Imaging Science in Dentistry 2024; 54: 81-91
https://doi.org/10.5624/isd.20230245

Deep learning-based automatic segmentation of the mandibular canal on panoramic radiographs: A multi-device study

Moe Thu Zar Aung[1,2], Sang-Heon Lim[3], Jiyong Han[3], Su Yang[4], Ju-Hee Kang[5], Jo-Eun Kim[1], Kyung-Hoe Huh[1], Won-Jin Yi[1,3,4,*], Min-Suk Heo[1,*], Sam-Sun Lee[1]

[1]*Department of Oral and Maxillofacial Radiology, School of Dentistry and Dental Research Institute, Seoul National University, Seoul, Korea*
[2]*Department of Oral Medicine, University of Dental Medicine, Mandalay, Myanmar*
[3]*Interdisciplinary Program in Bioengineering, Graduate School of Engineering, Seoul National University, Seoul, Korea*
[4]*Department of Applied Bioengineering, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea*
[5]*Department of Oral and Maxillofacial Radiology, Seoul National University Dental Hospital, Seoul, Korea*

ABSTRACT

Purpose: The objective of this study was to propose a deep-learning model for the detection of the mandibular canal on dental panoramic radiographs.

Materials and Methods: A total of 2,100 panoramic radiographs (PANs) were collected from 3 different machines: RAYSCAN Alpha (n = 700, PAN A), OP-100 (n = 700, PAN B), and CS8100 (n = 700, PAN C). Initially, an oral and maxillofacial radiologist coarsely annotated the mandibular canals. For deep learning analysis, convolutional neural networks (CNNs) utilizing U-Net architecture were employed for automated canal segmentation. Seven independent networks were trained using training sets representing all possible combinations of the 3 groups. These networks were then assessed using a hold-out test dataset.

Results: Among the 7 networks evaluated, the network trained with all 3 available groups achieved an average precision of 90.6%, a recall of 87.4%, and a Dice similarity coefficient (DSC) of 88.9%. The 3 networks trained using each of the 3 possible 2-group combinations also demonstrated reliable performance for mandibular canal segmentation, as follows: 1) PAN A and B exhibited a mean DSC of 87.9%, 2) PAN A and C displayed a mean DSC of 87.8%, and 3) PAN B and C demonstrated a mean DSC of 88.4%.

Conclusion: This multi-device study indicated that the examined CNN-based deep learning approach can achieve excellent canal segmentation performance, with a DSC exceeding 88%. Furthermore, the study highlighted the importance of considering the characteristics of panoramic radiographs when developing a robust deep-learning network, rather than depending solely on the size of the dataset. ***(Imaging Sci Dent 2024; 54: 81-91)***

KEY WORDS: Mandibular Canal; Panoramic Radiography; Deep Learning; Artificial Intelligence

Introduction

The mandibular canal houses the inferior alveolar artery, vein, and nerve,[1] the last of which provides motor innervation to the facial muscles and sensory innervation to the lower teeth, chin, and lower lip. Injury to the inferior alveolar nerve can lead to partial numbness or a total loss of sensation in the lower lip, tongue, chin, and buccal mucosa.[2,3] To prevent such complications, it is necessary to precisely determine the location of the mandibular canal, which encases the nerve within a thin layer of cortical bone.[4,5]

To identify this anatomical structure, radiographic im-

Moe Thu Zar Aung and Sang-Heon Lim contributed equally to this work as first authors.

This work was supported by the Korea Medical Device Development Fund grant, funded by the Korean government (the Ministry of Science and ICT; the Ministry of Trade, Industry and Energy; the Ministry of Health & Welfare; and the Ministry of Food and Drug Safety) (Project Number: 1711194231, RS-2023-KD000011; Project Number: 1711174552, KMDF_PR_20200901_0147; and Project Number: 1711196792, RS-2023-00253380). This work was also supported by the National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIT) (No. 2023R1A2C200532611).
Received November 13, 2023; Revised January 9, 2024; Accepted January 10, 2024
Published online February 22, 2024

*Correspondence to : Prof. Min-Suk Heo
Department of Oral and Maxillofacial Radiology, School of Dentistry, Seoul National University. 101 Daehak-ro, Jongno-Gu, Seoul 03080, Korea
Tel) 82-2-6256-3051, E-mail) hmslsh@snu.ac.kr
Prof. Won-Jin Yi
Department of Oral and Maxillofacial Radiology, School of Dentistry and Dental Research Institute, Seoul National University, Seoul, Korea
Tel) 82-2-6256-3052, E-mail) wjyi@snu.ac.kr

Imaging Science in Dentistry · pISSN 2233-7822 eISSN 2233-7830

aging techniques are necessary. Panoramic radiography, in particular, is readily accessible and can be used to detect and locate anatomical structures using straightforward methods at a comparatively low cost.[6] Furthermore, it can produce images of a quality sufficient for most dental radiographic needs and enable a holistic assessment of the maxillomandibular complex while exposing patients to relatively low levels of ionizing radiation.[6,7]

Manual detection of the path of the mandibular canal is time-consuming and labor-intensive. Panoramic radiographs typically display low contrast due to the inclusion of excessive non-target tissue during the image reconstruction process, which can impact inter-observer variability.[8] Consequently, an automated mandibular canal segmentation system must be developed to alleviate the burden on radiologists.

In efforts to address these limitations, several recent studies have been published regarding deep learning-based canal segmentation. Previous research has applied deep learning methods to automatically segment the mandibular canal on panoramic dental images. These studies have been focused on various aspects, including ambiguity classification, visualization of impacted third molars, and panoptic segmentation of the mandibular canal along with other structures using deep learning.[9-11]

However, these investigations have not considered the mental foramen region during training and/or have not employed datasets representing over 2,000 participants.[9-11] The mental foramen is a critical anatomical landmark that must be accurately identified and preserved to avoid complications.[12] Consequently, incorporating the mental foramen into mandibular canal segmentation tasks could greatly contribute to diagnosis and analysis.

To ensure the development of a reliable deep-learning model, it is essential not only to construct a substantial set of image data but also to train the network with a diverse array of panoramic dental images. The shape, size, and contrast of the mandibular canal as observed on images can vary due to factors such as the detector of the imaging device, post-processing techniques, and other related variables. Consequently, it is important to classify the types of imaging devices used to acquire the training dataset and to analyze their impact on network performance. In consideration of these characteristics, several multicenter studies have focused on generalizing deep learning networks.[13-15] These studies have demonstrated the potential applicability of deep learning networks in clinical practice.

However, to the best of the authors' knowledge, no multi-device research has been conducted on the use of deep learning for canal segmentation on dental panoramic radiographs. Consequently, the aim of this study was to develop an automated method for segmenting the mandibular canal on panoramic radiographs obtained with various devices using deep learning techniques. The primary contributions of this research are twofold, as it represents: 1) a study carried out using data from 3 distinct devices, including an analysis of network performance across these devices to enhance the generalizability and robustness of the method, and 2) an investigation into an appropriate training strategy for a deep learning network, aimed at increasing the reliability of an artificial intelligence-based approach for automated canal segmentation on dental panoramic radiographs.

Materials and Methods

The study received approval from the institutional review board (IRB) of Seoul National University Dental Hospital, Dental Life Science Research Institute, based on the results of their deliberation (IRB No. ERI23015).

Data acquisition and preprocessing

A total of 2,100 panoramic images from patients who visited Seoul National University Dental Hospital between January 2021 and February 2023 were collected and categorized into 3 groups: PAN A consisted of panoramic radiographs from 350 male and 350 female patients, acquired using the RAYSCAN Alpha machine (Ray Corp, Seoul, Korea); PAN B comprised panoramic radiographs from another set of 350 male and 350 female patients, obtained with the OP-100 device (Imaging Instrumentarium, Tuusula, Finland); and PAN C included panoramic radiographs from a final group of 350 male and 350 female patients, captured using the CS8100 machine (Carestream Dental, Atlanta, GA, USA).

The target population of this study was strictly limited to patients with permanent dentition. Consequently, the age of the participants ranged from 18 to 40 years. The exclusion criteria omitted panoramic radiographs of patients who had received atypical treatments, including partial or complete mandibulectomy, orthognathic surgery, and reparative procedures in the posterior region of the mandible. Additionally, patients with class II and III impacted mandibular teeth were not considered in this analysis. Furthermore, individuals with a history of mandibular fractures, cystic lesions or tumors in the mandible, bone diseases such as osteomyelitis or osteoporosis, or syndromic conditions were excluded due to the potential for compromised visibility of the entire

Table 1. Different characteristics of panoramic dental imaging devices.

Parameters	PAN A	PAN B	PAN C
Manufacturer	RAYSCAN Alpha	OP-100	CS8100
Tube voltage	60-90 kVp	57-85 kVp	60-90 kVp
Tube current	4-17 mA	2-16 mA	2-15 mA
Irradiation time	14 s	16.8-17.6 s	2-12.5 s
Frequency	60-220 kHz	75-150 kHz	140 kHz

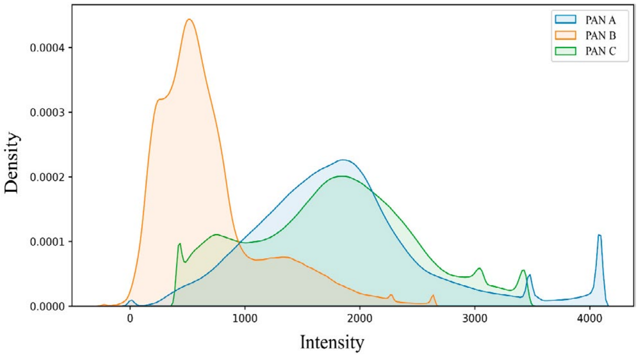

Fig. 1. Kernel density estimation plot for digital panoramic radiographs taken with 3 different imaging devices. PAN: panoramic radiograph, PAN A: RAYSCAN Alpha, PAN B: OP-100, PAN C: CS8100.

mandibular canal tract.

All panoramic radiographs were captured by expert radiographers, in accordance with the manufacturer's recommendations and in consideration of factors such as the patient's body size, jaw width (narrow or wide), degree of obesity, and other relevant characteristics. The RAYSCAN Alpha and CS8100 devices are digital radiography machines, while the OP-100 is a computed radiography device capable of acquiring digital images via Fuji Computed Radiography (Fuji Corp, Tokyo, Japan). The specifications of these 3 scanners are detailed in Table 1.

An experienced oral radiologist manually annotated panoramic radiographs for mandibular canal segmentation using a software application (3D Slicer for Windows 10, ver. 4.10.2; Massachusetts Institute of Technology, Cambridge, MA, USA).[16] The images were resampled to a resolution of 1024 × 512 (width × height) and normalized using min-max intensity scaling, with values ranging from 0 to 1. Additionally, a total of 2,100 panoramic radiographs were divided into 2 datasets: an internal dataset consisting of 1,800 images for network training and validation, and a test set comprising 300 images to assess the network's performance. The internal dataset included 600 images from each of the 3 groups. In a similar fashion, the test set contained 100 images from each group, totaling 300 images. Fig. 1 displays the kernel density estimation plot for all images ($n = 2{,}100$) used in this study.[17,18] The kernel density estimation plot reveals that the pixel intensities of PAN A and PAN C exhibited similar distributions, whereas PAN B displayed a markedly different pattern.

In this study, augmentation techniques were applied to the training datasets to enhance the generalizability of the network. The Albumentations augmentation method (ver. 1.3.0; https://sourceforge.net/projects/albumentations.mirror/files/1.3.0/) was used to introduce random transformations to the training set. The parameters for these transformations were as follows: a 15% probability of a horizontal flip, a 50% probability of rotation within a range of −10° to 10°, a 50% probability of shifting contrast by 0% to 30%, and a 5% probability of applying Gaussian blur. These data augmentation strategies were applied consistently to each deep learning process in the study.

Study design and network architecture

In this study, a U-Net-based architecture was employed for mandibular canal segmentation. Four distinct network architectures—ResNet50, ResNet152, SEResNet152, and EfficientNetB4—were adapted to the backbone of U-Net.[19-21] The objective was to identify the most relevant backbone for canal segmentation. To select an appropriate network architecture, 5-fold cross-validation was performed on a dataset comprising 1,800 internal images. U-Net-based semantic segmentation networks designed for canal segmentation were utilized for deep learning analysis. These networks were sourced from open-source code available at https://github.com/qubvel/segmentation. The network architecture consisted of an encoder followed by a decoder, each with 5 resolution steps. A 2-dimensional convolution kernel was used to construct the networks. Fig. 2B illustrates a representative U-Net-based architecture, while the actual network architecture employed in the experiments was a modified version with alterations to the U-Net encoder structure.

Network optimization was achieved through the joint minimization of both binary focal loss and Dice loss, utilizing a batch size of 8:

$$Loss_{Dice} = \frac{2 \times \sum p_{True} \times p_{Pred}}{\sum p^2{}_{True} + \sum p^2{}_{red}},$$

$$Loss_{Focal} = -\alpha \times p_{True} \times (1 - p_{Pred})^{\gamma} \log(p_{Pred}) - (1 - p_{True}) \times \alpha \times p_{Pred}{}^{\gamma} \log(1 - p_{Pred})$$ [22]

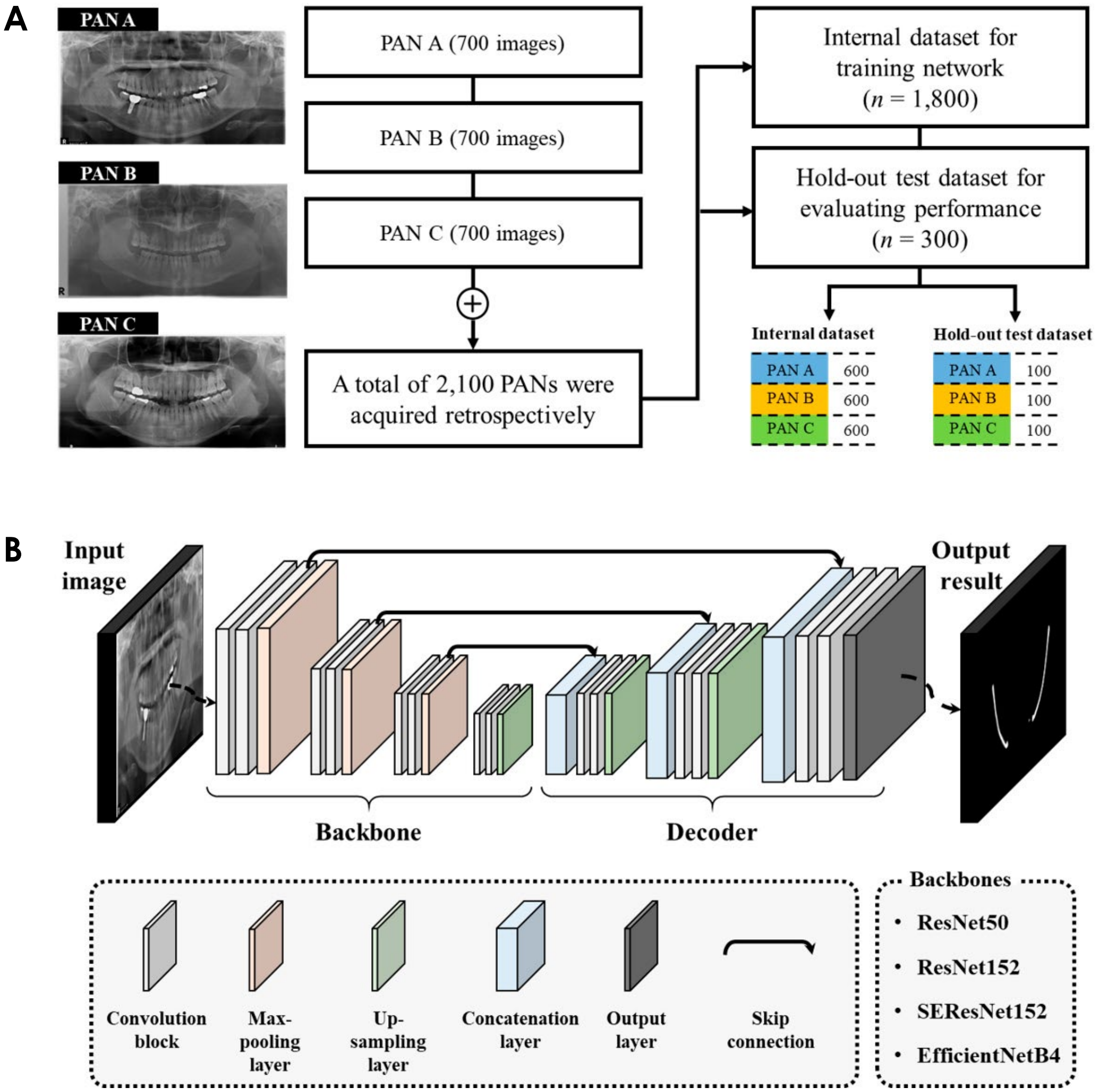

Fig. 2. A. Data enrollment criteria. B. Representative architecture of U-Net. The actual network architecture employed in the experiments was a modified version, with changes made to the encoder structure of U-Net. PAN: panoramic radiograph, PAN A: RAYSCAN Alpha, PAN B: OP-100, PAN C: CS8100.

Here, p_{True} represents the pixel value of the ground truth, while p_{Pred} denotes the network's inference result (probability) for the corresponding pixel. α is the weighting factor in balanced cross-entropy, and γ is the focusing parameter used for the smooth adjustment of weights. The values for α and γ were set to 0.25 and 2.0, respectively. Notably, both α and γ are empirically determined hyper-parameters in the binary focal loss equation.

The deep learning networks were implemented using Python3, employing Keras with a TensorFlow (ver. 2.10.0; https://github.com/tensorflow/tensorflow/releases/tag/v2.10.0) backend, and were run on an NVIDIA RTX A6000 graphics processing unit (48 GB; NVIDIA, Santa Clara, CA, USA). Training of the networks was conducted with an Adam optimizer. The initial learning rate was set at 0.001 and was decreased by a factor of 0.2 every 5 epochs upon reaching a plateau, over a total of 200 epochs, with a batch size of 8.

Evaluation of network performance for segmentation

In this study, SPSS (version 26.0; IBM Corp., Armonk, NY, USA) was utilized for statistical analysis, specifically employing the Friedman test. A significance level of $P < 0.05$ was considered to indicate statistical significance. To evaluate the performance of automated segmentation, the precision, recall, and Dice similarity coefficient (DSC) were calculated using the Scikit-learn Python library (version 1.2.2; https://scikit-learn.org/1.2/) and the Python programming language (version 3.9.16; https://www.python.org/downloads/release/python-3916/). The network was assessed on a pixel-wise basis, applying the following equations:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN}$$

The rates of true positives, false positives, true negatives, and false negatives were calculated at the pixel level. Network predictions were deemed positive when the probability exceeded 0.5.

Results

Table 2 presents the assessment outcomes for the 4 distinct backbones integrated into the U-Net-based architecture. The images from the test set were excluded from this analysis, as the network was trained and evaluated solely using the internal dataset. To assess performance across the entire internal dataset, which comprised 1,800 images, 5-fold cross-validation was conducted. The dataset was partitioned into training, validation, and test sets in a 6:2:2 ratio for each group of images. The U-Net with the EfficientNetB4 backbone exhibited the best segmentation performance, achieving precision, recall, and DSC metrics of 84.0%, 84.2%, and 83.9%, respectively. Consequently, EfficientNetB4 was chosen as the backbone for the subsequent multi-device study.

The validation set was generated by randomly allocating 30% of the images from the training set during the network training phase. For example, when the training dataset contained 600, 1200, and 1800 images, then the corresponding validation sets consisted of 180, 360, and 540 images, respectively. In contrast, the test set comprised a constant number of 300 images. In this multi-device study, 7 types of training datasets were introduced, representing all possible combinations of the 3 scanners. The networks trained with these 7 training datasets fell into 3 categories: 1) single-device networks; 2) dual-device networks; and 3) a multi-device network. The single-device network was trained exclusively with images from 1 scanner, either PAN A, PAN B, or PAN C, with the training set containing 600 images. The dual-device network combined datasets from 2 different scanners, utilizing all 3 possible pairings of the 2 groups for training sets and resulting in 1200 images per set (n = 1,200). Finally, the multi-device network was trained using a comprehensive dataset that included images from all 3 scanners, totaling 1,800 images.

Table 3 presents the evaluation metrics for 3 single-device networks. The networks trained with PAN A, PAN B, and PAN C achieved average DSC values of 75.3%, 65.7%, and 85.4%, respectively. These findings demonstrate that employing PAN C for network training resulted in the best segmentation performance, with average precision, recall, and DSC values of 85.8%, 85.1%, and 85.4%, respectively.

The evaluation metrics for the dual-device networks are presented in Table 4. When compared to the training method that utilized only single groups, the dual-device networks demonstrated a significant improvement in performance, as reflected by the segmentation metrics. The networks trained using the combinations of PAN A and PAN B, PAN A and PAN C, and PAN B and PAN C achieved average DSC values of 87.9%, 87.8%, and 88.4%, respectively. Moreover, the network trained with a dataset combining PAN B and PAN C outperformed those trained with the other 2 combinations of image groups, achieving precision, recall, and DSC values of 90.3%, 86.8%, and 88.4%, respectively. The dual-device networks exhibited superior performance to the single-device networks. However, these

Table 2. Evaluation metrics from 4 backbones for mandibular canal segmentation

Backbone	ResNet50	ResNet152	SEResNet152	EfficientNetB4
Precision	0.793 ± 0.090 (0.784-0.802)	0.826 ± 0.089 (0.817-0.835)	0.820 ± 0.087 (0.811-0.829)	0.840 ± 0.081 (0.836-0.844)
Recall	0.804 ± 0.098 (0.794-0.815)	0.819 ± 0.086 (0.810-0.828)	0.839 ± 0.078 (0.831-0.847)	0.842 ± 0.074 (0.839-0.845)
DSC	0.800 ± 0.083 (0.787-0.804)	0.820 ± 0.079 (0.812-0.829)	0.827 ± 0.074 (0.820-0.835)	0.839 ± 0.070 (0.836-0.843)

DSC: dice similarity coefficient

Table 3. Evaluation results for the single-device cohort networks

Training dataset	Hold-out validation	Test dataset			
		PAN A (n=100)	PAN B (n=100)	PAN C (n=100)	Average (n=300)
PAN A (n = 600)	Precision	0.760 ± 0.075 (0.745-0.775)	0.690 ± 0.075 (0.664-0.715)	0.766 ± 0.092 (0.748-0.785)	0.739 ± 0.107 (0.726-0.751)
	Recall	0.821 ± 0.078 (0.805-0.836)	0.693 ± 0.130 (0.667-0.719)	0.797 ± 0.099 (0.777-0.816)	0.770 ± 0.118 (0.757-0.784)
	DSC	0.788 ± 0.073 (0.774-0.803)	0.690 ± 0.126 (0.665-0.715)	0.781 ± 0.093 (0.762-0.799)	0.753 ± 0.109 (0.740-0.765)
PAN B (n = 600)	Precision	0.683 ± 0.118 (0.659-0.706)	0.656 ± 0.118 (0.633-0.680)	0.679 ± 0.136 (0.652-0.706)	0.673 ± 0.125 (0.658-0.687)
	Recall	0.673 ± 0.123 (0.648-0.697)	0.664 ± 0.130 (0.638-0.690)	0.600 ± 0.131 (0.574-0.626)	0.646 ± 0.132 (0.631-0.661)
	DSC	0.677 ± 0.118 (0.653-0.700)	0.659 ± 0.122 (0.635-0.683)	0.635 ± 0.130 (0.610-0.661)	0.657 ± 0.124 (0.643-0.671)
PAN C (n = 600)	Precision	0.883 ± 0.052 (0.873-0.893)	0.838 ± 0.083 (0.821-0.854)	0.852 ± 0.075 (0.838-0.867)	0.858 ± 0.074 (0.849-0.866)
	Recall	0.889 ± 0.061 (0.876-0.901)	0.815 ± 0.097 (0.780-0.834)	0.849 ± 0.087 (0.831-0.866)	0.851 ± 0.088 (0.841-0.861)
	DSC	0.885 ± 0.054 (0.875-0.896)	0.825 ± 0.086 (0.808-0.842)	0.850 ± 0.079 (0.834-0.866)	0.854 ± 0.078 (0.845-0.862)

DSC: dice similarity coefficient

findings suggest that the greater network performance cannot be attributed solely to the increased volume of training data; the performance also varied according to the distinct characteristics of the data in the training set.

Table 5 presents the evaluation metrics for the multi-device network. This network attained the highest performance in mandibular canal segmentation among the 7 different networks evaluated, achieving an average precision of 90.6%, recall of 87.4%, and DSC of 88.9%. Although the multi-device network surpassed all dual-device networks in overall performance, it demonstrated marginally lower results compared to the 2 types of dual-device networks when assessed on the test set corresponding to PAN B.

Fig. 3 displays the automated segmentation results of the 7 independent networks for the 3 panoramic radiograph samples in the test set. In the networks trained on a single device, errors frequently occurred along the trajectory of the mandibular canal. Conversely, in networks trained on dual- or multi-device setups, errors were predominantly localized to the areas surrounding the mental foramen or mandibular foramen.

Fig. 4 presents box plots depicting the DSC metrics for the 7 networks, representing all possible combinations of the 3 groups. The Friedman test was conducted to analyze differences in the DSC metric distributions across the networks, while the Bonferroni correction was applied to assess 2-sided statistical significance. The statistical significance values across the 7 networks are presented in Table 6. The Friedman test revealed no statistically significant differences among the dual- and multi-device networks.

Discussion

As a cutting-edge machine learning method, deep learning has garnered substantial attention in clinical research.[9] This approach has also gained traction in dental and maxillofacial radiology, with a growing body of research employing deep learning techniques for the localization and segmentation of the bilateral mandibular canals, each housing an inferior alveolar artery, vein, and nerve. However, several previous studies of mandibular canal segmentation using deep learning have encountered limitations with generalizability, as the networks were typically trained and evaluated on patient data from a single imaging device. To address this issue, the present study incorporated network

Table 4. Evaluation metrics for dual-device cohort networks

Training dataset	Hold-out validation	Test dataset PAN A (n = 100)	 PAN B (n = 100)	 PAN C (n = 100)	Average (n = 300)
PAN A and B (n = 1200)	Precision	0.904 ± 0.060 (0.892-0.916)	0.895 ± 0.062 (0.882-0.907)	0.898 ± 0.082 (0.882-0.914)	0.899 ± 0.069 (0.891-0.907)
	Recall	0.868 ± 0.068 (0.855-0.881)	0.860 ± 0.070 (0.846-0.874)	0.855 ± 0.089 (0.837-0.873)	0.861 ± 0.076 (0.852-0.870)
	DSC	0.885 ± 0.061 (0.873-0.897)	0.877 ± 0.064 (0.864-0.889)	0.876 ± 0.084 (0.859-0.892)	0.879 ± 0.071 (0.871-0.887)
PAN A and C (n = 1200)	Precision	0.914 ± 0.039 (0.907-0.922)	0.875 ± 0.075 (0.860-0.890)	0.893 ± 0.066 (0.880-0.906)	0.894 ± 0.064 (0.887-0.901)
	Recall	0.889 ± 0.058 (0.877-0.900)	0.830 ± 0.100 (0.810-0.850)	0.875 ± 0.076 (0.860-0.890)	0.865 ± 0.084 (0.855-0.874)
	DSC	0.901 ± 0.047 (0.891-0.910)	0.851 ± 0.086 (0.834-0.868)	0.884 ± 0.069 (0.870-0.898)	0.878 ± 0.072 (0.870-0.887)
PAN B and C (n = 1200)	Precision	0.915 ± 0.049 (0.905-0.925)	0.889 ± 0.005 (0.879-0.900)	0.906 ± 0.064 (0.893-0.918)	0.903 ± 0.057 (0.897-0.910)
	Recall	0.884 ± 0.068 (0.870-0.897)	0.872 ± 0.074 (0.857-0.887)	0.847 ± 0.075 (0.833-0.862)	0.868 ± 0.074 (0.859-0.876)
	DSC	0.898 ± 0.058 (0.887-0.910)	0.880 ± 0.061 (0.868-0.892)	0.875 ± 0.067 (0.862-0.889)	0.884 ± 0.063 (0.877-0.892)

DSC: dice similarity coefficient

Table 5. Evaluation results for multi-device cohort network

Training dataset	Hold-out validation	Test dataset PAN A	 PAN B	 PAN C	Average
PAN A, B, and C (n = 1800)	Precision	0.921 ± 0.044 (0.913-0.930)	0.893 ± 0.055 (0.882-0.904)	0.904 ± 0.066 (0.891-0.917)	0.906 ± 0.057 (0.900-0.913)
	Recall	0.889 ± 0.044 (0.880-0.898)	0.858 ± 0.071 (0.844-0.872)	0.873 ± 0.069 (0.860-0.872)	0.874 ± 0.064 (0.866-0.881)
	DSC	0.905 ± 0.042 (0.896-0.913)	0.875 ± 0.061 (0.862-0.887)	0.888 ± 0.066 (0.875-0.901)	0.889 ± 0.059 (0.883-0.896)

DSC: dice similarity coefficient

training and evaluation using a dataset from multiple devices, constituting a key step toward assessing the generalizability and clinical utility of deep learning methods.

This study introduced a multi-device approach for training a deep learning network, utilizing 2,100 panoramic radiographs obtained from 3 different imaging devices. A U-Net architecture with an EfficientB4Net backbone was used to develop networks for mandibular canal segmentation. These networks were optimized using the 7 possible combinations of the 3 distinct groups of data. Additionally, the networks were evaluated using a hold-out test dataset. The results indicated that the network trained on the combined datasets from all 3 groups (n = 1,800) outperformed the other networks, achieving an average precision of 90.6%, recall of 87.3%, and DSC of 88.9%. Notably, this study's multi-device methodology leveraged panoramic radiographs with diverse characteristics. Deep learning methods are highly dependent on the training set, and a more narrowly defined training dataset can lead to increased bias due to the characteristics of the data used during training.

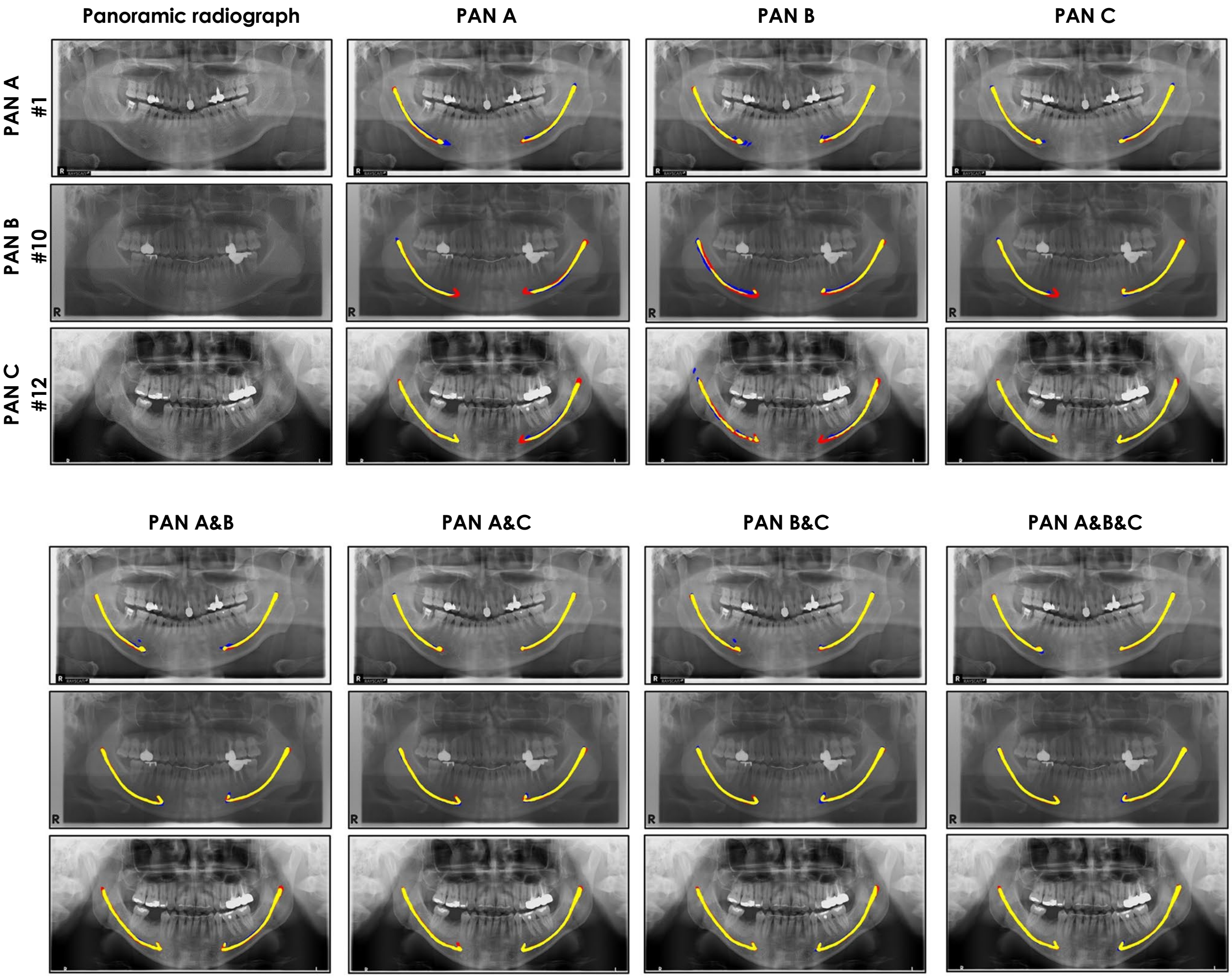

Fig. 3. Canal segmentation results for 3 examples of panoramic dental radiographs, with the combinations of groups used as training sets indicated above each column. The red, blue, and yellow regions denote false negatives, false positives, and true positives, respectively. PAN: panoramic radiograph, PAN A: RAYSCAN Alpha, PAN B: OP-100, PAN C: CS8100.

Moreover, for deep learning networks to be successfully applied in real-world clinical settings, it is crucial to use datasets that reflect a wide range of imaging characteristics.

In recent years, several studies have utilized a range of deep learning architectures to develop automated networks for mandibular canal segmentation on panoramic dental images. To the best of the authors' knowledge, however, no deep learning-based studies have been conducted to establish a semantic segmentation network for panoramic radiographs in a multi-device context. In a study focused on developing a mandibular canal segmentation network through ambiguity classification, the same imaging equipment and radiation parameters—tube voltage, tube current, and exposure time—were used for network training.[9] That study employed a dataset of 1,366 panoramic radiographs from a single device and achieved an average DSC of 85.7% for mandibular canal segmentation. Another study described panoptic segmentation of the mandibular canal and 6 other structures using a deep neural network applied to panoramic radiographs.[11] The network was trained with 51 panoramic radiographs from a single device, achieving an average intersection-over-union value of 63.9%. These 2 prior studies have limitations relative to the present research: 1) they relied on panoramic radiographs from a single imaging device, potentially limiting the generalizability of their findings to other devices; and 2) the number of radiographs used was relatively small, which may have affected the statistical power and reliability of their results. Additionally, another prior study proposed a transfer learning-based method for canal segmentation. This approach used cropped patches from panoramic images that included the region of the impacted third molar and the adjacent

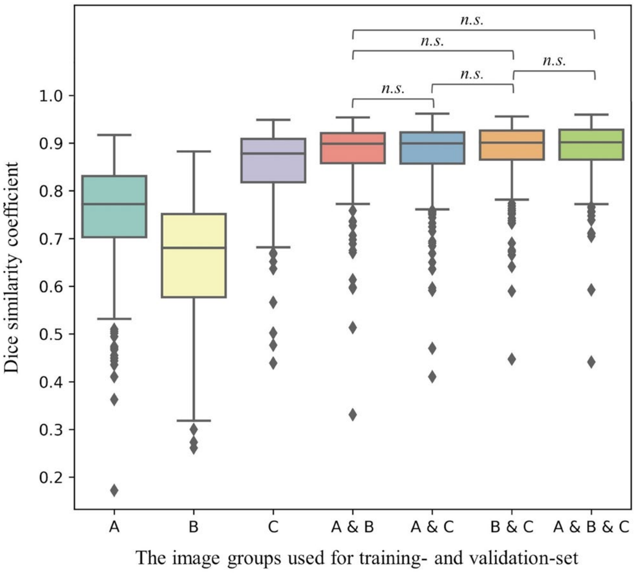

Fig. 4. Box plots of Dice similarity coefficient metrics. The x-axis indicates the groups of internal datasets included in the training set. Symbols above the plots signify the absence of significant differences between the 2 networks. The boxes in the graph represent the quartiles of the Dice similarity coefficient metrics, with the horizontal line inside each box indicating the median. n.s.: not significant, PAN A: RAYSCAN Alpha, PAN B: OP-100, PAN C: CS8100.

mandibular canal, collected from 2 different hospitals. The network was trained on a dataset of 2,940 patches from both institutions. The trained network achieved an average DSC of 85.7% for canal segmentation. The strength of that study lies in its large dataset and inclusion of data from 2 medical centers, which likely improved the generalizability of the trained network. However, unlike the present study, it used only specific regions of interest from panoramic images for network training. This required additional preprocessing to extract and prepare these regions for training, which could be viewed as a limitation of the method. Nevertheless, a potential limitation of the present study is the composition of the multicenter dataset, which consisted exclusively of panoramic radiographs from a single type of scanner. This could have introduced bias into the network and impacted its generalizability.

The present statistical analysis encompassed 7 different networks, each trained with 1 of the 7 types of training sets that comprised all possible combinations of image groups. Performance evaluation of the single-device networks revealed that the network trained with images from PAN C exhibited the best performance, followed by the networks trained on PAN A and PAN B images. This discrepancy in

Table 6. The Friedman test results. The asymptotic significances of the seven networks were computed to assess the differences in their performance.

Group − Group (Training dataset)	Test statistic	Standard error	Standard test statistic	Significance	Adjusted significance
B − A	.703	.176	3.988	.000	.001
B − C	− 2.697	.176	− 15.289	.000	.000
B − A and B	− 3.683	.176	− 20.883	.000	.000
B − A and C	− 3.797	.176	− 21.525	.000	.000
B − B and C	− 3.913	.176	− 22.187	.000	.000
B − A and B and C	− 4.130	.176	− 23.415	.000	.000
A − C	− 1.993	.176	− 11.301	.000	.000
A − A and B	− 2.980	.176	− 16.895	.000	.000
A − A and C	− 3.093	.176	− 17.538	.000	.000
A − B and C	− 3.210	.176	− 18.199	.000	.000
A − A and B and C	− 3.427	.176	− 19.427	.000	.000
C − A and B	− .987	.176	− 5.594	.000	.000
C − A and C	− 1.100	.176	− 6.236	.000	.000
C − B and C	− 1.217	.176	− 6.898	.000	.000
C − A and B and C	− 1.433	.176	− 8.126	.000	.000
A and B − A and C	− .113	.176	− .643	.521	1.000
A and B − B and C	− .230	.176	− 1.304	.192	1.000
A and B − A and B and C	− .447	.176	− 2.532	.011	.238
A and C − B and C	− .117	.176	− .661	.508	1.000
A and C − A and B and C	− .333	.176	− 1.890	.059	1.000
B and C − A and B and C	− .217	.176	− 1.228	.219	1.000

performance can likely be attributed to the fact that PAN B displayed a significantly different intensity distribution from the other groups, as shown in Fig. 1. These differences in intensity distribution among the image groups are believed to have influenced the results. Moreover, the network trained with images from PAN C outperformed the model trained on images from PAN A, despite the absence of marked differences in image intensity distribution and radiation conditions between them. The superior performance of the model trained on PAN C may be due to the post-processing techniques used by each vendor, which can introduce subtle variations in the radiographs. Regarding the dual-device networks, the network trained with a dataset that included PAN B demonstrated superior performance. This finding aligns with the results from the networks trained on single image groups, where the inclusion of PAN C in the training dataset resulted in the highest average DSC. Therefore, the combination of PAN B and PAN C constituted the optimal training dataset among the 3 dual-device networks.

The average DSC metrics indicated that the multi-device network outperformed the others in the segmentation of the mandibular canal. This network demonstrated enhanced performance with the test sets for PAN A and PAN C when compared to the dual-device networks. However, a slight decline was observed in the average DSC for the test set corresponding to PAN B. Given that the images from PAN A and PAN C shared similar intensity distributions, which differed from those of PAN B, it is believed that the multi-device network may exhibit a bias toward images from PAN A and PAN C. Additionally, the Friedman test revealed no statistically significant difference in performance between the dual-device networks (mean DSC: 87.9%, 87.8%, and 88.4%) and the multi-device network (mean DSC: 88.9%). These results suggest that training a network with images that share similar characteristics may lead to a bias in segmentation performance for those specific images. This underscores the importance of developing robust networks that account for not only variations across multiple centers and devices, but also the distinct characteristics of image data obtained with various imaging devices.

This study had some limitations. First, the annotations were derived from a single examiner. Although an experienced rater conducted the manual delineation, the potential exists for bias due to individual subjectivity and opinion. Furthermore, optimal canal segmentation cannot be fully investigated using only the U-Net-based network, despite its use with 4 different backbones. Future studies employing state-of-the-art networks should be conducted to explore the optimal network architecture. Second, despite the successful application of the dual- and multi-device networks for accurate mandibular canal segmentation, the number of false positives reported by the network remains a substantial challenge. These errors were typically observed in and around the region of the mental foramen. Therefore, it may be assumed that a deep learning-based approach, utilizing a boundary- and continuity-aware training strategy for mandibular canal segmentation, could address these issues.[23] Third, the evaluation datasets used in this study were limited to a single ethnic group, ensuring no demographic differences, and the patients' ages ranged from 18 to 40 years. Consequently, future research should aim to validate the auto-segmentation model on datasets that include a broader range of ages and ethnic diversity.

In conclusion, this study indicates that the examined deep learning method demonstrated reliable performance for mandibular canal segmentation on panoramic radiographs from multiple devices, achieving a DSC greater than 88%. Furthermore, the multi-device study revealed that developing a robust network requires the acquisition of a dataset that captures the diverse characteristics of panoramic radiographs, rather than simply increasing the volume of training data for deep learning.

Conflicts of Interest: None

References

1. Iwanaga J, Ibaragi S, Takeshita Y, Asaumi J, Horner K, Gest TR, et al. Mandibular canal versus inferior alveolar canal: a Delphi study. Clin Anat 2021; 34: 1095-100.
2. Agbaje JO, Van de casteele E, Hiel M, Verbaanderd C, Lambrichts I, Politis C. Neuropathy of trigeminal nerve branches after oral and maxillofacial treatment. J Maxillofac Oral Surg 2016; 15: 321-7.
3. Kushnerev E, Yates JM. Evidence-based outcomes following inferior alveolar and lingual nerve injury and repair: a systematic review. J Oral Rehabil 2015; 42: 768-802.
4. Ai CJ, Jabar NA, Lan TH, Ramli R. Mandibular canal enlargement: clinical and radiological characteristics. J Clin Imaging Sci 2017; 7: 28.
5. Jung YH, Cho BH. Radiographic evaluation of the course and visibility of the mandibular canal. Imaging Sci Dent 2014; 44: 273-8.
6. Shah N, Bansal N, Logani A. Recent advances in imaging technologies in dentistry. World J Radiol 2014; 6: 794-807.
7. Mallya S, Lam E. White and Pharoah's oral radiology: principles and interpretation. 8th ed. St. Louis: Elsevier, 2018.
8. Puciło M, Puciło A, Safranow K, Nowicka A. The influence of age, sex, and tooth type on the anatomical relationship between tooth roots and the mandibular canal. Imaging Sci Dent 2021; 51: 373-82.

9. Yang S, Li A, Li P, Yun Z, Lin G, Cheng J, et al. Automatic segmentation of inferior alveolar canal with ambiguity classification in panoramic images using deep learning. Heliyon 2023; 9: e13694.
10. Ariji Y, Mori M, Fukuda M, Katsumata A, Ariji E. Automatic visualization of the mandibular canal in relation to an impacted mandibular third molar on panoramic radiographs using deep learning segmentation and transfer learning techniques. Oral Surg Oral Med Oral Pathol Oral Radiol 2022; 134: 749-57.
11. Cha JY, Yoon HI, Yeo IS, Huh KH, Han JS. Panoptic segmentation on panoramic radiographs: deep learning-based segmentation of various structures including maxillary sinus and mandibular canal. J Clin Med 2021; 10: 2577.
12. Hadilou M, Gholami L, Ghojazadeh M, Emadi N. Prevalence and extension of the anterior loop of the mental nerve in different populations and CBCT imaging settings: a systematic review and meta-analysis. Imaging Sci Dent 2022; 52: 141-53.
13. Grøvik E, Yi D, Iv M, Tong E, Nilsen LB, Latysheva A, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study. NPJ Digit Med 2021; 4: 33.
14. Kiljunen T, Akram S, Niemelä J, Löyttyniemi E, Seppälä J, Heikkilä J, et al. A deep learning-based automated ct segmentation of prostate cancer anatomy for radiation therapy planning - a retrospective multicenter study. Diagnostics (Basel) 2020; 10: 959.
15. Vesal S, Gayo I, Bhattacharya I, Natarajan S, Marks LS, Barratt DC, et al. Domain generalization for prostate segmentation in transrectal ultrasound images: a multi-center study. Med Image Anal 2022; 82: 102620.
16. Fedorov A, Beichel R, Kalpathy cramer J, Finet J, Fillion robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012; 30: 1323-41.
17. Soleimanpour N, Mohammadi M. Probabilistic load flow by using nonparametric density estimators. IEEE Trans Power Syst 2013; 28: 3747-55.
18. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020; 11: 4294.
19. Targ S, Almeida D, Lyman K. Resnet in resnet: generalizing residual architectures [Internet]. arXiv 2016; 1603.08029 [cited 2023 Nov 10]. Available from https://arxiv.org/abs/1603.08029.
20. Tan M, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks [Internet]. arXiv 2020; 1905.11946v5 [cited 2023 Nov 10]. Available from https://arxiv.org/abs/1905.11946.
21. Hu J, Shen L, Albaqnie S, Sun G, Wu E. Squeeze-and-excitation networks [Internet]. arXiv 2019; 1709.01507v4 [cited 2023 Nov 10]. Available from https://arxiv.org/abs/1709.01507.
22. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020; 42: 318-27.
23. Jeoun BS, Yang S, Lee SJ, Kim TI, Kim JM, Kim JE, et al. Canal-Net for automatic and robust 3D segmentation of mandibular canals in CBCT images using a continuity-aware contextual network. Sci Rep 2022; 12: 13460.