Journal of Imaging Informatics in Medicine (2024) 37:1863–1873
https://doi.org/10.1007/s10278-024-01006-z

Automated Segmentation and Diagnostic Measurement for the Evaluation of Cervical Spine Injuries Using X-Rays

Jae Hyuk Shim[1] · Woo Seok Kim[2] · Kwang Gi Kim[1] · Gi Taek Yee[3] · Young Jae Kim[1] · Tae Seok Jeong[2]

Received: 18 August 2023 / Revised: 21 December 2023 / Accepted: 22 December 2023 / Published online: 20 February 2024

Abstract

Accurate assessment of cervical spine X-ray images through diagnostic metrics plays a crucial role in determining appropriate treatment strategies for cervical injuries and evaluating surgical outcomes. Such assessment can be facilitated through the use of automatic methods such as machine learning and computer vision algorithms. A total of 852 cervical X-rays obtained from Gachon Medical Center were used for multiclass segmentation of the craniofacial bones (hard palate, basion, opisthion) and cervical spine (C1–C7), incorporating architectures such as EfficientNetB4, DenseNet201, and InceptionResNetV2. Diagnostic metrics automatically measured using computer vision algorithms were compared with manually measured metrics through Pearson's correlation coefficient and paired *t*-tests. The three models demonstrated high average dice coefficient values for the cervical spine (C1, 0.93; C2, 0.96; C3, 0.96; C4, 0.96; C5, 0.96; C6, 0.96; C7, 0.95) and lower values for the craniofacial bones (hard palate, 0.69; basion, 0.81; opisthion, 0.71). Comparison of manually measured metrics and automatically measured metrics showed high Pearson's correlation coefficients in McGregor's line ($r=0.89$), space available cord ($r=0.94$), cervical sagittal vertical axis ($r=0.99$), cervical lordosis ($r=0.88$), lower correlations in basion-dens interval ($r=0.65$), basion-axial interval ($r=0.72$), and Powers ratio ($r=0.62$). No metric showed adjusted significant differences at $P<0.05$ between manual and automatic metric measuring methods. These findings demonstrate the potential of multiclass segmentation in automating the measurement of diagnostic metrics for cervical spine injuries and showcase the clinical potential for diagnosing cervical spine injuries and evaluating cervical surgical outcomes.

Keywords Cervical spine · Craniofacial bones · Machine learning · Computed radiography · Computer vision

Abbreviations

MG	McGregor's line
BDI	Basion-dens interval
BAI	Basion-axial interval
SAC	Space available cord
cSVA	Cervical sagittal vertical axis
CL	Cervical lordosis

Introduction

Cervical spine injuries resulting from high-energy trauma can have severe consequences, including sensory function loss due to the proximity to vital nerves [1, 2]. Prompt treatment is crucial for successful stabilization and preventing mortality, especially in cases such as traumatic atlanto-occipital dislocation (TAOD). X-rays are commonly used to visualize the cervical spine to assess for fractures and ligament injuries, determining injury severity and preparing appropriate surgical procedures [3]. While X-rays can be an outdated modality for evaluating cervical injuries compared to the use of brain CT, the wide accessibility of X-ray imaging in addition to its cost-effectiveness can be a valuable tool for initial screening of cervical injuries. Assessment of cervical injury using X-rays can be done by measuring the

Jae-Hyuk Shim and Woo Seok Kim contributed equally to this work.

✉ Kwang Gi Kim
kimkg@gachon.ac.kr

✉ Gi Taek Yee
gtyee@gilhospital.com

[1] Department of Biomedical Engineering, Gil Medical Center, Gachon University College of Medicine, Incheon, Korea

[2] Department of Traumatology, Gil Medical Center, Gachon University College of Medicine, Incheon, Korea

[3] Department of Neurosurgery, Gil Medical Center, Gachon University College of Medicine, Incheon, Korea

relative distances or angles of various cervical spinal structures and craniofacial bones [4, 5]. However, each diagnostic metric may show inconsistent results depending on the type of injury sustained by the patient and may require multiple combinations of diagnostic metrics for proper evaluation [3]. Additionally, there can be cases where diagnosis is difficult due to poor X-ray resolution amplifying bone superimposition, obfuscating the necessary spinal structures for measuring diagnostic metrics.

To address such issues, methods involving machine learning and algorithms to automate cervical spine segmentation and diagnostic metric measurement can be a promising solution to aid in cervical injury diagnosis. Many studies have previously attempted to incorporate machine learning algorithms to segment the cervical spine automatically [6, 7]. However, most of them limited their scope to lower cervical segments due to X-ray obfuscation from bone superimposition and did not include craniofacial bones (hard palate, basion, opisthion) required for evaluating cervical spine injuries such as TAOD with the powers ratio [2, 9]. Additionally, there is a lack of research on automatically measuring cervical diagnostic metrics using segmented cervical spine regions.

In this study, we aimed to implement multiclass segmentation of the cervical regions (hard palate, basion, opisthion, C1–C7) using U-Net architectures with EfficientNet-B4, DenseNet201, and InceptionResNetV2 backbones trained on X-ray images of normal and pre/postoperative patients [8]. Subsequently, we developed algorithms to automatically measure diagnostic metrics McGregor line (MG), basion-dens interval (BDI), basion-atlas interval (BAI), Powers ratio, space-available-cord (SAC), cervical sagittal vertical axis (cSVA), and cervical lordosis (CL) using the multiclass segmentations [10–12]. We then compared the automatically generated measurements with manual measurements obtained from manually segmented regions to verify its performance. By demonstrating the reliability and efficiency of our automated approach compared to manual measurements, we aim to contribute to the development of automated tools for accurate diagnosis of cervical spine injuries and prediction of surgical outcomes.

Materials and Methods

Data Acquisition

This retrospective study protocol was approved by the institutional review board at Gachon University Gil Medical Center (GDIRB2022-190). Methods used in this study were all in accordance with the relevant guidelines and regulations of the declaration of Helsinki, and informed consent was waived by the institutional review board at Gachon University Gil Medical Center due to the retrospective nature of the study.

A total of 1062 X-ray images of unique patients in one of two groups were obtained from the Gachon University Gil Medical Center between 2009 and 2022 for this study. One group consisted of normal subjects that did not show visible signs of traumatic cervical injury or were suffering from different degenerative conditions ($n=954$). The second consisted of pre/postoperative subjects that showed symptoms of traumatic cervical injury or cervical degeneration with visible spinal fractures or with spinal implants ($n=108$). Around 210 images were excluded for complete obfuscation of bones, most commonly C7, either from overlap with the chest ($n=155$) or spinal implants ($n=33$), and attached bone structures ($n=22$), most commonly the anterior arc of C1 attached to the dens of C2. As a result, 777 normal images and 75 pre/postoperative images were used for the study.

The image resolutions ranged from 0.107 mm per pixel to 0.194 mm per pixel, with an average width of 1712.0 pixels and height of 2111.7 pixels. Each X-ray image was converted to a resolution of 1.0 mm per pixel and resized to 512×512 pixels. Contrast-limited adaptive histogram equalization (CLAHE) was applied to enhance image quality; then, the images were normalized to values between 0 and 1. The dataset was first randomly divided into two groups, 752 images for training (702 normal, 50 pre/postoperative) and 100 images for testing (75 normal, 25 pre/postoperative). Stratified k-fold cross validation was done for the training data to create 5 k-folds with even distribution of 560 normal and 40 pre/post operative images for the training set and 140 normal and 10 pre/post operative images for the validation set.

The delineation of the segmentation targets, hard palate, basion, opisthion, and C1–C7 vertebrae and measurement of diagnostic metrics, MG, BDI, BAI, Powers ratio, SAC, cSVA, and CL were done by eight graduate students and validated by two medical doctors (W.S.K. and T.S.J., with 8 years and 12 years of experience in traumatic brain and spine injury) with an example segmentation shown in Fig. 1. Clinical information of patients for each X-ray was available for each subject during the delineation. The measured diagnostic metrics were adjusted for changes in resolution (to 1.0 mm per pixel) and size (512×512 pixels) to ensure consistency across the dataset.

Network

The three U-Net architectures pretrained on ImageNet with backbones EfficientNetB4, DenseNet201, and InceptionResNetv2 served as the primary frameworks for training and validating our cervical spine X-ray dataset [8, 13–16]. The neural networks were trained on a server instance equipped with two NVIDIA RTX 2080tis, providing a total of 22 GB

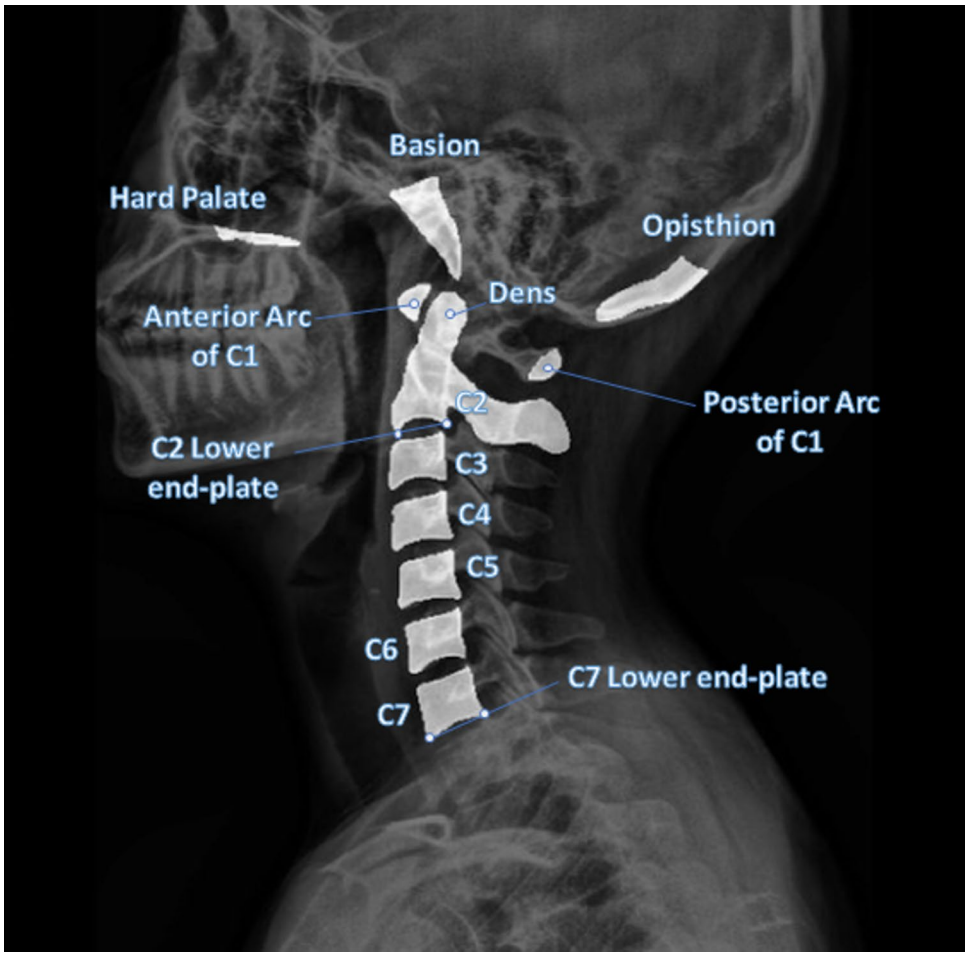

Fig. 1 Cervical regions segmented for measuring metrics used in diagnosing cervical injuries

of GPU memory. The training configuration included a batch size of 4, 200 epochs (with early stopping if validation accuracy did not improve for 40 epochs), the Adam optimizer, and a learning rate of 0.001 (reduced by a factor of 0.1 if there was no improvement in validation accuracy for 10 epochs). The loss function employed for training the multiclass segmentation model was based on dice coefficient weights, with each class (including the background) assigned a weight of 0.091.

Evaluation

Each multiclass model (EfficientNetB4, DenseNet201, InceptionResNetV2) trained on 5 different cross-validation k-folds predicted 10 label images (corresponding to each segmentation contour: hard palate, basion, opisthion, C1 ~ C7) for each of the 100 predicted X-ray images, resulting in a total of 15,000 images. The predicted segmentation masks were evaluated against their respective ground truth masks using the segmentation metrics python package [17]. This package utilizes volume and distance-based methods to compute voxel-based metrics such as dice (F-1), precision, recall, as well as distance-based metrics like Hausdorff distance.

The diagnostic metrics (MG, Powers ratio, BDI, BAI, SAC, cSVA, CL) were automatically measured using python scripts of computer vision algorithms (OpenCV) on the predicted segmentation masks. Procedures for measuring diagnostic metrics are provided in Appendix S1. Visual guidelines for measuring the diagnostic metrics are shown in Fig. 2. To determine the agreeability of the measurements using automatic and manual methods, the metrics automatically measured using OpenCV python scripts were compared to those manually measured by graduate students trained by medical doctors using mean-squared error, Pearson's correlation coefficient, and paired *t*-test with *P* values (significant at $P < 0.05$) adjusted for false discovery rate using the Holm-Hochberg procedure. Linear regression scatter plots assessing the relationship between manually measured metric with automatically measured metric were also made for each diagnostic metric obtained with the three segmentation models.

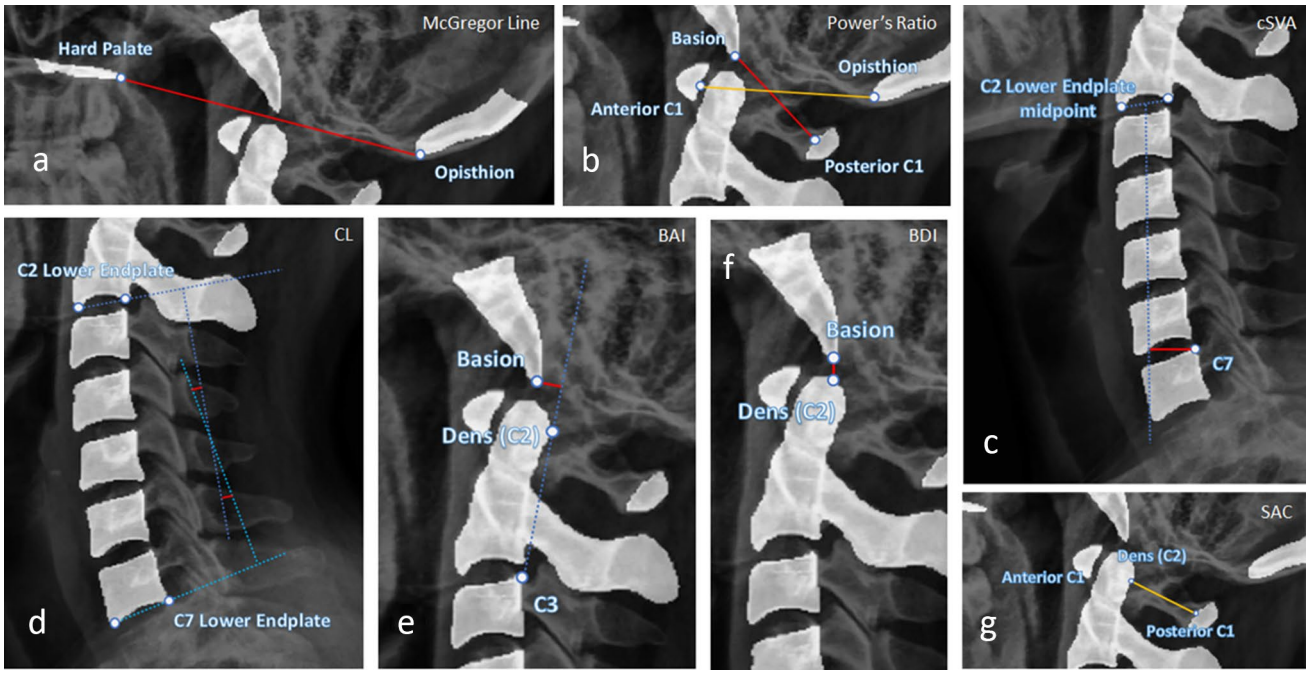

Fig. 2 Visual guidelines for measuring diagnostic metrics. Dots represent the specific points used to measure each respective metric both manually and automatically using OpenCV algorithms. **a** McGregor line: length of the red line. **b** Powers ratio: ratio of the lengths of the red and orange line. **c** Cervical sagittal vertical axis: length of the red line. **d** Cervical lordosis: angle between two dotted lines that extend from the lines drawn from the C2 lower endplate and the C7 lower endplate. **e** Basion-atlas interval: length of the red-line. **f** Basion-dens interval: length of the red line. **g** SAC: length of the orange line

Results

The demographics of patients with each group (normal, preoperative, postoperative) are shown in Table 1. The average age of the subjects was 55.7 (± 14.1) years, with 589 males and 472 females. Around 210 images were excluded due to bones in X-rays that were obfuscated or attached. A flowchart of the exclusion and inclusion procedure is shown in Fig. 3.

The segmentation metrics, dice, Jaccard, precision, recall, and Hausdorff distance (HD) measured through the segmentation metrics python package for each segmented mask class (hard palate, basion, opisthion, C1 ~ C7) obtained using the three segmentation models (EfficientNetB4, DenseNet201, InceptionResNetV2) are shown in Table 2. The dice coefficient for the hard palate region was 0.69 ± 0.10, 0.69 ± 0.10, and 0.68 ± 0.11 across the models, respectively, while for the basion region, it remained high at an average of 0.81 ± 0.07. Opisthion and the cervical vertebrae regions C1 through C7 showed excellent model agreement, with dice coefficients ranging from 0.93 ± 0.02 to 0.96 ± 0.01, indicating robust segmentation performance in these regions. The Jaccard index and precision metrics followed similar trends, with the Jaccard index averaging 0.54 ± 0.12 for the hard palate and 0.69 ± 0.09 for the basion across models. Precision remained high, particularly in the cervical regions, averaging 0.97 ± 0.02. Recall rates were slightly more variable, with an average of 0.62 ± 0.16 for the hard palate, improving to 0.96 ± 0.02 for the cervical regions. The Hausdorff distance showed more variability, with larger values observed for the hard palate (11.54 ± 13.23) and opisthion (13.39 ± 8.92), while the cervical regions C1 through C7 demonstrated considerably lower average distances.

The comparison between manually measured diagnostic metrics, MG, BDI, BAI, Powers, ADI, SAC, cSVA, and CL using manually delineated segmentations and automatically measured diagnostic metrics using predicted segmentations

Table 1 Demographics of patients in each subgroup

Characteristic	Normal	Postoperative
No. of patients	954	108
Sex		
M	509	96
F	445	22
Age (years)	55.4 ± 14.2	60.3 ± 11.3
Age range	26–97	27–86

Age (years) are mean ± standard deviation

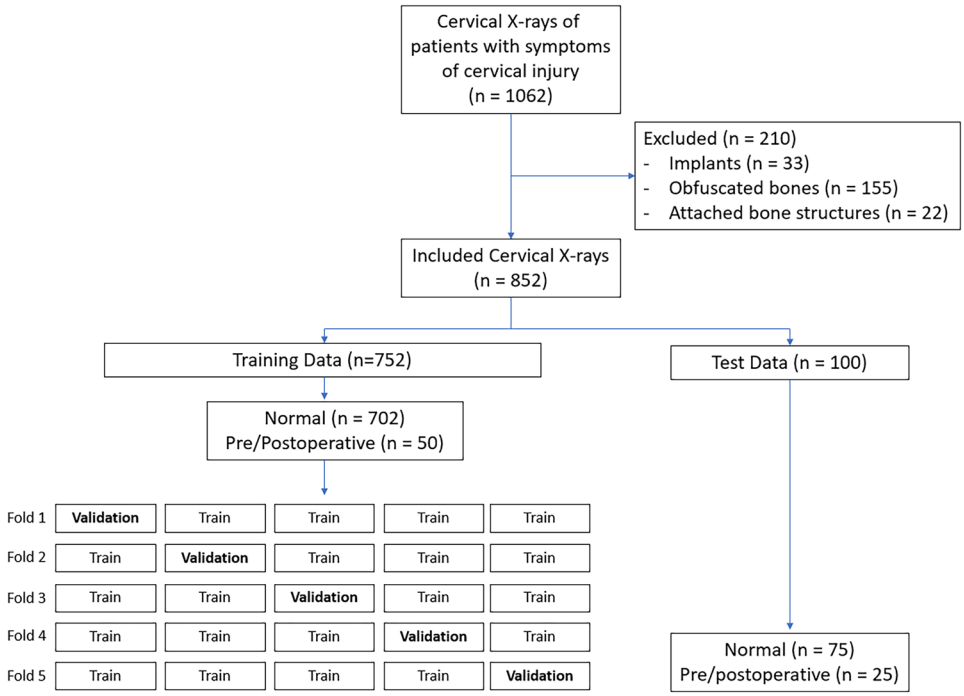

Fig. 3 Flowchart of X-rays that were included and excluded from the study, as well as how the included X-rays were randomly divided into training, validation, and testing datasets

is shown in Table 3. For MG, models EfficientNetB4, DenseNet201, InceptionResNetV2 demonstrated strong correlations ($r=0.90$, 0.88, 0.89) between predicted and manual values with a MSE of 48.33, 63.46, and 56.26, respectively, and no significant difference was found post-adjustment (adjusted $P=1.78$, 0.60, 0.68). BDI revealed significant difference in predictions with EfficientNetB4, DenseNet201, and InceptionResNetV2 ($P=0.02$, 0.01, 0.02), but showed no significance post-adjustment ($P=0.19$, 0.09, 0.12). However, the three models exhibited varying correlations from $r=0.72$ (DenseNet201) to $r=0.58$ (InceptionResNetV2). While the three models exhibited no significant differences for BAI (adjusted $P=1.79$, 1.75, 1.34 for EfficientNetB4, DenseNet201, InceptionResNetV2), correlation of BAI was lower for InceptionResNetV2 ($r=0.66$) than with EfficientNetB4 ($r=0.75$) and DenseNet201 ($r=0.74$). For Powers ratio, there was no significant difference in each segmentation model, ($P=1.78$, 1.75, 1.47, for EfficientNetB4, DenseNet201, InceptionResNetV2), but EfficientNetB4 showed higher correlation ($r=0.66$) than in DenseNet201 ($r=0.60$) and in InceptionResNetV2 ($r=0.59$). The SAC metric consistently exhibited high correlation across all models ($r=0.95$, 0.93, 0.95 for EfficientNetB4, DenseNet201, InceptionResNetV2) and non-significant P-values for all three models ($P=1.79$, 1.75, 1.47), indicating the segmentation models' strong predictive capability for SAC. Similarly, cSVA showed high correlation across all models ($r=0.99$ for all three models) and non-significant P-values ($P=1.79$, 1.75, 1.47 for EfficientNetB4, DenseNet201, InceptionResNetV2), nearing perfect agreement. CL showed high correlation for DenseNet201 ($r=0.91$) over EfficientNetB4 ($r=0.87$) and InceptionResNetV2 ($r=0.86$) and showed non-significant P-values for all three models ($P=1.79$, 1.75, 1.47). Linear regression scatter plots detailing the correlation between manual metrics and automatically measured metrics are shown in Fig. 4.

Discussion

In this study, we utilized U-Net neural networks with EfficientNet-B4, DenseNet201, and InceptionResNetV2 backbones trained on X-rays of the cervical spine for multiclass segmentation of the hard palate, basion, opisthion, and cervical spine [8, 13]. Additionally, we automatically measured metrics used in the diagnosis of cervical spine injuries and evaluation of cervical surgery outcomes with the results from the multiclass segmentation obtained using the three architectures. To our knowledge, no previous study has evaluated the performance of a multiclass cervical spine and craniofacial segmentation model then automatically measured diagnostic metrics, MG, BDI, BAI, Powers ratio, SAC, cSVA, and CL, using the predicted segmentations. Automated methods of measuring diagnostic metrics for cervical spine injuries as presented in this study can be a valuable tool for evaluating the cervical spine more promptly and consistently.

The lower average dice coefficients observed in the three segmentation models for the craniofacial multiclass segmentations, hard palate (0.69), basion (0.81), and opisthion

Table 2 Computed segmentation metrics of each U-Net multiclass segmentation model predicted segmentation mask compared with manually segmented ground truths

	Hard palate	Basion	Opisthion	C1	C2	C3	C4	C5	C6	C7
EfficientNetB4										
Dice coefficient	0.69 ± 0.1	0.81 ± 0.06	0.71 ± 0.1	0.93 ± 0.02	0.96 ± 0.01	0.95 ± 0.06	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.02
Jaccard	0.54 ± 0.12	0.69 ± 0.09	0.56 ± 0.12	0.87 ± 0.03	0.93 ± 0.02	0.91 ± 0.07	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.91 ± 0.03
Precision	0.82 ± 0.11	0.83 ± 0.12	0.82 ± 0.14	0.93 ± 0.04	0.96 ± 0.02	0.96 ± 0.06	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.03
Recall	0.62 ± 0.15	0.82 ± 0.11	0.65 ± 0.15	0.93 ± 0.03	0.96 ± 0.02	0.94 ± 0.06	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.03	0.94 ± 0.03
Hausdorff distance	11.36 ± 12.43	7.85 ± 3.85	13.19 ± 7.64	1.34 ± 0.49	2.48 ± 0.95	1.71 ± 1.9	1.32 ± 1.06	1.35 ± 0.51	2.08 ± 13.39	1.9 ± 1.83
DenseNet201										
Dice coefficient	0.69 ± 0.1	0.81 ± 0.07	0.7 ± 0.1	0.93 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.01
Jaccard	0.54 ± 0.12	0.68 ± 0.09	0.55 ± 0.12	0.87 ± 0.03	0.93 ± 0.02	0.93 ± 0.01	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.91 ± 0.03
Precision	0.82 ± 0.11	0.83 ± 0.12	0.82 ± 0.14	0.94 ± 0.03	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.03
Recall	0.63 ± 0.16	0.81 ± 0.12	0.65 ± 0.15	0.93 ± 0.03	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.02	0.95 ± 0.03
Hausdorff distance	12.11 ± 20.12	8.29 ± 6.74	14.05 ± 11.58	1.36 ± 1.45	2.46 ± 0.96	1.26 ± 0.44	1.24 ± 0.38	1.29 ± 0.44	1.51 ± 1.67	2.01 ± 3.1
InceptionResNetV2										
Dice coefficient	0.68 ± 0.11	0.81 ± 0.07	0.71 ± 0.1	0.93 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.02
Jaccard	0.53 ± 0.13	0.69 ± 0.1	0.56 ± 0.12	0.87 ± 0.03	0.93 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.91 ± 0.03
Precision	0.82 ± 0.11	0.83 ± 0.12	0.83 ± 0.14	0.93 ± 0.04	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.02
Recall	0.61 ± 0.16	0.81 ± 0.12	0.66 ± 0.15	0.93 ± 0.03	0.96 ± 0.02	0.95 ± 0.02	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.03	0.95 ± 0.03
Hausdorff distance	11.15 ± 7.15	7.75 ± 4.1	12.93 ± 7.53	1.33 ± 0.51	2.51 ± 1.34	1.32 ± 0.58	1.57 ± 6.47	1.37 ± 1.17	1.49 ± 0.57	1.88 ± 1.92
Model averages										
Dice coefficient	0.69 ± 0.1	0.81 ± 0.07	0.71 ± 0.1	0.93 ± 0.02	0.96 ± 0.01	0.96 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.02
Jaccard	0.54 ± 0.12	0.69 ± 0.09	0.56 ± 0.12	0.87 ± 0.03	0.93 ± 0.02	0.92 ± 0.03	0.93 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.91 ± 0.03
Precision	0.82 ± 0.11	0.83 ± 0.12	0.82 ± 0.14	0.93 ± 0.04	0.96 ± 0.02	0.97 ± 0.03	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.03
Recall	0.62 ± 0.16	0.82 ± 0.12	0.65 ± 0.15	0.93 ± 0.03	0.96 ± 0.02	0.95 ± 0.03	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.03	0.95 ± 0.03
Hausdorff distance	11.54 ± 13.23	7.96 ± 4.9	13.39 ± 8.92	1.35 ± 0.82	2.49 ± 1.08	1.43 ± 0.97	1.38 ± 2.64	1.34 ± 0.71	1.69 ± 5.21	1.93 ± 2.29

Data presented as mean ± standard deviation; Hausdorff distance in millimeters

Table 3 Comparison of averages, mean squared error (MSE), Pearson's correlation coefficient, and paired *t*-test (significance at $P<0.05$) of manually and automatically measured diagnostic metrics

		MG	BDI	BAI	Powers ratio	SAC	cSVA	CL
EfficientNetB4	Predict averages	159.58 ± 13.22	11.43 ± 2.56	9.39 ± 3.92	0.70 ± 0.06	36.38 ± 4.17	33.36 ± 16.54	14.01 ± 8.19
	Manual averages	156.90 ± 15.02	9.78 ± 3.49	9.20 ± 3.80	0.68 ± 0.07	35.96 ± 4.25	32.04 ± 16.36	14.75 ± 8.00
	MSE	48.33	8.03	8.19	< 0.01	1.97	6.64	17.17
	Correlation	0.90	0.65	0.75	0.66	0.95	0.99	0.87
	P-value	0.26	0.02	0.95	0.25	0.45	0.49	0.59
	Adjusted *P*-value	1.78	0.19	1.79	1.78	1.79	1.79	1.79
DenseNet201	Predict averages	160.23 ± 13.20	10.99 ± 3.15	9.08 ± 3.97	0.69 ± 0.05	36.44 ± 4.23	33.91 ± 17.05	14.26 ± 8.41
	Manual averages	156.90 ± 15.02	9.78 ± 3.49	9.20 ± 3.80	0.68 ± 0.07	35.96 ± 4.25	32.04 ± 16.36	14.75 ± 8.00
	MSE	63.46	7.71	7.90	< 0.01	2.68	7.75	12.55
	Correlation	0.88	0.72	0.74	0.60	0.93	0.99	0.91
	P-value	0.10	0.01	0.83	0.35	0.44	0.43	0.68
	Adjusted *P*-value	0.60	0.09	1.75	1.75	1.75	1.75	1.75
InceptionResNetV2	Predict averages	159.96 ± 12.94	10.87 ± 2.41	9.04 ± 3.65	0.69 ± 0.06	36.42 ± 4.13	33.82 ± 17.20	13.41 ± 8.00
	Manual averages	156.90 ± 15.02	9.78 ± 3.49	9.20 ± 3.80	0.68 ± 0.07	35.96 ± 4.25	32.04 ± 16.36	14.75 ± 8.00
	MSE	56.26	9.14	9.47	< 0.01	1.89	7.36	18.62
	Correlation	0.89	0.58	0.66	0.59	0.95	0.99	0.86
	P-value	0.11	0.02	0.88	0.27	0.49	0.47	0.37
	Adjusted *P*-value	0.68	0.12	1.34	1.47	1.47	1.47	1.47
Model averages	Predict averages	159.92 ± 0.27	11.10 ± 0.24	9.17 ± 0.16	0.69 ± 0.00	36.41 ± 0.02	33.70 ± 0.24	13.89 ± 0.36
	Manual averages	156.90 ± 15.02	9.78 ± 3.49	9.20 ± 3.80	0.68 ± 0.07	35.96 ± 4.25	32.04 ± 16.36	14.75 ± 8.00
	MSE	56.02	8.29	8.52	< 0.01	2.18	7.25	16.11
	Correlation	0.89	0.65	0.72	0.62	0.94	0.99	0.88

Data presented as mean ± standard deviation for predict, manual averages

MG McGregor's Line, *BDI* basion-dens interval, *BAI* basion-atlas interval, *SAC* space-available-cord, *cSVA* cervical sagittal vertical axis, *CL* cervical lordosis

(0.71) can be attributed to the challenges in delineating masks with superimposed boundaries in X-ray images compared to the well-defined boundaries of the cervical vertebrae (C1–C7) [18]. The segmentation performance of these craniofacial bones, particularly in terms of recall (average of 0.62, 0.82, 0.65 for hard palate, basion, opisthion across three models), was affected due to the inherent challenges of delineating masks with indistinct boundaries. However, the X-rays showed visibly distinct boundaries in the landmarks of each craniofacial bones that were used for measuring diagnostic metrics, such as MG, BDI, BAI, and Powers ratio as shown in Fig. 5. It is likely that the deficits in the dice coefficient are mainly due to areas away from the landmarks, where the lack of well-defined boundaries makes consistent segmentation difficult, both manually and automatically. Conversely, the cervical vertebrae (C1–C7) consistently demonstrated higher dice coefficients (0.93–0.97), benefiting from clear boundaries that facilitated accurate manual delineation as well as multiclass prediction.

The diagnostic cervical spine metrics measured using both manual and automated segmentations were evaluated for agreement using mean square errors (MSE), Pearson's correlation coefficient, and paired *t*-tests. As shown in Table 2 and Fig. 4, the cervical sagittal vertical axis (cSVA), which relies on points of reference from C2 and C7 vertebrae, exhibited the highest correlation coefficient of $r=0.99$ and coefficient of determination of $r^2=0.99$ across all three segmentation models, indicating a strong linear relationship between manual and automated measurements. Similarly, relatively fair correlations were found in other metrics obtained using C1–C7 vertebrae like SAC ($r=0.90$, 0.88, 0.89; $r^2=0.90$, 0.87, 0.90 for EfficientNetB4, DenseNet201, InceptionResNetV2) and CL ($r=0.87$, 0.91, 0.86; $r^2=0.76$, 0.83, 0.74) between the three segmentation models, likely due to the consistent and accurate segmentation performance of the C1–C7 vertebrae. Notably, the McGregor line (MG) demonstrated high correlation coefficient values across models, with $r=0.89$, $r^2=0.81$ for EfficientNetB4; 0.88, 0.77 for DenseNet201; and 0.89, 0.80 for InceptionResNetV2, despite the low dice coefficient of the hard palate (0.69) and opisthion (0.71). As shown in Fig. 5, the clear boundaries observed in the regions of the hard palate and the opisthion where points of reference were used to calculate the MG allowed for accurate MG measurements using the predicted

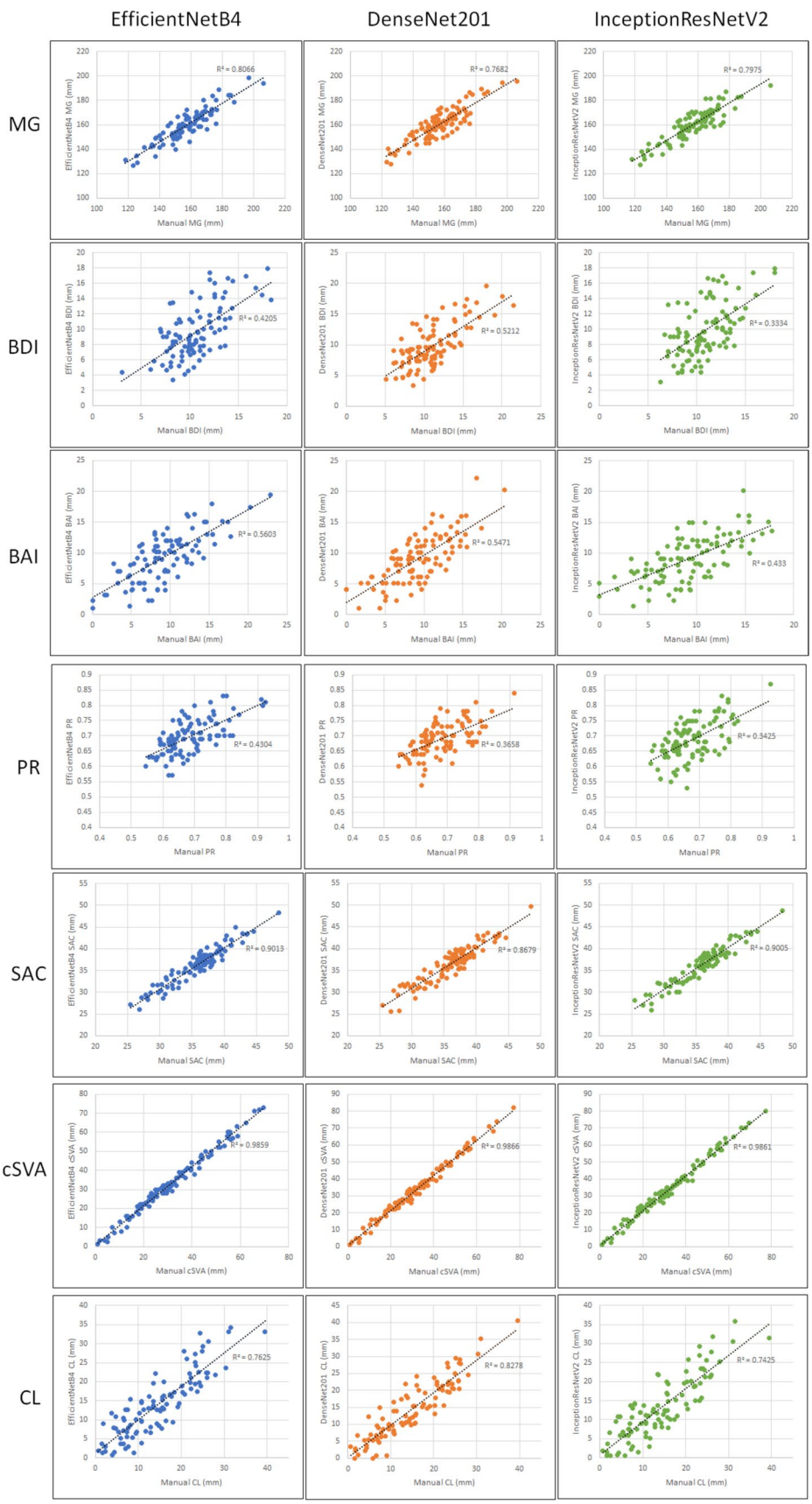

Fig. 4 Linear regression charts with coefficient of determination of manually measured metrics and automatically measured metrics across three models

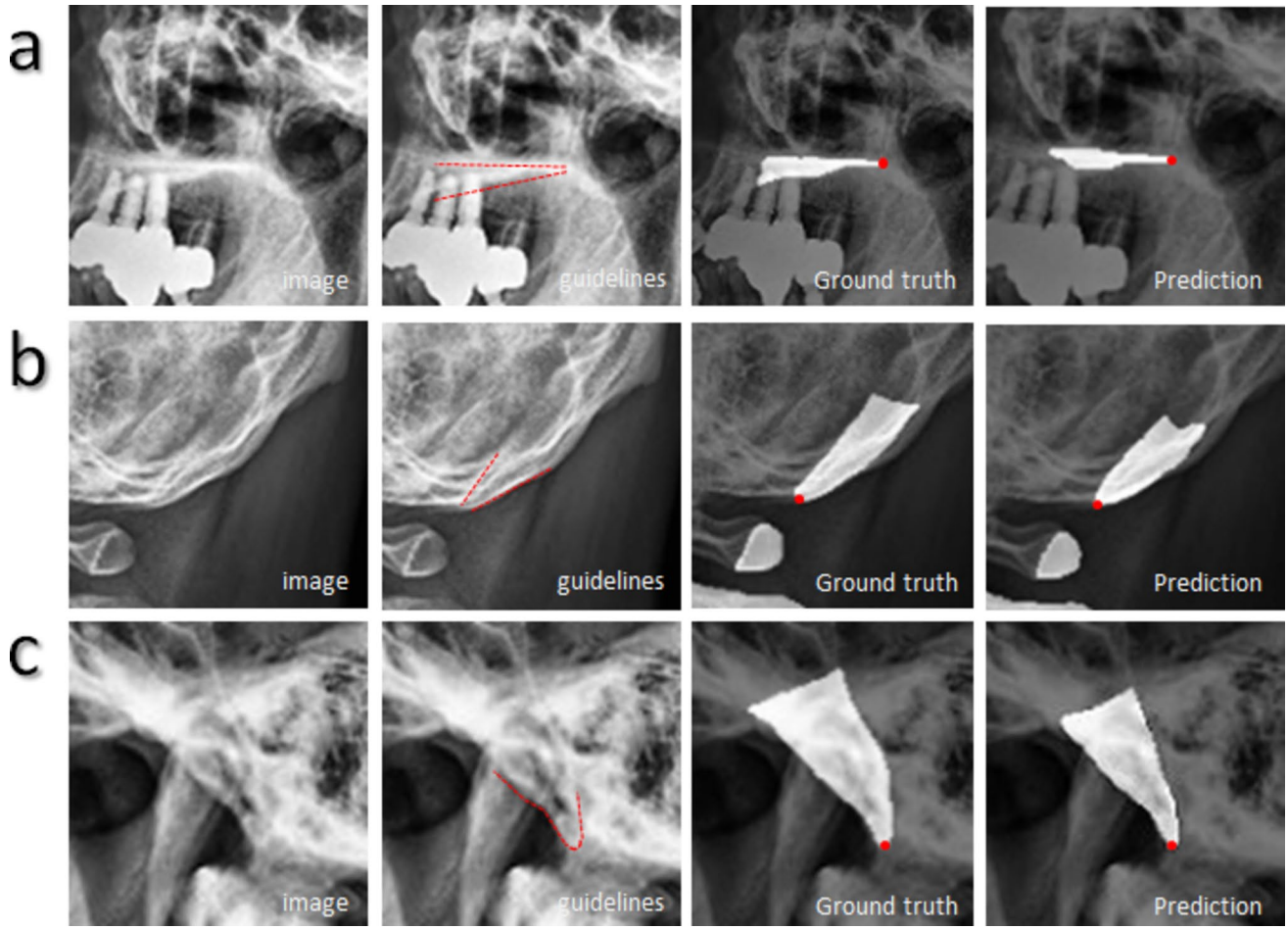

Fig. 5 Examples of craniofacial bone boundaries that can be outlined consistently. In many X-rays, there are guidelines (red dotted lines) that can be used to outline important areas of the craniofacial bones used to locate points (red dots) for measuring diagnostic metrics. However, as shown in ground truth and prediction images, there lacks a definite boundary for outlining the outer areas of each craniofacial bone, resulting in segmentation inconsistencies. Figure shows the image, guidelines used to outline boundaries, point of reference used to measure diagnostic metrics of the **a** hard palate, **b** opisthion, and **c** basion

segmentations. On the other hand, weaker correlation coefficients were observed for measurements obtained from points of reference using the basion mask, such as BDI ($r = 0.65$, $r^2 = 0.42$ for EfficientNetB4, 0.72, 0.52 for DenseNet201, 0.58, 0.33 for InceptionResNetV2), BAI ($r = 0.75$, 0.74, 0.66; $r^2 = 0.56$, 0.55, 0.43), and Powers ratio ($r = 0.66$, 0.60, 0.59; $r^2 = 0.43$, 0.37, 0.34). This weaker correlation is likely due to superimposition making boundaries used to obtain points of reference difficult to discern depending on the quality of X-rays as shown in Fig. 6. The diagnostic metrics obtained using manual and automatic methods were also compared for significant differences using paired *t*-test. All of the false-positive corrected *P* values for each metric comparison exceeded the significant value of 0.05 across all three models, suggesting that there were no statistically significant distinctions between the metrics acquired using manual and automatic methods. The consistent agreement of manual and automatic results shown through correlation and paired *t*-tests highlights the reliability of the proposed automatic method, affirming the clinical potential for the automated framework to aid in diagnosing cervical injuries and evaluating surgical outcomes.

There are several limitations to acknowledge for this study. Firstly, the segmentation performance of the craniofacial bones was relatively lower due to the lack of well-defined outer boundaries and challenges posed by obfuscation in the X-ray images, making it difficult to establish clear guidelines for segmentation. Further improvements can be made by defining the boundaries of the craniofacial bones, such as extending the outer boundaries to encompass the entire cranium or the edges of the image. Second, although the study utilized a diverse dataset with images of varying conditions, it is difficult to estimate the segmentation model's performance in settings with different

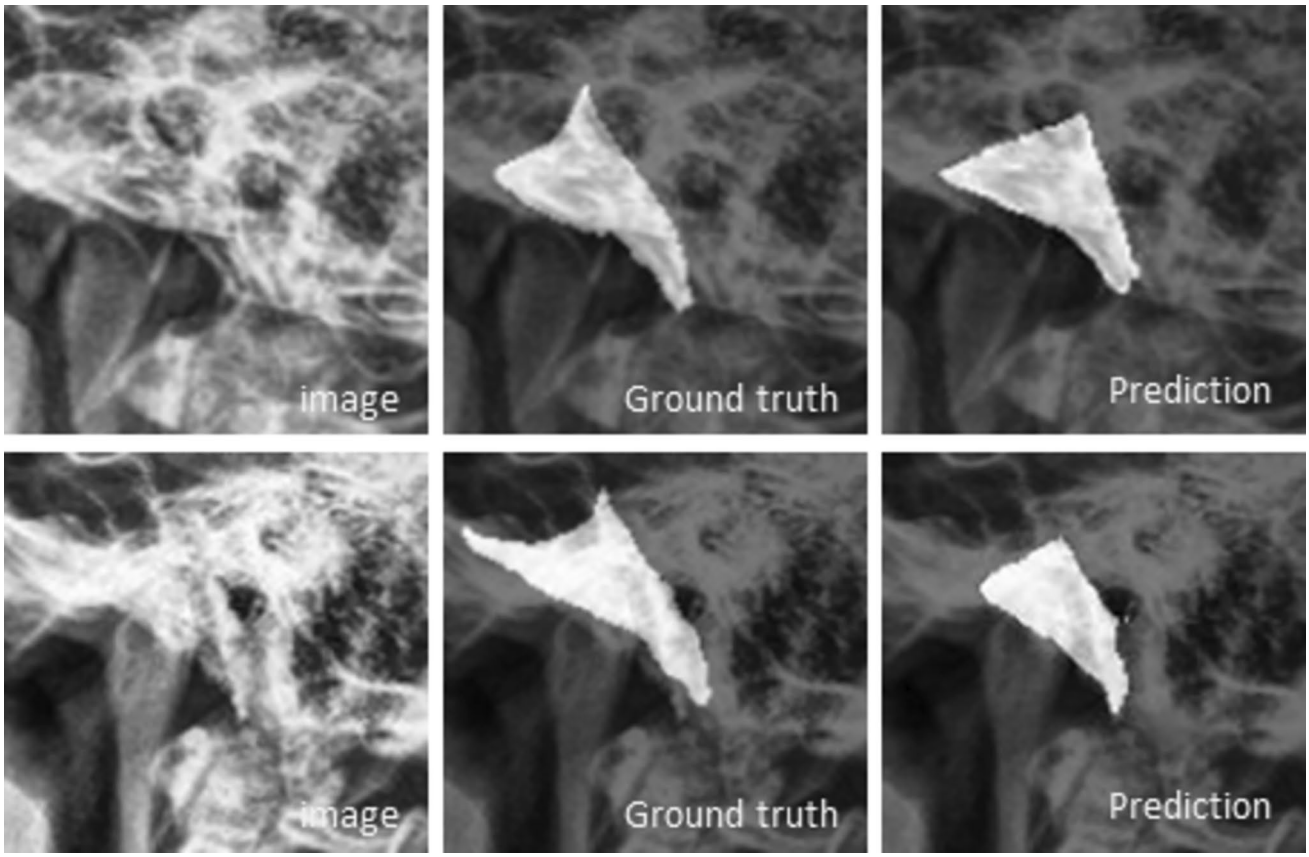

Fig. 6 Examples of obfuscated boundaries of the basion from bone superimposition that make consistent segmentation through both manual and automatic methods difficult

X-ray imaging machines and populations. A follow-up study utilizing cervical images from different sources can help the generality of the segmentation model. Third, the imbalanced dataset (954 normal, 108 pre/postoperative images), with an overrepresentation of normal subjects and a limited number of subjects with cervical injuries, may impact the model's real-world performance. Additionally, due to the low number of preoperative images, it was difficult to apply the measured metrics to clinically diagnose traumatic cervical injuries. To ensure the reliability and robustness of the segmentation model and its diagnostic implications, further validation studies using datasets with more preoperative cervical images are necessary.

In conclusion, the results of this study demonstrate the effectiveness of multiclass segmentation in accurately segmenting the cervical spine using X-rays and predicting diagnostic metrics. The high correlation coefficients, small MSE values, and no significant differences in any of the metrics indicated that the automated measurements closely aligned with manual measurements. The reliable performance of the segmentation model, especially in delineating the cervical vertebrae, highlights its potential as a valuable tool for assisting healthcare professionals in diagnosing cervical spine injuries and predicting surgical outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-01006-z.

Author Contributions Guarantors of Integrity of entire study, K.G.K., G.T.Y.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.H.S., W.S.K.; clinical studies, W.S.K., T.S.J.; experimental studies, J.H.S.; statistical analysis, J.H.S.; and manuscript editing, all authors.

Funding This work was supported by the GRRC program of Gyeonggi province (GRRC-Gachon2020(B01), AI-based Medical Image Analysis), and by Gachon University (GCU-202205980001).

Data Availability The models and code generated for this study are available from the corresponding author upon reasonable request.

Declarations

Ethics Approval This retrospective study protocol was approved by the institutional review board at Gachon University Gil Medical Center (GDIRB2022-190). Methods used in this study were all in accordance with the relevant guidelines and regulations of the declaration of Helsinki.

Consent to Participate Informed consent was waived by the institutional review board at Gachon University Gil Medical Center due to the retrospective nature of the study.

Competing Interests The authors declare no competing interests.

References

1. Cooper Z, Gross JA, Lacey JM, Traven N, Mirza SK, Arbabi S. Identifying survivors with traumatic craniocervical dissociation: a retrospective study. Journal of Surgical Research. 2010 May 1;160(1):3–8.
2. Kim YJ, Yoo CJ, Park CW, Lee SG, Son S, Kim WK. Traumatic atlanto-occipital dislocation (AOD). Korean Journal of Spine. 2012;9(2):85.
3. Joaquim AF, Schroeder GD, Vaccaro AR. Traumatic Atlanto-Occipital Dislocation—A Comprehensive Analysis of All Case Series Found in the Spinal Trauma Literature. International Journal of Spine Surgery. 2021 Aug 1;15(4):724-39.
4. Glaun GD, Phillips JH. Occipitocervical Dissociation in Three Siblings: A Pediatric Case Report and Review of the Literature. Journal of the American Academy of Orthopaedic Surgeons. Global Research & Reviews. 2018;2(5).
5. Yang SY, Boniello AJ, Poorman CE, Chang AL, Wang S, Passias PG. A review of the diagnosis and treatment of atlantoaxial dislocations. Global spine journal. 2014 Aug;4(3):197–210.
6. Al Arif SM, Knapp K, Slabaugh G. Fully automatic cervical vertebrae segmentation framework for X-ray images. Computer methods and programs in biomedicine. 2018 Apr 1;157:95–111.
7. Rehman F, Shah SI, Gilani SO, Emad D, Riaz MN, Faiza R. A novel framework to segment out cervical vertebrae. In2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE) 2019 Mar 6 (pp. 190–194). IEEE.
8. Yakubovskiy PI. Segmentation Models. 2019. GitHub repository. Available from: https://github.com/qubvel/segmentation_models. Accessed August 21, 2022.
9. Lee HJ, Hong JT, Kim IS, Kwon JY, Lee SW. Analysis of measurement accuracy for craniovertebral junction pathology: most reliable method for cephalometric analysis. Journal of Korean Neurosurgical Society. 2013 Oct 31;54(4):275-9.
10. Singh AK, Fulton Z, Tiwari R, Zhang X, Lu L, Altmeyer WB, Tantiwongkosi B. Basion–Cartilaginous Dens Interval: An Imaging Parameter for Craniovertebral Junction Assessment in Children. American Journal of Neuroradiology. 2017 Dec 1;38(12):2380-4.
11. Rojas CA, Bertozzi JC, Martinez CR, Whitlow J. Reassessment of the craniocervical junction: normal values on CT. American journal of neuroradiology. 2007 Oct 1;28(9):1819-23.
12. Benke M, Yu WD, Peden SC, O'Brien JR. Occipitocervical junction: imaging, pathology, instrumentation. Am J Orthop. 2011 Oct 1;40(10):E205-15.
13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234–241). Springer, Cham.
14. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. InInternational conference on machine learning 2019 May 24 (pp. 6105–6114). PMLR.
15. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. InProceedings of the AAAI conference on artificial intelligence 2017 Feb 12 (Vol. 31, No. 1).
16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 4700–4708).
17. Jia J. A package to compute segmentation metrics: seg-metrics. 2020. https://github.com/Ordgod/segmentation_metrics. https://doi.org/10.5281/zenodo.3995075.
18. Izzetti R, Nisi M, Aringhieri G, Crocetti L, Graziani F, Nardi C. Basic knowledge and new advances in panoramic radiography imaging techniques: A narrative review on what dentists and radiologists should know. Applied Sciences. 2021 Aug 26;11(17):7858.