Dijlah Journal of Medical Sciences (DJMS) Vol. 3, No. 3, September, 2026, pp. 251-259
P-ISSN: 3078-3178, E-ISSN: 3078-8625, paper ID: 22

ORIGINAL ARTICLE

Paired Comparison of Automated and Conventional Mandibular Segmentation

Khawla H. Rasheed*
College of Dentistry, Ibn Sina University for Medical and Pharmaceutical Sciences, Baghdad, Iraq

Article Info

Article history:

Received August, 05, 2026
Revised August, 23, 2026
Accepted September, 10, 2026

Keywords:

Artificial Intelligence,
Computed Tomography,
Mandibular Segmentation,
Surface Distance,
Total Segmentation

ABSTRACT

Mandibular segmentation is a critical step in three-dimensional virtual surgical planning, but conventional threshold-based segmentation can require substantial operator interaction, particularly near the dentition and temporomandibular joints. This paired study compared automated and conventional mandibular segmentation in 30 computed tomography (CT) cases from the Public Domain Database for Computational Anatomy (PDDCA). Automated segmentation was performed in 3D Slicer version 5.12.3 with TotalSegmentator version 2.14.0 using the craniofacial-structures task; the separate lower-teeth label was excluded. Conventional segmentation was performed independently by the same investigator, who had approximately six years of experience in three-dimensional medical-image processing and segmentation. The PDDCA reference was withheld during both segmentation pathways. Across the 30 paired cases, automated versus conventional segmentation yielded: mean Dice similarity coefficient was 0.821 ± 0.021 for automated segmentation and 0.828 ± 0.015 for conventional segmentation; 95th percentile Hausdorff distance (HD95) was 2.392 ± 0.241 versus 3.446 ± 0.943 mm, average symmetric surface distance (ASSD) was 0.951 ± 0.094 versus 1.152 ± 0.165 mm, and maximum symmetric surface deviation was 4.992 ± 1.051 versus 17.737 ± 8.686 mm. Mean signed volume difference was +33.06% versus +23.86%, respectively, Matched total workflow time was 23.38 ± 1.94 min for the automated pathway and 106.80 ± 13.38 min for the conventional pathway (mean paired difference = -83.42 min; 95% CI, -88.54 to -78.29 min; $p < 0.001$), corresponding to a 78.1% reduction.

Corresponding Author:

* Khawla H. Rasheed
College of Dentistry, Ibn Sina University for Medical and Pharmaceutical Sciences, Baghdad, Iraq
Email: khawlarasheed.g@ibnsina.edu.iq

1- INTRODUCTION

Virtual surgical planning has become an established component of mandibular reconstruction and orthognathic surgery because three-dimensional anatomical models permit preoperative simulation of osteotomy, segment repositioning, reconstruction, and patient-specific transfer strategies [1-3]. Computer-assisted workflows may also reduce planning or operative time in selected craniomaxillofacial applications [4-6]. The reliability of these workflows begins with segmentation; errors introduced while converting computed tomography (CT) data into a mandibular model can propagate to later measurements, guide design, reconstruction, and accuracy assessment. Mandibular segmentation remains technically demanding at the condyles, thin cortical margins, dental region, and areas affected by metal or beam-hardening artifacts. Reviews describe substantial operator dependence in conventional approaches and rapid development of deep learning alternatives [7-9]. Previous systems based on three-dimensional neural networks, shape priors, coarse-to-fine learning, and attention mechanisms, and external validation has reported high volumetric agreement [10-13]. However, high algorithm-level accuracy does not automatically establish that an automated output is appropriate for a specific virtual surgical-planning workflow [14-16]. The present study used a paired within-case design in which the same CT case was processed independently by automated and conventional segmentation and both generated masks were evaluated only afterward against the same withheld reference. The objectives were to compare workflow efficiency and geometric accuracy while explicitly accounting for reference anatomy, dentition handling, blinding, image geometry, quality control, and complementary overlap and surface-distance metrics.

2- MATERIALS AND METHODS

2.1 Study design, protocol development, and analysis unit

This study used a paired within-case design comprising 30 CT cases. Each case was processed independently by (1) an automated segmentation pathway and (2) a conventional manual/semi-automatic pathway, as shown in figure (1). The original PDDCA Mandible.nrrd file was withheld during both segmentation procedures, and the automated output was not displayed during conventional segmentation. Each CT case therefore served as its own control. The first five cases constituted a technical-development phase used to lock the file structure, crop rule, mandible-only target, handling of the separate lower-teeth label, temporomandibular-joint (TMJ) separation rule, connected-component cleaning, reference-grid mapping, metric implementation, and timing definitions. The locked protocol was then applied unchanged to the full 30-case cohort.

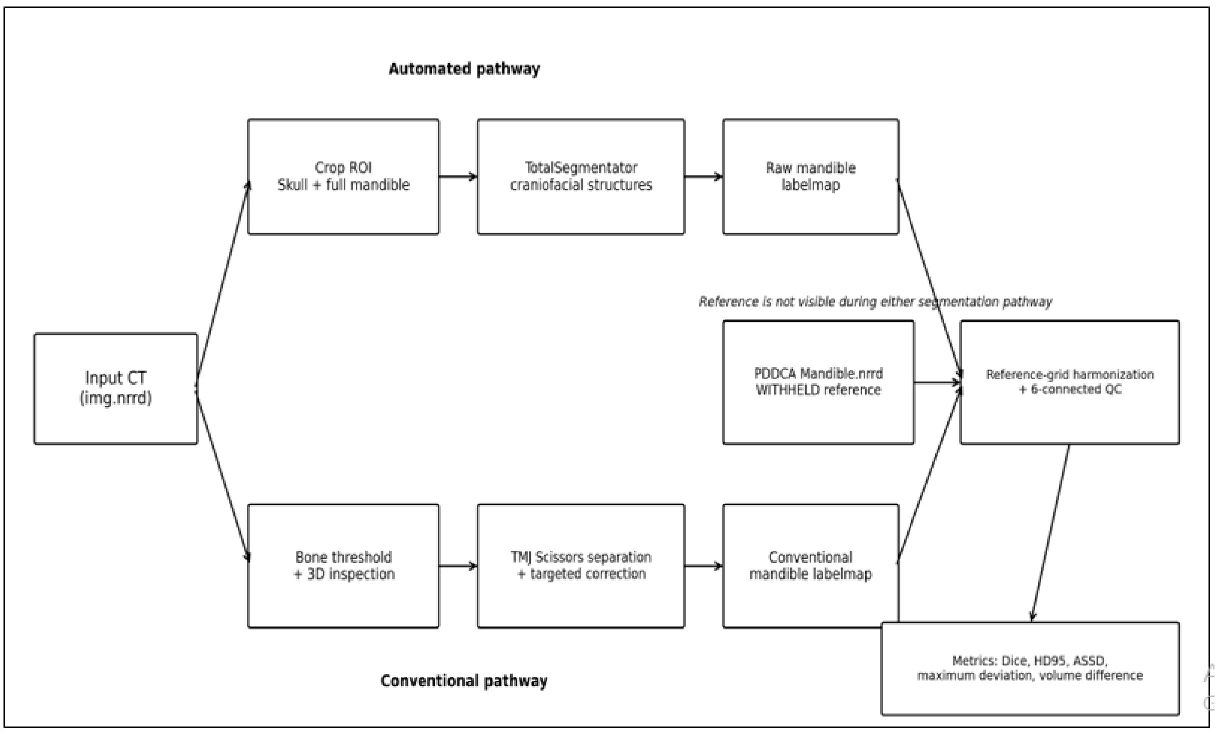

Fig (1): Paired segmentation and evaluation workflow. The same CT was processed independently by the automated and conventional pathways. The PDDCA reference was withheld during both segmentations and introduced only after mask finalization for reference-grid harmonization, standardized quality control, and quantitative comparison

2.2 Dataset, reference standard, operator, and blinding

The study used de-identified head-and-neck CT data from the Public Domain Database for Computational Anatomy (PDDCA), originating from the 2015 Head and Neck Auto-Segmentation Challenge [17]. In the challenge protocol, the mandible was contoured from the chin to the coronoid processes and condyles, with particular attention to separating mandibular bone from the teeth. The supplied reference was therefore treated as an independent geometric reference rather than a definitive virtual-surgical-planning ground truth. The challenge reference contours were manually delineated and quality-assured; case-specific inter-observer variability for the mandible reference is not available. One investigator with approximately six years of experience in three-dimensional medical-image processing and segmentation handled both study pathways: she initiated and managed the automated workflow and performed the conventional segmentation. The investigator was aware of the comparative study objective but was blinded to the PDDCA reference and to the opposing segmentation output while each mask was being generated.

2.3 Software and workstation environment

Segmentation was performed in 3D Slicer version 5.12.3 on a Windows workstation with approximately 16 GB of system memory and Intel UHD integrated graphics without a dedicated NVIDIA graphics processor. Automated inference used the SlicerTotalSegmentator extension with TotalSegmentator version 2.14.0. The craniofacial-structures task generated separate labels that included the mandible and lower teeth, in addition to other craniofacial structures [17, 18, 19]. Only the mandible label was retained for the primary analysis; the separately generated lower-teeth label was excluded. The pretrained model weights distributed with the installed TotalSegmentator package were used without retraining, fine-tuning, or modification. Because a separate mandible-only validation statistic for this exact software/task version was not available from the installed task documentation, mandible performance was evaluated independently in the present study against the withheld PDDCA reference. Work flow stages are summarized in figure (2).

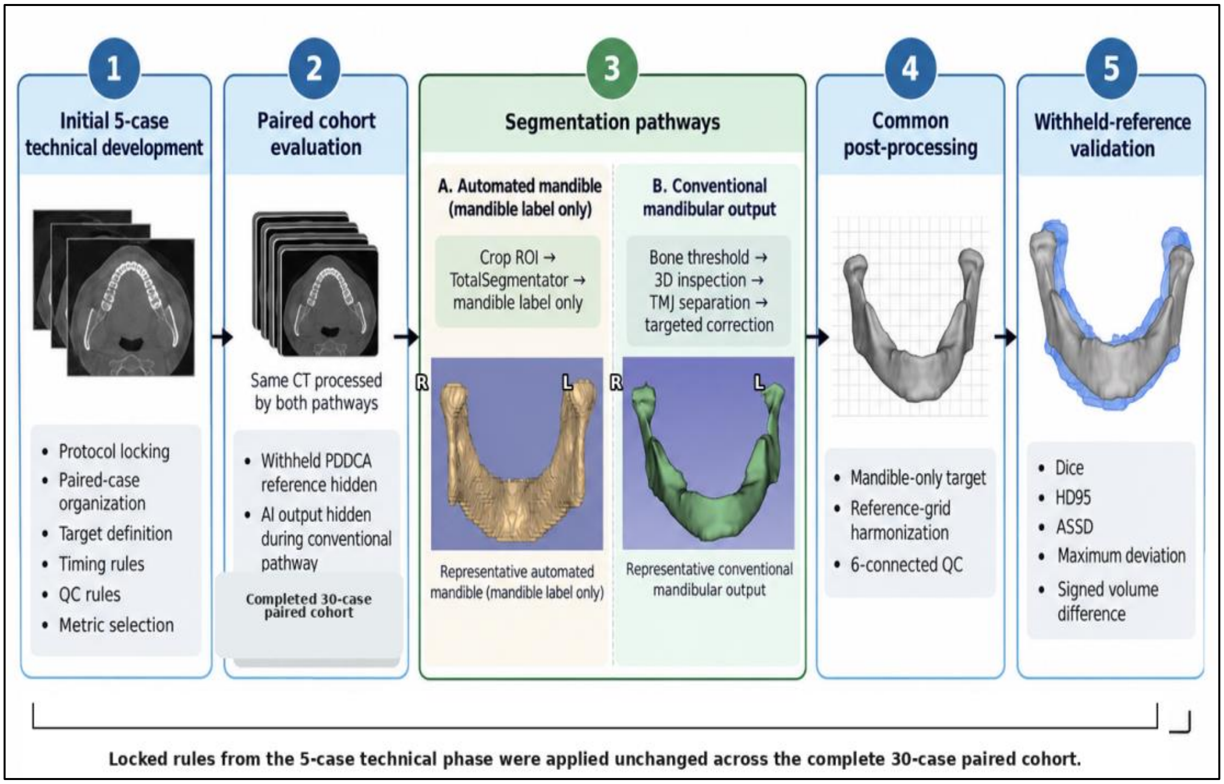

Fig (2): Stage-based workflow development and validity assessment. The first five cases were used to lock the technical protocol, after which the same rules were applied unchanged to the complete 30-case paired cohort. Stage 3 shows the two representative frontal-view mandibular renderings: the automated mandible in tan and the conventional mandible in green. Patient right (R) and left (L) are labeled on both renderings. The colors distinguish the two pathways and do not encode accuracy. The PDDCA reference remained withheld until both segmentation outputs were finalized

2.4 Automated segmentation workflow

Only the original img.nrrd file was loaded for automated segmentation. Because full-volume inference placed a substantial memory load on the available workstation, a standardized Crop Volume step was applied before inference. The crop was defined anatomically rather than by a fixed distance below the mandibular border: it included the entire mandible, both condyles and TMJ regions, and the upper neck immediately below the mandibular inferior border, while excluding the chest and shoulders. Interpolated cropping was disabled, spacing scale was maintained at 1.00, and isotropic spacing was disabled to avoid unnecessary preprocessing resampling. Total Segmentation was run with the craniofacial-structures task in Normal quality mode. The mandible label was retained as the automated raw output and teeth lower was excluded. No investigator-applied Paint, Erase, Scissors, or explicit post-segmentation smoothing filter was applied to the automated mandible. Neural-network inference and label resampling can nevertheless influence voxel-boundary appearance; these operations were not interpreted as equivalent to an investigator-applied smoothing step. The output was exported as a binary NRRD labelmap. When an output had been saved on a cropped grid, it was computationally mapped to the original PDDCA reference geometry before metric calculation.

2.5 Conventional segmentation workflow

For conventional segmentation, the scene was reopened with the original CT only; neither the PDDCA reference nor the automated mask was displayed. A new Segment Editor segment was created for the mandible. A lower threshold of 250 HU was used as the locked starting value to suppress soft tissue while retaining cortical bone. The threshold was not optimized against the PDDCA reference. Three-dimensional inspection followed, and any case-specific threshold adjustment was based only on the source CT and was to be recorded in the case log. A recurring segmentation challenge was continuity between the mandibular condylar region and adjacent temporal or skull bone after thresholding. When an Islands operation alone did not separate the mandible, the Scissors tool (Erase inside, free-form) was used on magnified sagittal or coronal views to sever the connecting bridge while preserving the condylar head. Limited targeted Paint/Erase correction was permitted only for obvious CT-based contour errors. No explicit smoothing filter was applied. The accepted segment was exported as a binary labelmap on the original CT geometry.

2.6 Standardized quality control and connected-component cleaning

Before quantitative comparison, all generated masks underwent the same predefined connected-component quality-control rule. Small isolated components unrelated to the principal mandibular object were removed by retaining the largest six-connected three-dimensional component. For every case, the number of removed components, removed voxel count, and removed physical volume were recorded separately for the automated and conventional pathways. In all cases 10 disconnected components totaling 10 voxels about (35.31 mm³) were removed from the automated masks, whereas 20 disconnected components totaling 55 voxels (203.72 mm³) were removed from the conventional masks. The largest removed fraction was 0.22 % of the generated mask volume. The PDDCA reference masks were not modified.

2.7 Reference-grid harmonization and quantitative metrics

All comparisons were performed on the PDDCA reference grid. When a generated mask was stored on a cropped grid, it was mapped to the reference geometry using the NRRD origin, spacing, and direction matrix with nearest-neighbor label resampling. All surface calculations were performed in physical millimeters so that anisotropic voxel spacing was respected. Five geometric outcomes were calculated. The Dice similarity coefficient was calculated as twice the intersection volume divided by the sum of the generated-segmentation and reference volumes. A one-voxel internal boundary was extracted using a six-connected neighborhood. Surface-to-surface distances were calculated in both directions and pooled. The 95th percentile of the pooled distances was reported as the 95th percentile Hausdorff distance; the arithmetic mean was reported as the average symmetric surface distance; and the maximum was reported as the maximum symmetric surface deviation. Volume was calculated from the foreground voxel count multiplied by the physical voxel volume. Signed volume difference was defined as $100 \times (V_{seg} - V_{ref})/V_{ref}$, with positive values indicating over-segmentation. Dice and surface-distance measures were interpreted together because they captured complementary error properties [20, 21, 22].

2.8 Workflow timing

Matched total-workflow boundaries were used for all 30 timing observations. All timings were recorded prospectively by the same investigator using a timer. Each predefined workflow step was timed separately as it was performed and was entered into the case timing log. Timing began with the first pathway-specific processing action after CT loading and ended after export of the final binary mask and completion of the predefined quality-control step. Automated total time included crop preparation, inference, export, geometry mapping when required, and connected-component quality control. Conventional total time included thresholding, manual separation/correction, export, and the same quality-control step. Active operator time and unattended computer-processing time were recorded separately, and total workflow time was derived from the recorded step times. The initial software-installation/model-initialization run was excluded as a setup run. Earlier inference-only automated timing values were not used in the revised comparison. The previously reported automated value of 23.0 ± 5.52 min was based on the narrower inference-only timing definition and was discarded from the revised analysis. All revised timing results were derived from the prospectively recorded step-level timer log.

2.9 Statistical analysis

The paired case was the unit of analysis (n = 30). For each outcome, the distribution of within-case differences was inspected graphically and assessed using the Shapiro-Wilk test; the W statistic and p-value were reported. Approximately normal paired differences were analyzed using a paired t-test with Cohen's dz and a 95% confidence interval. Non-normal paired differences were analyzed using the Wilcoxon signed-rank test with matched-pairs rank-biserial correlation and a 95% confidence interval. Holm adjustment was applied across the five prespecified geometric outcomes. Total workflow time was analyzed as a distinct prespecified efficiency hypothesis, and a sensitivity Holm correction across all six outcomes was also reported. Statistical significance was set at $p < 0.05$.

3- RESULTS AND DISCUSSION

3.1 Cohort characteristics and processing feasibility

All 30 CT cases completed segmentation pathways, quality control, reference-grid harmonization, and final paired analysis. The locked workflow was applied consistently across the cohort, and no case was excluded after protocol locking. Among the 30 included CT cases, 15 (50.0%) were dentate, 13 (43.3%) were partially edentulous, and 2 (6.7%) were edentulous. Dental or metallic artifacts were absent in 22 cases (73.3%), mild in 5 (16.7%), and moderate in 3 (10.0%); no case showed severe artifact. No obvious mandibular pathology was identified in any of the 30 cases. Age, sex, and clinical diagnosis were not available in the distributed dataset. Case characteristics are summarized in table 1.

Table (1): Characteristics of the CT cases included in the study (n = 30)

Characteristic	Category	n (%)
Dentition status	Dentate	15 (50.0%)
	Partially edentulous	13 (43.3%)
	Edentulous	2 (6.7%)
Dental/metal artifact	None	22 (73.3%)
	Mild	5 (16.7%)
	Moderate	3 (10.0%)
	Severe	0 (0.0%)
Visible mandibular pathology	None visible	30 (100.0%)
	Visible pathology	0 (0.0%)
Age	Not available in the distributed dataset	—
Sex	Not available in the distributed dataset	—
Clinical diagnosis	Not available in the distributed dataset	—
Case selection	Purposive selection of cases with available CT, PDDCA mandible reference, and verified image-reference geometry	—

3.2 Volumetric overlap and volume agreement

Across the 30 paired cases, the automated and conventional pathways were compared for volumetric overlap, absolute mandibular volume, and signed volume difference relative to the PDDCA reference. The Dice similarity coefficient was 0.821 ± 0.021 for automated segmentation and 0.828 ± 0.015 for conventional segmentation. The volume findings were interpreted together with surface-distance outcomes and were not used alone to define clinical acceptability.

3.3 Surface-distance accuracy and worst-case review

Across the 30 paired cases, HD95, ASSD, and maximum symmetric surface deviation were calculated for both segmentation pathways. The 95th percentile Hausdorff distance (HD95) was 2.392 ± 0.241 versus 3.446 ± 0.943 mm, average symmetric surface distance (ASSD) was 0.951 ± 0.094 versus 1.152 ± 0.165 mm, and maximum symmetric surface deviation was 4.992 ± 1.051 versus 17.737 ± 8.686 mm for automated and conventional segmentation respectively. Mean signed volume difference was +33.06% versus +23.86%, respectively for each pathway.] Maximum deviation was retained as a complementary safety-sensitive geometric outcome rather than reclassified post hoc as a co-primary endpoint. Spatial review was performed across the complete 30-case cohort to determine whether discrepancies were diffuse or concentrated in specific mandibular regions. Worst-performing conventional cases by maximum surface deviation in the verified five-case analysis are shown in table (2).

Table (2): Worst-performing conventional cases by maximum surface deviation in the verified thirty-case analysis

Case	Auto max (mm)	Conv max (mm)	Conv Dice	Conv vol. diff. (%)	Interpretive feature
Case 01	3.16	26.88	0.804	33.7	Gross localized boundary excursion
Case 07	5.71	23.77	0.851	1.0	Large local error despite near-neutral global volume
Case 22	5.09	20.00	0.79	27.2	Localized error with concurrent volume excess

3.4 Workflow timing

Across 30 matched cases, total workflow time was 23.38 ± 1.94 min for automated segmentation and 106.80 ± 13.38 min for conventional segmentation. The mean paired difference was -83.42 min (95% CI, -88.54 to -78.29 min; $p < 0.001$), with Cohen's dz = -6.08 (95% noncentral-t CI, -7.67 to -4.48), representing a 78.1% reduction in total workflow time. Active operator time was 6.52 ± 1.49 min for the automated pathway and 69.50 ± 9.81 min for the conventional pathway, a 90.6% reduction in hands-on time. Unattended computer-processing time was 16.87 ± 1.74 min for automated segmentation and 37.30 ± 7.52 min for conventional segmentation. The paired total-time difference was approximately normal ($W = 0.975$, $p = 0.676$), supporting use of the paired t-test. Although the revised automated total mean (23.38 min) is numerically close to the previously reported inference-only mean (23.0 min), the two values were obtained using different timing definitions. The revised total was independently derived from the prospective step-level timer log and included both active operator time (6.52 min on average) and unattended computer-processing time (16.87 min on average); the component means sum to approximately 23.39 min because of rounding. The earlier inference-only value was discarded and was not reused in the revised comparison.

3.5 Final paired statistical comparison

The final paired statistical analysis was performed on all 30 cases is shown in table (3).

Table (3): Final paired statistical comparison across the 30 CT cases

Outcome	Automated	Conventional	Paired diff. (95% CI)	Shapiro-Wilk W (p)	p / Holm p	Cohen dz (95% bootstrap CI)
Dice coefficient	0.821 ± 0.021	0.828 ± 0.015	-0.007 (-0.033, +0.020)	0.899 (0.402)	0.519 / 0.519	-0.32 (-1.27, +2.01)
HD95	2.392 ± 0.241 mm	3.446 ± 0.943 mm	-1.054 mm (-2.202, +0.093)	0.940 (0.665)	0.063 / 0.236	-1.14 (-4.41, -0.68)
ASSD	0.951 ± 0.094 mm	1.152 ± 0.165 mm	-0.201 mm (-0.415, +0.012)	0.973 (0.894)	0.059 / 0.236	-1.17 (-5.01, -0.35)
Maximum symmetric surface deviation	4.992 ± 1.051 mm	17.737 ± 8.686 mm	-12.745 mm (-24.332, -1.157)	0.974 (0.903)	0.038 / 0.189	-1.37 (-5.57, -0.70)
Signed volume difference	33.06 ± 8.02%	23.86 ± 13.17%	+9.20 pp (-2.05, +20.45)	0.947 (0.714)	0.086 / 0.236	+1.01 (+0.54, +2.85)

3.6 Comparison with published evidence

Published algorithm-development and validation studies frequently report high mandible Dice values. A 2025 systematic review reported a pooled Dice of approximately 0.94 for artificial-intelligence-based maxillofacial segmentation [15], while other external and development studies reported values around or above 0.94 [13, 14, 16]. Direct numerical comparison with the present 30-case external paired workflow evaluation should be made cautiously because the current study used an unmodified automated output, heterogeneous CT geometry, and a radiotherapy-oriented PDDCA anatomical reference rather than a surgical-planning ground truth. Dice, volume difference, and surface-distance measures can rank segmentation outputs differently because they capture different error properties [21, 22]. The observed combination of similar Dice values and different signed volume errors is therefore not contradictory; excess segmentation may occur in spatially different regions while preserving similar overall overlap. For virtual surgical planning, localized discrepancies near the condyle, ramus, dental region, or inferior border may be clinically important even when global overlap appears similar. The final analysis therefore treats the metrics as a multi-dimensional error profile rather than relying on a single score. The present study evaluated an existing automated tool rather than developing a new neural network. U-Net, 3D U-Net, V-Net, and related architectures established the foundations of contemporary medical-image segmentation [23-26], whereas nnU-Net standardized many configuration decisions [19]. Total Segmentator extends automated segmentation to broad anatomical coverage [18], and its craniofacial-structures task provides separate mandible and dental labels. The methodological contribution of the present study was the controlled paired comparison, reference withholding, explicit dentition handling, geometry harmonization, standardized quality control, and joint assessment of workflow efficiency and geometric accuracy.

3.7 Limitations and generalizability

Several limitations must be considered. First, PDDCA was created for head-and-neck radiotherapy auto-segmentation rather than virtual surgical planning. Its mandible contour includes the chin, coronoid processes, and condyles and emphasizes separation from the teeth, but the target may still differ from dental or surgical-planning conventions [17]. Second, case-specific inter-observer variability of the mandible reference is unavailable. Third, one experienced investigator performed the conventional workflow, so inter-operator reproducibility was not assessed. Fourth, 250 HU was used as a fixed starting threshold and may not generalize to osteoporotic, sclerotic, or artifact-heavy scans. Fifth, largest-component cleaning could remove a clinically meaningful disconnected fracture fragment. The 30-case cohort included 15 dentate, 13 partially edentulous, and 2 edentulous cases; dental or metal artifacts were absent in 22 cases, mild in 5, and moderate in 3, with no severe artifacts observed. No obvious mandibular pathology was identified in any included case. Age, sex, and clinical diagnosis were not available in the distributed dataset. Finally, PDDCA is a 2015 radiotherapy dataset; its age, heterogeneous CT protocols, scanner characteristics, cancer population, dentition, and dental hardware limit direct generalizability when evaluating a current TotalSegmentator installation.

4- CONCLUSION

In this 30-case paired comparison, automated segmentation reduced total workflow time from 106.80 ± 13.38 min to 23.38 ± 1.94 min, a 78.1% reduction, and reduced active operator time from 69.50 ± 9.81 min to 6.52 ± 1.49 min. In the current study, complete geometric measurements, automated and conventional mandibular segmentation showed similar global overlap, while automated segmentation had lower HD95, ASSD, and maximum surface deviation and conventional segmentation showed lower mean volume excess. None of the 30 geometric outcomes remained statistically significant after Holm adjustment, although the conventional maximum-deviation pattern revealed clinically important, localized failures in several cases. These findings support the use of complementary overlap, volume, and surface metrics rather than Dice alone. The findings support interpretation of overlap, volume, and surface-distance metrics together rather than reliance on Dice alone.

REFERENCES

[1] Barr, M. L., Haveles, C. S., Rezzadeh, K. S., Nolan, I. T., Castro, R., Lee, J. C., Steinbacher, D., & Pfaff, M. J. (2020). Virtual surgical planning for mandibular reconstruction with the fibula free flap: A systematic review and meta-analysis. *Annals of Plastic Surgery*, *84*(1), 117–122. https://doi.org/10.1097/SAP.0000000000002006

[2] Wilde, F., Cornelius, C.-P., & Schramm, A. (2014). Computer-assisted mandibular reconstruction using a patient-specific reconstruction plate fabricated with computer-aided design and manufacturing techniques. *Craniomaxillofacial Trauma & Reconstruction*, *7*(2), 158–166. https://doi.org/10.1055/s-0034-1371356

[3] Annino, D. J., Jr., Sethi, R. K., Hansen, E. E., Horne, S., Dey, T., Rettig, E. M., Uppaluri, R., Kass, J. I., & Goguen, L. A. (2022). Virtual planning and 3D-printed guides for mandibular reconstruction: Factors impacting accuracy. *Laryngoscope Investigative Otolaryngology*, *7*(6), 1798–1807. https://doi.org/10.1002/lio2.830

[4] El-Mahallawy, Y., Abdelrahman, H. H., & Al-Mahalawy, H. (2023). Accuracy of virtual surgical planning in mandibular reconstruction: Application of a standard and reliable postoperative evaluation methodology. *BMC Oral Health*, *23*, Article 119. https://doi.org/10.1186/s12903-023-02811-8

[5] Alkhayer, A., Piffkó, J., Lippold, C., & Dirksen, D. (2020). Accuracy of virtual planning in orthognathic surgery: A systematic review. *Head & Face Medicine*, *16*, Article 34. https://doi.org/10.1186/s13005-020-00250-2

[6] Chen, Z., Li, J., Huang, X., & Wang, X. (2021). A meta-analysis and systematic review comparing the effectiveness of traditional and virtual surgical planning for orthognathic surgery: Based on randomized clinical trials. *Journal of Oral and Maxillofacial Surgery*, *79*(2), 471.e1–471.e19. https://doi.org/10.1016/j.joms.2020.09.005

[7] Alkaabi, S., Al-Moraissi, E. A., Al-Shamiri, H. M., & Gulfam, M. (2022). Virtual and traditional surgical planning in orthognathic surgery—Systematic review and meta-analysis. *British Journal of Oral and Maxillofacial Surgery*, *60*(9), 1184–1191. https://doi.org/10.1016/j.bjoms.2022.07.007

[8] Nilsson, J., Hindocha, N., & Thor, A. (2020). Time matters—Differences between computer-assisted surgery and conventional planning in cranio-maxillofacial surgery: A systematic review and meta-analysis. *Journal of Cranio-Maxillofacial Surgery*, *48*(2), 132–140. https://doi.org/10.1016/j.jcms.2019.11.024

[9] Qiu, B., Ding, G., Zhao, J., Chen, Y., & Jiang, H. (2021). Automatic segmentation of mandible from conventional methods to deep learning—A review. *Journal of Personalized Medicine*, *11*(7), Article 629. https://doi.org/10.3390/jpm11070629

[10] Verhelst, P.-J., Van Dessel, J., Sun, Y., Jacobs, R., & Willems, G. (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

[11] Qiu, B., Ding, G., Zhao, J., Chen, Y., & Jiang, H. (2021). Robust and accurate mandible segmentation on dental CBCT scans affected by metal artifacts using a prior shape model. *Journal of Personalized Medicine*, *11*(5), Article 364. https://doi.org/10.3390/jpm11050364

[12] Qiu, B., Ding, G., Zhao, J., Chen,Y., & Jiang, H. (2021). Mandible segmentation of dental CBCT scans affected by metal artifacts using coarse-to-fine learning model. *Journal of Personalized Medicine*, *11*(6), Article 560. https://doi.org/10.3390/jpm11060560

[13] Xu, J., Tong, T., & Ding, X. (2021). Automatic mandible segmentation from CT image using 3D fully convolutional neural network based on DenseASPP and attention gates. *International Journal of Computer Assisted Radiology and Surgery*, *16*(10), 1785–1794. https://doi.org/10.1007/s11548-021-02447-5

[14] Kargilis, D. C., et al. (2025). Deep learning segmentation of mandible with lower dentition from cone beam CT. *Oral Radiology*, *41*(1), 1–9. https://doi.org/10.1007/s11282-024-00770-6

[15] Alahmari, M., et al. (2025). Accuracy of artificial intelligence-based segmentation in maxillofacial structures: A systematic review. *BMC Oral Health*, *25*, Article 350. https://doi.org/10.1186/s12903-025-05730-y

[16] Beyer, M., et al. (2026). An innovative AI-based dual segmentation application for head surgery. *International Journal of Oral and Maxillofacial Surgery*, *55*(4), 489–495. https://doi.org/10.1016/j.ijom.2025.11.005

[17] Raudaschl, P. F., et al. (2017). Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Medical Physics*, *44*(5), 2020–2036. https://doi.org/10.1002/mp.12197

[18] Wasserthal, J., et al. (2023). TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, *5*(5), e230024. https://doi.org/10.1148/ryai.230024

[19] 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

[20] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

[21] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, *15*, Article 29. https://doi.org/10.1186/s12880-015-0068-x

[22] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863. https://doi.org/10.1109/34.232073

[23] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (Lecture Notes in Computer Science, Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[24] 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. H. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2016* (Lecture Notes in Computer Science, Vol. 9901, pp. 424–432). Springer. https://doi.org/10.1007/978-3-319-46723-8_49

[25] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE. https://doi.org/10.1109/3DV.2016.79

[26] Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025