differences between the segmentation results and the ground truth for all scanners and protocols.

Materials and Methods

Ethical approval and informed consent

Ethical approval for this study was obtained from the Research Ethics Committee of the Dental School, National and Kapodistrian University of Athens (Protocol No. 335; initial approval 02/05/2017, renewed 16/11/2021). All procedures adhered to applicable guidelines and regulatory standards. Because the specimens were anonymized and their donors could not be identified, the committee issued a waiver for informed consent. Handling of all human tissues complied with the corresponding local legal requirements.

Sample

This study utilized ten human dry skulls obtained from the Municipal Cemetery of Serres, Greece, with formal approval from the local authorities (Municipality of Serres, Protocol No. 4044/12.07.2018). The specimens belonged to unidentified individuals who had died 8 to 12 years prior to sample collection. At the time of retrieval, no claims had been made by relatives, and the skulls had been buried collectively according to cemetery protocol. Each skull was visually inspected by two evaluators to confirm that it represented a fully developed adult cranium with no major defects, pathological conditions, or significant age related deterioration.

Data acquisition

<u>Optical reference models</u>

All ten specimens were scanned using a structured-light 3D surface scanner (Artec Space Spider, Artec3D, Luxembourg; Software: Artec Studio 12, Version 12.1.6.16) to acquire high-resolution reference surface models that comprised the ground truth models for the study. Prior to scanning session, the optical scanner was calibrated according to the manufacturer's instructions. To ensure dimensional stability between optical and tomographic scanning, all skulls were hydrated by immersion in tap water for 15 minutes before scanning. Scanning was performed immediately after removal from water, following gentle patting with tissue paper. The water temperature ranged

between 22–25 °C, with an approximate pH of 7.5. The optical scanner precision [25], as well as the effect of hydration on the anatomical form of facial skeletal bones [16] has been thoroughly evaluated previously on the same specimens and the subsequent surface models have been used for the present study. Raw optical scan data were processed in Artec Studio 16 (Version 16.0.5.114). Partial scans were merged into complete skull models through automated alignment or a combination of initial manual landmark approximation followed by automated registration. The resulting ten reference models were imported into Viewbox 4 software (dHAL Software, Kifisia, Greece) for subsequent comparisons with the tomographic surface based models.

Tomographic Data

The tomographic scans included in this study were performed on the same ten skulls. These were previously used in publications evaluating their trueness [15,26]. Soft tissue simulation was achieved using a 3D-printed head shell filled with water (PETG, MasterFill Premium PETG Pro, 3DHUB, Greece), aiming to similar hydration conditions to those during optical scanning. As previously published [26], each skull was scanned using four tomographic protocols, generating ten volumes per acquisition setting (total: 40 volumes):

1. CT scanner (Revolution CT 256, GE Healthcare, USA; 251 Hellenic Airforce Hospital, Athens, Greece). Tube voltage: 120 kV, tube current: 490 mA in the area of interest (automatically configured based on tissue mass and density), exposure time: 1 s, slice thickness: 0.625 mm, voxel size: 0.49 to 0.62 × 0.49 to 0.62 × 0.31 (interslice) mm, FOV: full head (displayed FOV: 25 cm).

2. CBCT scanner I (Newtom VGiMK4, Verona, Italy; Dental School, National and Kapodistrian University of Athens Greece). Tube voltage: 110 kV, tube current: 4–5 mA (automatically configured based on tissue mass and density), exposure time: 4 s, voxel size: 0.3 × 0.3 × 0.3 mm, FOV: ⌀15 × 15 cm.

3. CBCT scanner II—regular dose settings (Planmeca F, Planmeca Promax 3D Mid 2018, Helsinki, Finland; Digital Iatriki Apeikonisi, Athens, Greece). Tube voltage: 100 kV, tube current: 8 mA, exposure time: 12 s, voxel size: 0.2 × 0.2 × 0.2 mm, FOV: ⌀20 × 17 cm.

4. CBCT scanner II—ultra-low dose settings (Planmeca U). Tube voltage: 100 kV, tube current: 8 mA, exposure time: 6s, voxel size: 0.2 × 0.2 × 0.2 mm, FOV: ⌀20 × 17 cm.

All tomographic DICOM datasets were converted in Neuroimaging Informatics Technology Initiative (NIfTI) files using ITK-Snap software (v. 4.0.1, http://www.itksnap.org/) [27]. Those NIfTI files were automatically segmented by one operator (GD) using the publicly available nnU-Net DentalSegmentator weights with no additional pre- or post-processing [8,28]. The resulting segmentation masks were then imported as NifTI files into Viewbox 4 and automatically reconstructed as 3D surface models (STL) using the software's built-in isosurface reconstruction (marching-cubes) algorithm.

Reproducibility assessment

To evaluate reproducibility, all four tomographic images from two randomly selected skulls were segmented twice using the DentalSegmentator software. Given the fully automated nature of the segmentation process, identical results were expected across repeated runs. In the event of discrepancies, a detailed reproducibility analysis would be conducted. Colour-coded distance maps

between best-fit approximated, repeatedly segmented surface models were generated to assess this outcome.

Surface Model Superimposition and Trueness Assessment

A best-fit registration procedure was performed to superimpose each CBCT/CT derived 3D model onto its corresponding optical reference model. A variant of the iterative closest point (ICP) algorithm [29] was used with the following parameters: 100% estimated overlap of meshes, matching point to plane, exact nearest neighbour search, 100% point sampling, 50 iterations. The reference areas used for superimposition and the predefined measurement regions are illustrated in Figure 1.

Measurement areas were consistently placed on the ground truth models and included bilateral areas on the forehead, zygomatic processes, and maxillary complex (Figure 1). Each single selection covered an area of 60 mm^2 and consisted of approximately 900 vertices. Bilateral areas were treated as single measurement units.

Trueness of the DentalSegmentator derived surface models was quantified by calculating the congruence between the superimposed models using Mean Absolute Distance (MAD) and Standard Deviation of Absolute Distances (SDAD). Both metrics were calculated for each of the three predefined measurement regions.

Colour-coded distance maps were generated to visually illustrate deviations over the entire surface model. Perfect trueness corresponded to regions showing zero deviations.

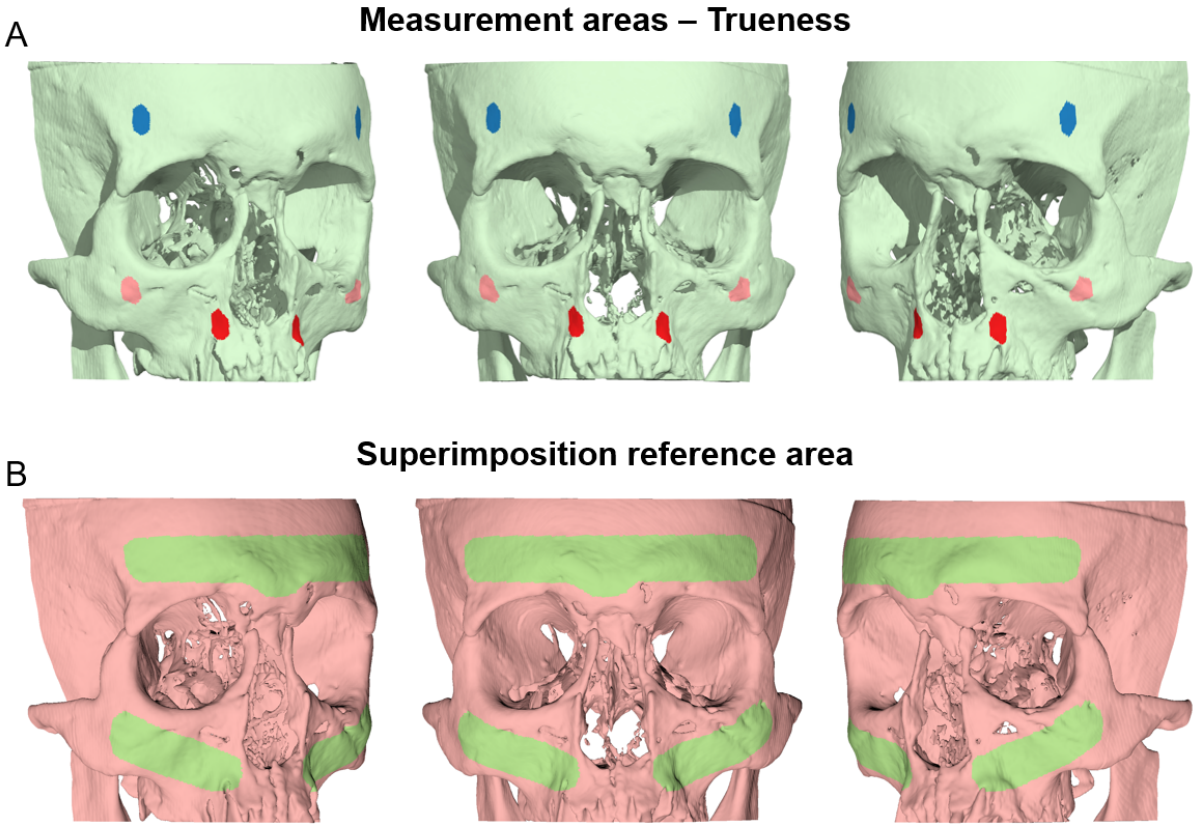

Figure 1. A. Three measurement areas defined in each skull through bilaterally selected mesh surfaces at the forehead (blue), the zygomatic process (pink), and the maxillary process (red) for trueness assessment, respectively. Each circular area shown in the images consists of 1,000 triangles. B. Reference area used for all surface-based superimpositions performed in the study (green).

Statistical analysis

The IBM SPSS statistics for Windows (Version 29.0.IBM Corp., Armonk, NY, USA) was used for descriptive and comparative statistical analyses and associated graph generation. The significance level was set at 0.05. Data normality was assessed using the Shapiro–Wilk test and visual inspection of normality plots, revealing significant deviations for some variables. Therefore, non-parametric statistical methods were used.

Differences in the trueness of segmented facial surface models across the four acquisition settings were assessed using the Kruskal–Wallis test. If significant, pairwise comparisons were conducted with the Dunn's test, applying Bonferroni correction. Analyses were performed both for the combined measurement areas and for each area individually.

Results

Reproducibility

All repeatedly segmented facial surface models from the tomographic scans (n=8: 4 acquisition settings x 2 skulls) through DentalSegmentator were identical to the original segmentations (Figure 2). Therefore, no further reproducibility assessment was performed.

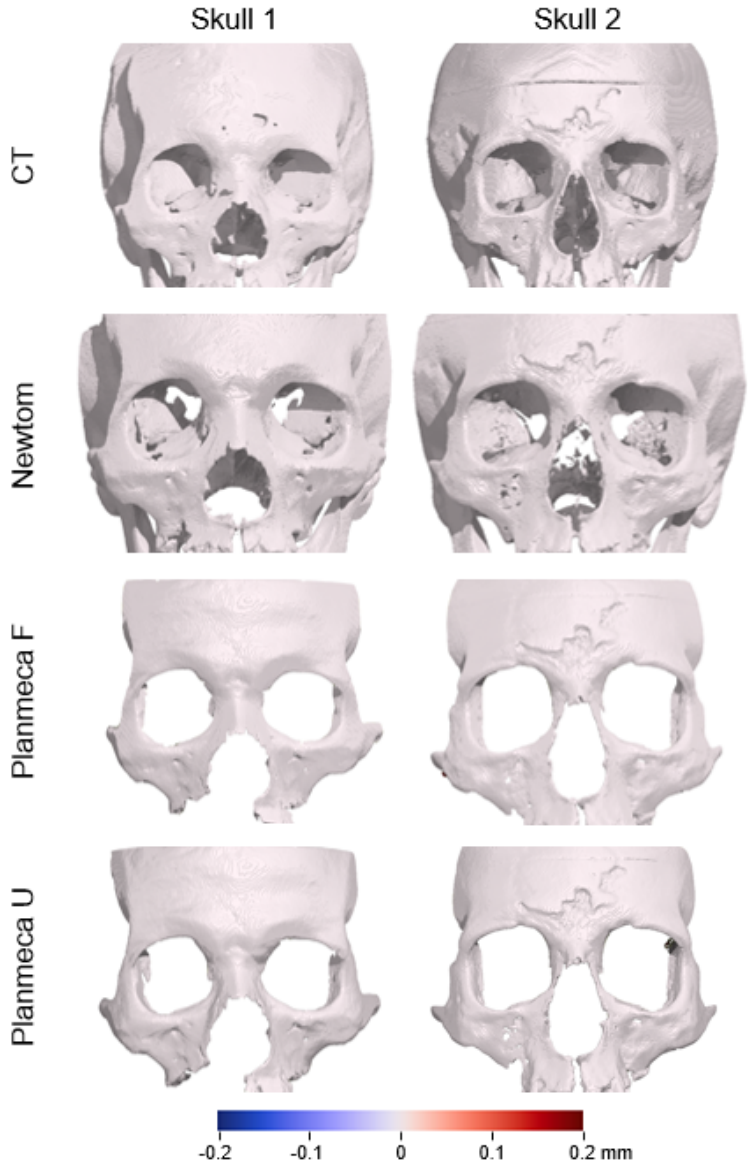

Figure 2. Colour-coded distance maps between repeatedly segmented surface models by DentalSegmentator, representative of the minimum, average, and maximum differences detected in the sample for each acquisition setting. The compared models retained their original spatial relation within the source tomographic volume (status before superimposition).

Trueness

Considering all acquisition settings and measurement areas, the overall trueness (MAD) of DentalSegmentator derived skeletal facial surfaces was 0.088 mm (IQR: 0.073), with an SDAD of 0.061 mm (IQR: 0.028). There were significant differences between acquisition settings in MAD values depicting trueness (Kruskal-Wallis Test: $p<0.001$), as well as in the respective SDAD values (Kruskal-Wallis Test: $p=0.003$), although these were of limited magnitude (Figure 3). The segmentation of CT scans provided slightly less accurate outcomes than those of the Planmeca scans. Detailed descriptive statistics and pairwise comparison test are provided in Table 1.

Table 1. Trueness of the DentalSegmentator facial skeletal surface models indicated by MAD and SDAD values per tomographic acquisition setting and associated pairwise comparisons.

	MAD		SDAD	
	Median	*IQR*	*Median*	*IQR*
CT	0.148	0.169	0.095	0.140
Newtom	0.101	0.065	0.070	0.019
Planmeca F	0.090	0.053	0.067	0.045
Planmeca U	0.087	0.046	0.067	0.036
	Pairwise Comparisons		*Pairwise Comparisons*	
	Sig.	*Adj. Sig.*[a]	*Sig.*	*Adj. Sig.*[a]
Planmeca F vs. Planmeca U	0.778	1.000	0.973	1.000
Planmeca F vs. Newtom	0.060	0.359	0.367	1.000
Planmeca F vs. CT	<0.001	0.001	0.001	0.008
Planmeca U vs. Newtom	0.110	0.658	0.385	1.000
Planmeca U vs. CT	<0.001	0.003	0.002	0.009
Newtom vs. CT	0.062	0.375	0.021	0.127

Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

When each measurement area was analysed as a single variable, there were only minor and very few statistically significant differences in trueness (MAD of the best-fit approximated models) between acquisition settings (Forehead, Newtom vs Planmeca F: p=0.026; Zygoma, CT vs Planmeca U: p=0.015). Similarly, for the SDAD there were no pairwise differences between acquisition settings although the Kruskal-Wallis test was significant for the Forehead area (p=0.024) (Figure 3).

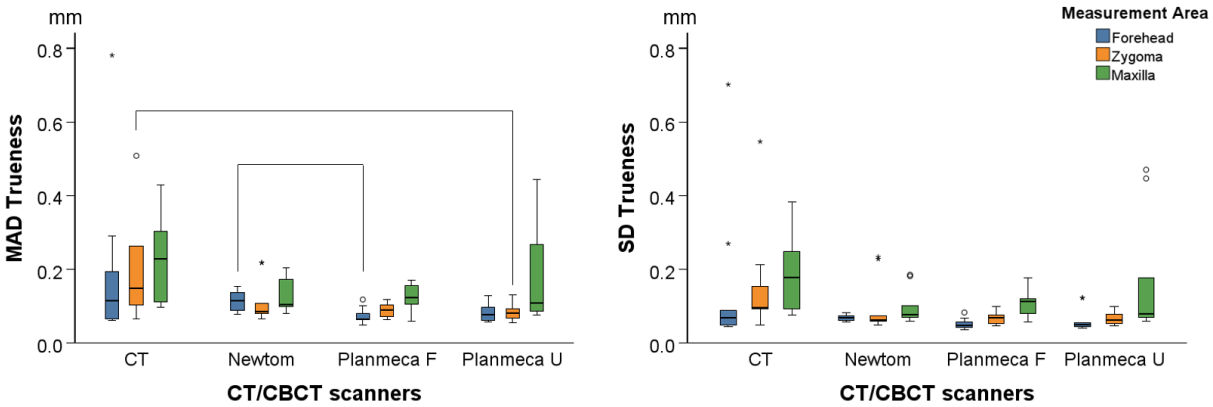

Figure 3. Box plots showing the trueness of the segmented facial surface models indicated by the distances of the tomographically derived models from the direct optical scans, following their best-fit approximation. The Mean Absolute Distances (MAD) and the standard deviations of the absolute distances (SD) between the superimposed surface models are shown. The lines connect variables that show significant differences (p<0.05) detected through Kruskal-Wallis, followed by Mann-Whitney U test (Bonferroni adjusted). Outliers are shown as black circles (further from the median more than 1.5 times the IQR).

The colour-coded distance maps between the best-fit approximated segmented and directly scanned facial surface models revealed consistent outcomes within and between acquisition settings, with deviations from the true model only locally reaching a maximum of 0.5 mm (Figure 4). Large deviations were often present at the sides of the skull, located distant to the used superimposition reference area.

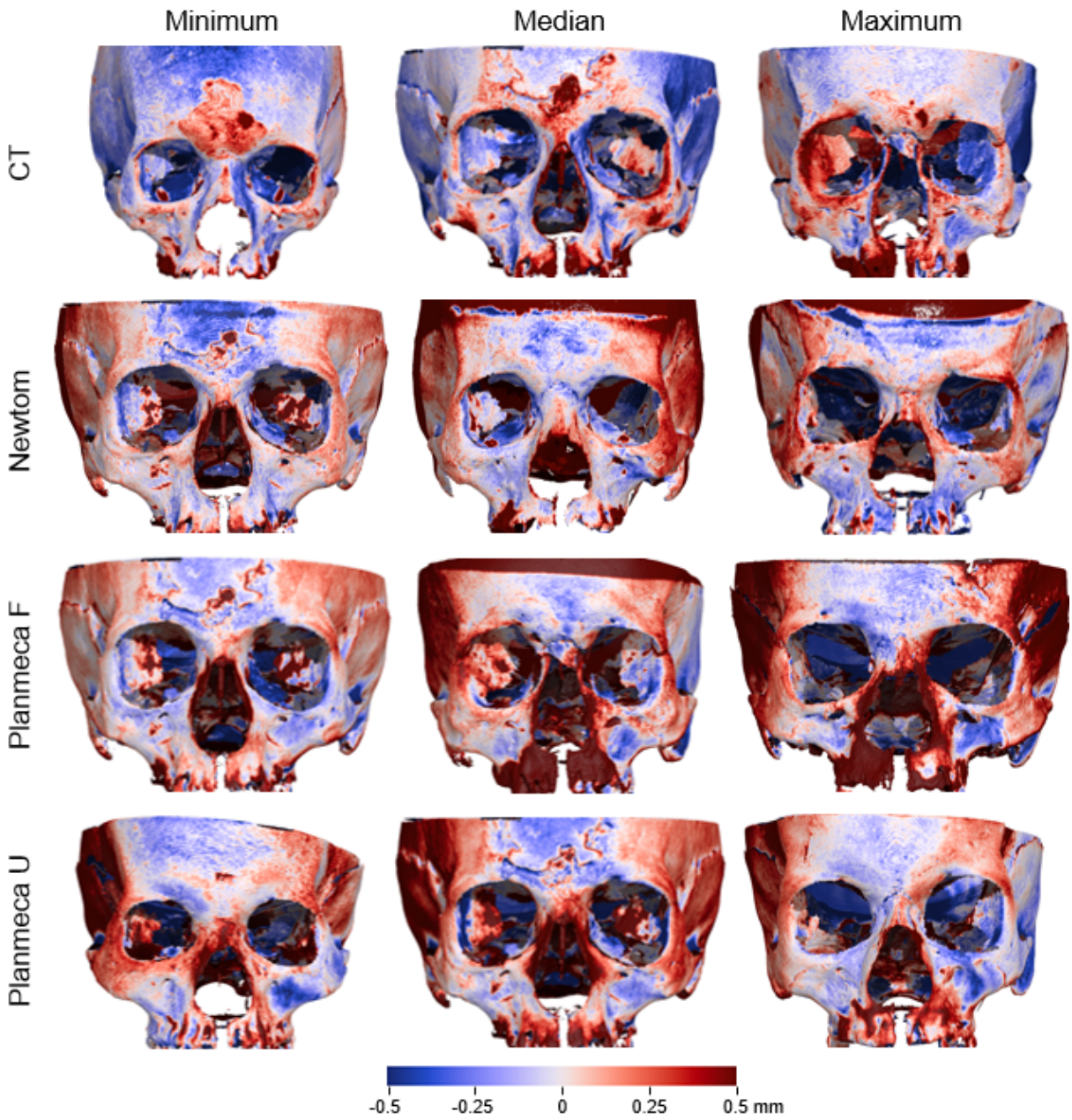

Figure 4. Colour-coded distance maps between best-fit approximated segmented surface models and directly obtained models through an optical surface scanner, representative of the minimum, average, and maximum deviations in trueness, detected in the sample for each acquisition setting.

Discussion

This ex vivo study evaluated the accuracy and generalizability of DentalSegmentator, an open-source deep learning model for automatic segmentation of skeletal facial surfaces from CT and CBCT scans

[8]. Overall, the tool produced highly accurate 3D facial bone models, with median mean absolute deviations (MAD) below 0.1 mm and standard deviations of absolute distances (SDAD) around 0.06 mm when compared with high-precision optical surface scans [25] used as ground truth. Although differences among tomographic scanners and acquisition settings, including the ultra-low-dose CBCT protocol, occasionally reached statistical significance, their magnitude was minimal and unlikely to be clinically meaningful. Colour-coded distance maps confirmed that local deviations rarely exceeded 0.5 mm and were mostly confined to lateral regions away from the superimposition reference area. These findings suggest that, for facial skeletal surfaces, DentalSegmentator is not a major source of geometric error and can be considered suitable for demanding clinical and research applications.

The present results are consistent with, and extend, previous work on the accuracy of tomographically derived facial surface models. In a closely related ex vivo study using the same skulls and a visually defined single-threshold segmentation approach, the overall trueness of CT- and CBCT-derived facial surfaces was approximately 0.12 mm, with local deviations again reaching up to 0.5 mm and no relevant differences between scanners or protocols [15,26]. In the current investigation, the overall MAD of the DentalSegmentator-derived surfaces (0.088 mm) was slightly lower than those previously reported for manual or semi-automatic threshold-based segmentations of the same tomographic data, while the spatial pattern and magnitude of local deviations remained similar. This indicates that, under realistic imaging conditions, the limiting factor for geometric accuracy is primarily the tomographic acquisition and not the deep learning-based segmentation itself.

These observations align well with the original multicenter evaluation of DentalSegmentator on clinical CT and CBCT datasets, where the model achieved high Dice similarity coefficients and normalized surface distances for multiple DMF structures [8]. In that work, segmentation quality was judged primarily through voxel-level overlap metrics against expert segmentations. By contrast, the present ex vivo study uses an independent physical ground truth and clinically interpretable surface distances, thereby complementing and strengthening the evidence base for the model. Together, these data support the notion that DentalSegmentator provides robust and geometrically accurate segmentations across a variety of scanners and protocols, at least for the facial skeleton.

From a generalizability perspective, the inclusion of one CT and two CBCT units, as well as a reduced-dose CBCT acquisition, is noteworthy. External benchmarks in medical image segmentation, such as the Medical Segmentation Decathlon and more recent large-scale frameworks, have demonstrated that robust models must cope with multi-center, multi-device, and protocol heterogeneity to perform reliably in real-world settings [30]. The impact of scanner type and acquisition parameters on the results of DMF AI segmentation has recently been quantified in a large-scale study demonstrating that, for a given model, segmentation accuracy can vary by up to 25% in Dice score depending on the scans characteristics [20]. The present findings indicate that DentalSegmentator maintains sub-0.15 mm median MAD across all tested settings, with the lowest accuracy associated with the CT scanner and slightly better outcomes for the CBCT devices, including the ultra-low-dose configuration. This pattern mirrors earlier ex vivo work showing that ultra-low-dose CBCT protocols can preserve the trueness of facial surface models at levels comparable to regular-dose protocols and CT, thereby supporting further dose reduction in clinical practice [15,26].

The use of an ex vivo experimental design on dry skulls embedded in soft-tissue simulants is an important methodological strength [16]. This approach is currently regarded as the most rigorous way to obtain a ground truth for skeletal surface geometry in radiology research, because it decouples

segmentation error from imaging limitations and allows direct comparison with high-precision surface scans [16,25]. In the present study, ground-truth models were acquired with a structured-light scanner whose precision has been shown to be on the order of tens of micrometres and whose model-generation error is negligible [25]. Hydration of the skulls and embedding in water-filled head shells followed previously validated protocols that mimic soft-tissue attenuation while inducing only limited geometric changes in the skeletal anatomy [16]. This combination of high-fidelity reference data and clinically realistic imaging conditions provides a solid basis for evaluating AI segmentation accuracy.

Another key contribution of this study lies in the choice of the evaluation metrics. Most AI-based segmentation studies in DMF imaging rely on volumetric overlap measures such as Dice or Jaccard coefficients, often complemented by global surface distances such as normalized surface distance, which are useful for benchmarking algorithms but can be difficult to interpret clinically [14,31]. Recent methodological work in medical imaging has emphasized that metric choice should reflect the clinical task and that overreliance on single, global metrics can obscure clinically relevant failure cases. In line with recommendations for AI in dental research, which call for transparent reporting and task-specific performance measures [11], the present study adopts local mean absolute distances and colour-coded surface maps as primary outcomes. These metrics are already widely used to assess craniofacial superimposition and 3D treatment changes, and their magnitude can be directly related to clinically acceptable thresholds for diagnosis, treatment planning, or 3D printing [32,33].

From a clinical standpoint, the observed level of error appears acceptable for most applications involving facial skeletal surfaces. Previous accuracy studies on CT- and CBCT-derived craniofacial models have reported surface deviations in the range of 0.1–0.2 mm for well-chosen acquisition protocols, with local deviations occasionally reaching 0.5 mm in areas of thin bone or at the periphery of the field of view [15,26]. The present findings fall within or slightly below this range, despite the fully automatic nature of the segmentation process. Such accuracy is likely sufficient for virtual treatment planning, orthognathic surgery simulations, assessment of facial morphology, 3D printing CAD-CAM applications and longitudinal 3D superimposition, where sub-millimetre discrepancies are typically considered negligible relative to biological variation and treatment effects [34–36]. However, for high accuracy-sensitive procedures, such as the design of cutting guides or patient-specific implants in regions of very thin cortical bone, careful visual inspection of the segmentation remains advisable.

The present study also complements recent work investigating the performance of DentalSegmentator and other deep learning tools for tooth segmentation in challenging clinical scenarios. In a study focusing on patients with multiple impacted teeth, DentalSegmentator and a commercial cloud-based solution achieved high normalized surface distances and low failure rates, outperforming two other commercial tools in several metrics [9]. Together with the current results on skeletal surfaces, these findings support the broader use of open-source deep learning models as reliable building blocks for constructing full virtual patients and for integrating AI into advanced orthodontic and orthognathic workflows.

Despite its strengths, this study has limitations that should be acknowledged. First, the sample size was limited to ten adult skulls, which restricts the variability of craniofacial morphology, bone density, and potential artefacts. While this is acceptable for a methodological validation study, larger and more heterogeneous datasets, ideally including pathological cases and different age groups, are needed to further probe model robustness. Second, the ex vivo design, although optimal for ground-truth acquisition, does not capture soft-tissue motion or patient positioning variability that are common in

clinical imaging [37]. Third, because only one deep learning model and restricted set tomographic scanners and acquisition protocols were assessed, the results may not be fully generalizable to all available systems or settings. Finally, this study thoroughly evaluated facial skeletal surface segmentation results but did not evaluate other anatomical parts such as anterior cranial base, mandible or teeth.

Future research should address these limitations by extending the evaluation to in vivo datasets with rigorously curated reference segmentations and by incorporating a broader range of scanners, voxel resolutions, and artefact conditions. In addition, comparative analyses involving diverse AI architectures, including medically adapted variants of foundation models or "segment-anything" frameworks [38], may help clarify their potential advantages for CBCT segmentation tasks. The integration of uncertainty estimation and automated quality-control procedures could further assist in identifying cases in which model outputs may be unreliable and therefore warrant manual verification. Finally, linking segmentation accuracy to downstream clinical tasks—for example, quantification of skeletal changes, prediction of surgical outcomes, or fabrication of guides and splints—would provide further assessment of clinical utility, consistent with recent calls to move beyond purely technical evaluations [19].

Conclusions

This ex vivo investigation shows that DentalSegmentator can accurately reconstruct skeletal facial surface models from ex vivo CT and CBCT scans acquired under realistic clinical conditions, with errors that are small in magnitude and comparable to, or lower than, those introduced by the imaging process itself. By combining rigorous ground-truth methodology, multi-scanner testing, and clinically meaningful evaluation metrics, the study adds an important layer of evidence supporting the deployment of open-source deep learning segmentation tools in craniofacial imaging and digital dentistry.

Funding: This research was funded by the Swiss Dental Association through a research grant (Protocol Number. 335-21) awarded to author N. Gkantidis and by the European Orthodontic Society, in the context of the W J B Houston Scholarship Award that was granted to author N. Gkantidis.

Declaration of Interest statement : The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Data: All data produced in the present study are available upon reasonable request to the authors

References

[1] J. Minnema, A. Ernst, M. Van Eijnatten, R. Pauwels, T. Forouzanfar, K.J. Batenburg, J. Wolff, A review on the application of deep learning for CT reconstruction, bone segmentation and surgical planning in oral and maxillofacial surgery, Dentomaxillofacial Radiology 51 (2022) 20210437. https://doi.org/10.1259/dmfr.20210437.

[2] F. Nogueira-Reis, N. Morgan, I.R. Suryani, C.P.M. Tabchoury, R. Jacobs, Full virtual patient generated by artificial intelligence-driven integrated segmentation of craniomaxillofacial structures from CBCT images, J Dent 141 (2024) 104829. https://doi.org/10.1016/j.jdent.2023.104829.

[3] L. Friedli, D. Kloukos, G. Kanavakis, D. Halazonetis, N. Gkantidis, The effect of threshold level on bone segmentation of cranial base structures from CT and CBCT images, Sci Rep 10 (2020) 7361. https://doi.org/10.1038/s41598-020-64383-9.

[4] B. Xiang, J. Lu, J. Yu, Evaluating tooth segmentation accuracy and time efficiency in CBCT images using artificial intelligence: A systematic review and Meta-analysis, J Dent 146 (2024) 105064. https://doi.org/10.1016/j.jdent.2024.105064.

[5] L. Zhang, W. Li, J. Lv, J. Xu, H. Zhou, G. Li, K. Ai, Advancements in oral and maxillofacial surgery medical images segmentation techniques: An overview, Journal of Dentistry 138 (2023) 104727. https://doi.org/10.1016/j.jdent.2023.104727.

[6] J.H. Park, M. Hamimi, J.J.E. Choi, C.M.S. Figueredo, A.B. Cameron, Comparisons of artificial intelligence automated segmentation techniques to manual segmentation techniques of the maxilla and maxillary sinus for CT or cone-beam CT scans-a systematic review, Dentomaxillofac Radiol 54 (2025) 529–539. https://doi.org/10.1093/dmfr/twaf042.

[7] S.R. Motamedian, S. Mohaghegh, M. Niazmand, H. Mohammad-Rahimi, N. Ahmadi, M. Yaseri, H.S.H. Boroojeni, Application of Artificial Intelligence in Orthognathic Surgery: A Scoping Review, Biomed Res Int 2025 (2025) 8284581. https://doi.org/10.1155/bmri/8284581.

[8] G. Dot, A. Chaurasia, G. Dubois, C. Savoldelli, S. Haghighat, S. Azimian, A.R. Taramsari, G. Sivaramakrishnan, J. Issa, A. Dubey, T. Schouman, L. Gajny, DentalSegmentator: Robust open source deep learning-based CT and CBCT image segmentation, Journal of Dentistry 147 (2024) 105130. https://doi.org/10.1016/j.jdent.2024.105130.

[9] E. Sinard, L. Gajny, M. de La Dure-Molla, R. Felizardo, G. Dot, Automated Cone Beam Computed Tomography Segmentation of Multiple Impacted Teeth With or Without Association to Rare Diseases: Evaluation of Four Deep Learning-Based Methods, Orthod Craniofac Res 28 (2025) 433–440. https://doi.org/10.1111/ocr.12890.

[10] M. Gillot, B. Baquero, C. Le, R. Deleat-Besson, J. Bianchi, A. Ruellas, M. Gurgel, M. Yatabe, N. Al Turkestani, K. Najarian, R. Soroushmehr, S. Pieper, R. Kikinis, B. Paniagua, J. Gryak, M. Ioshida, C. Massaro, L. Gomes, H. Oh, K. Evangelista, C.M. Chaves Junior, D. Garib, F. Costa, E. Benavides, F. Soki, J.-C. Fillion-Robin, H. Joshi, L. Cevidanes, J.C. Prieto, Automatic multi-anatomical skull structure segmentation of cone-beam computed tomography scans using 3D UNETR, PLoS One 17 (2022) e0275033. https://doi.org/10.1371/journal.pone.0275033.

[11] F. Schwendicke, T. Singh, J.-H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, S. Uribe, J. Krois, Artificial intelligence in dental research: Checklist for authors, reviewers, readers, Journal of Dentistry 107 (2021) 103610. https://doi.org/10.1016/j.jdent.2021.103610.

[12] S.E. Uribe, M.H. Hamdan, N.A. Valente, S. Yamaguchi, F. Umer, A. Tichy, R. Pauwels, F. Schwendicke, Evaluating dental AI research papers: Key considerations for editors and reviewers, J Dent 160 (2025) 105867. https://doi.org/10.1016/j.jdent.2025.105867.

[13] ISO 5725-2:2019, Accuracy (trueness and precision) of measurement methods and results, n.d.

[14] L. Maier-Hein, A. Reinke, P. Godau, M.D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M.A. Riegler, M. Wiesenfarth, A.E. Kavur, C.H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M.B. Blaschko, M.J. Cardoso, V. Cheplygina, B.A. Cimini, G.S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D.A. Hashimoto, M.M. Hoffman, M. Huisman, P. Jannin, C.E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C.I. Sánchez, S. Shetty, M. van Smeden, R.M. Summers, A.A. Taha, A. Tiulpin, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, P.F. Jäger, Metrics reloaded: recommendations for image analysis validation, Nat Methods 21 (2024) 195–212. https://doi.org/10.1038/s41592-023-02151-z.

[15] M. Ghamri, K. Dritsas, J. Probst, M. Jäggi, S. Psomiadis, R. Schulze, C. Verna, C. Katsaros, D. Halazonetis, N. Gkantidis, Accuracy of facial skeletal surfaces segmented from CT and CBCT radiographs, Sci Rep 13 (2023) 21002. https://doi.org/10.1038/s41598-023-48320-0.

[16] K. Dritsas, J. Probst, Y. Ren, C. Verna, C. Katsaros, D. Halazonetis, N. Gkantidis, Effect of hydration on the anatomical form of human dry skulls, Sci Rep 12 (2022) 22549. https://doi.org/10.1038/s41598-022-27042-9.

[17] P.A. Lopes, G.M. Santaella, C.A.S. Lima, K. de F. Vasconcelos, F.C. Groppo, Evaluation of soft tissues simulant materials in cone beam computed tomography, Dentomaxillofac Radiol 48 (2019) 20180072. https://doi.org/10.1259/dmfr.20180072.

[18] X. Liang, D. Nguyen, S.B. Jiang, Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion, Mach. Learn.: Sci. Technol. 2 (2020) 015007. https://doi.org/10.1088/2632-2153/abb214.

[19] F. Schwendicke, W. Samek, J. Krois, Artificial Intelligence in Dentistry: Chances and Challenges, J Dent Res 99 (2020) 769–774. https://doi.org/10.1177/0022034520915714.

[20] K. Xie, L. Johannes Gruber, M. Crampen, Y. Li, A. Ferreira, E. Tappeiner, M. Gillot, J. Schepers, J. Xu, T. Pankert, M. Beyer, N. Shahamiri, R. ten Brink, G. Dot, C. Weschke, N. van Nistelrooij, P.-J. Verhelst, Y. Guo, Z. Xu, J. Bienzeisler, A. Rashad, T. Flügge, R. Cotton, S. Vinayahalingam, R. Ilesan, S. Raith, D. Madsen, C. Seibold, T. Xi, S. Bergé, S. Nebelung, O. Kodym, O. Sundqvist, F.M. Thieringer, H. Lamecker, A. Coppens, T. Potrusil, J. Kraeima, M.

Witjes, G. Wu, X. Chen, A. Lambrechts, L.H. Soares Cevidanes, S. Zachow, A. Hermans, D. Truhn, V. Alves, J. Egger, R. Röhrig, F. Hölzle, B. Hinrichs-Puladi, Beyond benchmarks: Towards robust artificial intelligence bone segmentation in socio-technical systems, Expert Systems with Applications 299 (2026) 130031. https://doi.org/10.1016/j.eswa.2025.130031.

[21] M. Ruetters, H. Gehrig, S. Awounvo, T.-S. Kim, S. Doll, K. Alexandrou, A. Felten, C. Lux, S. Sen, Tooth segmentation by low-dose CBCT for orthodontic treatment planning : Explorative ex vivo validation, J Orofac Orthop (2024). https://doi.org/10.1007/s00056-024-00558-7.

[22] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, C. Kelly, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'Souza, S.A. Moinuddin, B. Garie, Y. McQuinlan, S. Ireland, K. Hampton, K. Fuller, H. Montgomery, G. Rees, M. Suleyman, T. Back, C.O. Hughes, J.R. Ledsam, O. Ronneberger, Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study, J Med Internet Res 23 (2021) e26151. https://doi.org/10.2196/26151.

[23] L.H.C. Cevidanes, G. Heymann, M.A. Cornelis, H.J. DeClerck, J.F.C. Tulloch, Superimposition of 3-dimensional cone-beam computed tomography models of growing patients, Am J Orthod Dentofacial Orthop 136 (2009) 94–99. https://doi.org/10.1016/j.ajodo.2009.01.018.

[24] N. Gkantidis, M. Schauseil, P. Pazera, B. Zorkun, C. Katsaros, B. Ludwig, Evaluation of 3-dimensional superimposition techniques on various skeletal structures of the head using surface models, PLoS One 10 (2015) e0118810. https://doi.org/10.1371/journal.pone.0118810.

[25] J. Probst, K. Dritsas, D. Halazonetis, Y. Ren, C. Katsaros, N. Gkantidis, Precision of a Hand-Held 3D Surface Scanner in Dry and Wet Skeletal Surfaces: An Ex Vivo Study, Diagnostics (Basel) 12 (2022) 2251. https://doi.org/10.3390/diagnostics12092251.

[26] M.K. Jaeggi, M. Ghamri, K. Dritsas, S. Psomiadis, C. Verna, D. Halazonetis, N. Gkantidis, Accuracy of anterior cranial base surfaces acquired from computed tomography imaging, Sci Rep 15 (2025) 24321. https://doi.org/10.1038/s41598-025-09104-w.

[27] P.A. Yushkevich, J. Piven, H.C. Hazlett, R.G. Smith, S. Ho, J.C. Gee, G. Gerig, User-guided 3D active contour segmentation of anatomical structures : significantly improved efficiency and reliability, Neuroimage 31 (2006) 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.

[28] G. Dot, DentalSegmentator nnU-Net pretrained model for CBCT image segmentation, (2024). https://doi.org/10.5281/ZENODO.10829675.

[29] P.J. Besl, N.D. Mckay, A Method for Registration of 3-D Shapes, Ieee T Pattern Anal 14 (1992) 239–256. https://doi.org/10.1109/34.121791.

[30] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, O. Ronneberger, R.M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P.F. Christ, R.K.G. Do, M.J. Gollub, S.H. Heckers, H. Huisman, W.R. Jarnagin, M.K. McHugo, S. Napel, J.S.G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J.A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K.

Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A.L. Simpson, L. Maier-Hein, M.J. Cardoso, The Medical Segmentation Decathlon, Nat Commun 13 (2022) 4128. https://doi.org/10.1038/s41467-022-30695-9.

[31] M. Büttner, R. Rokhshad, J. Brinz, J. Issa, A. Chaurasia, S.E. Uribe, T. Karteva, S. Chala, A. Tichy, F. Schwendicke, Core outcomes measures in dental computer vision studies (DentalCOMS), Journal of Dentistry 150 (2024) 105318. https://doi.org/10.1016/j.jdent.2024.105318.

[32] P.G.M. Knoops, C.A.A. Beaumont, A. Borghi, N. Rodriguez-Florez, R.W.F. Breakey, W. Rodgers, F. Angullia, N.U.O. Jeelani, S. Schievano, D.J. Dunaway, Comparison of three-dimensional scanner systems for craniomaxillofacial imaging, Journal of Plastic, Reconstructive & Aesthetic Surgery 70 (2017) 441–449. https://doi.org/10.1016/j.bjps.2016.12.015.

[33] S.T. Häner, G. Kanavakis, F. Matthey, N. Gkantidis, Valid 3D surface superimposition references to assess facial changes during growth, Sci Rep 11 (2021) 16456. https://doi.org/10.1038/s41598-021-95942-3.

[34] D.J. Clegg, A.J. Deek, C. Blackburn, C.A. Scott, J.R. Daggett, The Use and Outcomes of 3D Printing in Pediatric Craniofacial Surgery: A Systematic Review, J Craniofac Surg (2024). https://doi.org/10.1097/SCS.0000000000009981.

[35] M. Ghamri, G. Kanavakis, N. Gkantidis, Reliability of Different Anterior Cranial Base Reference Areas for Voxel-Based Superimposition, J Clin Med 10 (2021) 5429. https://doi.org/10.3390/jcm10225429.

[36] B.A. Sharaf, A. Abushehab, C.E. Michaelcheck, S.M. Hussein, W. Gibreel, O. Alser, A.Z. Kouzani, A.M. Pazelli, K.H. Lee, Y. Oh, M.A. Hawkes, H. Hallak, J.M. Morris, Virtual surgical planning in craniomaxillofacial surgery: A systematic review and meta-analysis of accuracy, operative time, and cost-effectiveness, J Plast Reconstr Aesthet Surg 105 (2025) 305–314. https://doi.org/10.1016/j.bjps.2025.04.037.

[37] L. Birklein, S. Niebler, E. Schömer, R. Brylka, U. Schwanecke, R. Schulze, Motion correction for separate mandibular and cranial movements in cone beam CT reconstructions, Med Phys 50 (2023) 3511–3525. https://doi.org/10.1002/mp.16347.

[38] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat Commun 15 (2024) 654. https://doi.org/10.1038/s41467-024-44824-z.