Comparison of 3 deep learning neural networks for classifying the relationship between the mandibular third molar and the mandibular canal on panoramic radiographs

Motoki Fukuda, DDS,[a] Yoshiko Ariji, DDS, PhD,[a] Yoshitaka Kise, DDS, PhD,[a] Michihito Nozawa, DDS,[a] Chiaki Kuwada, DDS,[a] Takuma Funakoshi, DDS,[a] Chisako Muramatsu, PhD,[b] Hiroshi Fujita, PhD,[c] Akitoshi Katsumata, DDS, PhD,[d] and Eiichiro Ariji, DDS, PhD[a]

Objective. The aim of this study was to compare time and storage space requirements, diagnostic performance, and consistency among 3 image recognition convolutional neural networks (CNNs) in the evaluation of the relationships between the mandibular third molar and the mandibular canal on panoramic radiographs.

Study Design. Of 600 panoramic radiographs, 300 each were assigned to noncontact and contact groups based on the relationship between the mandibular third molar and the mandibular canal. The CNNs were trained twice by using cropped image patches with sizes of 70 × 70 pixels and 140 × 140 pixels. Time and storage space were measured for each system. Accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) were determined. Intra-CNN and inter-CNN consistency values were calculated.

Results. Time and storage space requirements depended on the depth of CNN layers and number of learned parameters, respectively. The highest AUC values ranged from 0.88 to 0.93 in the CNNs created by 70 × 70 pixel patches, but there were no significant differences in diagnostic performance among any of the models with smaller patches. Intra-CNN and inter-CNN consistency values were good or very good for all CNNs.

Conclusions. The size of the image patches should be carefully determined to ensure acquisition of high diagnostic performance and consistency. (Oral Surg Oral Med Oral Pathol Oral Radiol 2020;130:336−343)

Panoramic radiography is one of the most common examinations for screening various lesions and conditions in the maxillofacial region. Cone beam computed tomography (CBCT) is recommended before mandibular third molar extraction when the third molar and mandibular canal are superimposed on panoramic images to minimize the risk of mandibular nerve damage during extraction.[1-9] However, determination of the relative positions of the third molar and the mandibular canal is sometimes difficult for inexperienced observers. Under these circumstances, there has recently been a surge in demand for computer-aided diagnosis (CAD) systems in the field of maxillofacial imaging in a push to ensure complete observations of the anatomy and to avoid overlooking critical diseases and conditions.[10,11]

Among several CAD techniques, deep learning (DL) systems using convolutional neural networks (CNNs) have received considerable attention in recent years.[12] CNNs have various functions, such as classification, object detection, and semantic segmentation.[13-15] Given an input image to be evaluated, the learning model for classification can yield the most appropriate output of several classes learned. Some CNNs for classification, such as AlexNet, GoogLeNet, and VGG-16, are freely available[16-18] in the Digits version 5.0 training system (Nvidia Corporation, Santa Clara, CA). These networks have won prizes in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC; https://arxiv.org/abs/1409.0575) and are commonly used in various fields. Research has addressed the application of CNNs to panoramic images.[19,20] However, no studies have offered a detailed comparison of their diagnostic performance and consistency with the same data sets. Such a comparison would clarify several characteristic features of CNNs and could contribute to appropriate network selection in future clinical applications and studies, although the performance and consistency would differ, depending on such factors as the size of the image patches.

Statement of Clinical Relevance

The deep learning technique appears to have potential for classifying the relationship between the mandibular third molar and the mandibular canal on panoramic radiographs and could be useful in minimizing the risk of nerve damage during extraction.

[a]Department of Oral and Maxillofacial Radiology, Aichi-Gakuin University School of Dentistry, Nagoya, Japan.
[b]Faculty of Data Science, Shiga University, Shiga, Japan.
[c]Department of Electrical, Electronic and Computer Faculty of Engineering, Gifu University, Gifu, Japan.
[d]Department of Oral Radiology, Asahi University School of Dentistry, Mizuho, Japan.
Received for publication Jan 8, 2020; returned for revision Mar 24, 2020; accepted for publication Apr 3, 2020.

2212-4403/$-see front matter
https://doi.org/10.1016/j.oooo.2020.04.005

The aims of the present study were to evaluate the diagnostic performance and consistency of 3 classification CNNs (AlexNet, GoogLeNet, and VGG-16) by using the same panoramic images, including the mandibular third molar and the mandibular canal, and to elucidate their differences, together with the effect of the size of the training image patches.

MATERIALS AND METHODS

Patients

Among approximately 6600 panoramic images obtained and stored in our hospital image database between December 2018 and May 2019, 2242 images were chosen by searching for the reference words "impacted tooth" on the imaging reports. Of these radiographs, 600 images were randomly selected by 2 radiologists (M.F. and E.A.), on the basis of the following criteria: (1) no contact or superimposition between the mandibular third molar and canal (300 images, which comprised the noncontact group); and (2) clear contact or superimposition (300 images, which comprised the contact group) (Figure 1). The radiologists were in complete agreement with regard to the relationship of the roots and canals in all 600 images; radiographs for which there was disagreement were excluded. These panoramic images were downloaded from the database in JPEG format with a matrix size of 1039 × 1378 pixels.

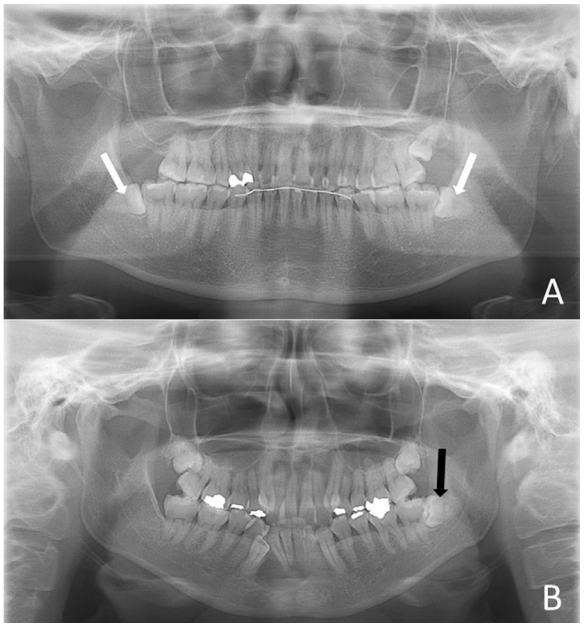

Fig. 1. Original panoramic radiographs before cropping. **A,** Radiograph from the noncontact group. The white arrows indicate third molars that exhibit no contact or superimposition between the roots and the mandibular canals. **B,** Radiograph from the contact group. The black arrow indicates a third molar that is in apparent contact with the canal.

All panoramic radiographs were made with the Veraviewepocs X550 PCR (J. Morita, Tokyo, Japan) with a tube voltage of 75 kV, tube current of 8 mA, and acquisition time of 16.2 seconds.

Preparation of data sets

Two sizes of image patches with square regions of interest of 70 × 70 and 140 × 140 pixels were cropped from the downloaded panoramic images. The center of the region of interest was placed at the area where the molar and the canal were situated most closely (Figure 2).

In total, 600 image patches were randomly classified into 400 training images, 100 validation images, and 100 test images. The training data set included 200 noncontact (class 0) and 200 contact (class 1) patches. The validation data set consisted of 50 noncontact and 50 contact patches. Data augmentation was performed on the training image patches by changing and adjusting image sharpness, brightness, and contrast by using image processing software (Irfan View version 4.44; http://www.Irfanview.com). This process is commonly used to increase the number of data sets for more effective training in the case of small size data sets. Consequently, the size of each training data set increased from 400 patches to 9000 patches.

The learning processes were performed twice for each CNN by using the same learning (training and validation) and test data sets. Two learning models were

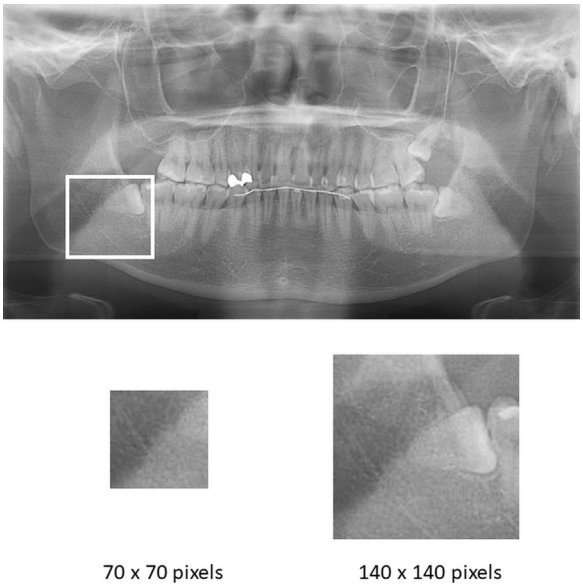

Fig. 2. Panoramic radiograph with a region of interest (ROI) indicated by the white box. The center of the ROI was placed at the area where the molar and canal were situated most closely. Cropping was performed in two different sizes: 70 × 70 pixels and 140 × 140 pixels from the 1039 × 1378 pixel panoramic image.

created for the 2 different sizes of image patches. AlexNet, GoogLeNet, and VGG-16 were trained over 300, 300, and 100 epochs, respectively. Consequently, 12 models were created and tested.

CNN architectures

Learning models were composed by using the 3 different networks: AlexNet, GoogLeNet, and VGG-16.[16-18] These 3 CNNs are publicly available on the DIGITS library version 5.0 (NVIDIA website: https://developer.ndivia.com/digits). They are all used to classify what is described in an image, having some differences in these architectures, such as layer depth and fine-tuned status (with or without). AlexNet was developed by Krizhevsky et al. in 2012 as a high-performance image recognition CNN. This network consists of 8 layers and was the simplest and shallowest of the 3 networks tested.[16] GoogLeNet was implemented by Szegydy et al. in 2015. This network has a characteristic 1 × 1 convolutional layer for dimension reduction, which decreases the number of learned CNN model parameters.[17] The phrase"1 × 1" indicates filter size of the convolutional layer, and it is widely known as a simple technique to decrease the number of channels in the field of computer science. VGG-16 was developed by Simonvan in 2015. This network is based on AlexNet and is composed of deeper layers. There are some variations in the VGG network: The number following VGG denotes the number of layers (e.g., VGG-16, VGG-19). In addition, VGG-16 in the DIGITS library is a fine-tuned network, including a pretrained architecture.[18]

Time and storage space requirement

The storage size of the trained model and the time requirement of the training process were compared among the 3 CNNs.

Evaluation of diagnostic performance

The test results for individual image patches with predicted classification values of 0 through 100 are shown in Figure 3. For the ground truth, the prediction percentage of the contact group corresponds to the true positive fraction (TPF = sensitivity) and that of the noncontact group corresponds to the false positive fraction (FPF = 1 − specificity). Receiver operating characteristic (ROC) curves were created, and the best cutoff points of the predictions were determined as the points on the curves closest to the upper left corner of the graph. For an ROC curve, such a point represents the point at which the sensitivity and specificity are maximized. Positive evaluations (assigning an image patch to the contact group) were determined when the predicted value was beyond the cutoff value. Accuracy, sensitivity, and specificity were calculated using these

Fig. 3. An example of the output display of the testing process. Path denotes the place and name of the patch, ground truth denotes the correct class, and top predictions denote the possibility value of each class.

cutoff values.[21] The area under the ROC curve (AUC) was obtained for each model, and these results were compared by using the χ^2 test, with the significance of difference established at $P < .01$.[22]

Evaluation of consistency

Intra- and inter-CNN consistencies were evaluated by using kappa statistics for the test results. For the intra-CNN consistency, the kappa values between the first and second models were obtained for both the 70 × 70 and 140 × 140 pixel image patches. For the inter-CNN consistency, the values were determined for the 3 pairs of 2 CNNs for both the first and second models. The kappa values were evaluated as follows: $\kappa < 0.2$ indicated poor consistency; 0.21 to 0.4 indicated fair consistency; 0.41 to 0.6 indicated moderate consistency; 0.61 to 0.8 indicated good consistency; and 0.81 to 1.0 indicated very good consistency.[23]

Ethical approval

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and followed the tenets of the Helsinki Declaration of 1964 and later versions. Informed consent was obtained from all patients for being included in the study. This study obtained ethical approval from Aichi-Gakuin University ethics committee (No. 496).

RESULTS

Time and storage space requirements

The time and storage space requirements for the learning process were 30 minutes and 63.8 GB, 2 hours and

Table I. Diagnostic performance

CNN	*AlexNet*				*GoogLeNet*				*VGG-16*			
Pixel size	*70 × 70 pixel*		*140 × 140 pixel*		*70 × 70 pixel*		*140 × 140 pixel*		*70 × 70 pixel*		*140 × 140 pixel*	
	First model	*Second model*	*First model*	*Second model*	*First model*	*Second model*	*First model*	*Second model*	*First model*	*Second model*	*First model*	*Second model*
Cutoff value	100	33	83	81	73	81	74	100	69	88	90	18
Accuracy	0.90±0.06	0.88±0.06	0.85±0.07	0.84±0.07	0.92±0.05	0.86±0.07	0.84±0.07	0.82±0.08	0.88±0.06	0.87±0.07	0.71±0.09	0.73±0.09
Sensitivity	0.88±0.06	0.88±0.06	0.82±0.08	0.80±0.08	0.88±0.06	0.84±0.07	0.82±0.08	0.76±0.08	0.88±0.06	0.88±0.06	0.62±0.10	0.80±0.08
Specificity	0.92±0.05	0.88±0.06	0.88±0.06	0.88±0.06	0.96±0.04	0.88±0.06	0.86±0.07	0.88±0.06	0.88±0.06	0.86±0.07	0.80±0.08	0.66±0.09
AUC	0.90	0.91	0.90*,†	0.89*,‡,§	0.93	0.88	0.87	0.83	0.91	0.91	0.75*,‡	0.75†,§
	(0.84-0.96)	(0.85-0.97)	(0.83-0.96)	(0.83-0.96)	(0.88-0.99)	(0.82-0.95)	(0.80-0.95)	(0.75-0.91)	(0.85-0.97)	(0.85-0.97)	(0.65-0.84)	(0.66-0.85)

*,†,‡,§: Values with the same superscript letter exhibit significant difference between them by the chi-squared test (P <.01).
The parentheses denote 95% confidence interval.
AUC, area under the curve; *CNN*, convolutional neural network.

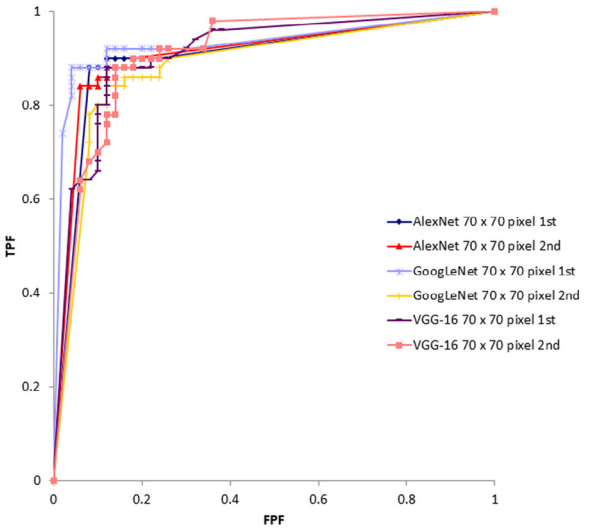

Fig. 4. Receiver operating characteristic (ROC) curves of the 70 × 70 size patch convolutional neural network (CNN) models. All CNNs were trained twice with exactly the same data set. The area under the receiver operating characteristic curve (AUC) values were: AlexNet first model 0.90 (0.84−0.96), second model 0.91 (0.85−0.97); GoogLeNet first model 0.93 (0.88−0.99), second model 0.88 (0.82−0.95); VGG-16 first model 0.91 (0.85−0.97), second model 0.91 (0.85−0.97). The values in parentheses denote the 95% confidence interval.

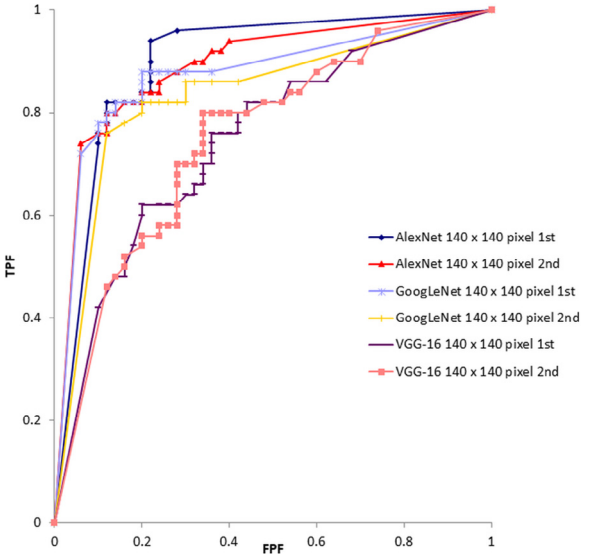

Fig. 5. Receiver operating characteristic (ROC) curves of 140 × 140 size patch convolutional neural network (CNN) models. All CNNs were trained twice with exactly the same data set. The area under the receiver operating characteristic curve (AUC) values were: AlexNet first model 0.90 (0.83−0.96), second model 0.89 (0.83−0.96); GoogLeNet first model 0.87 (0.80−0.95), second model 0.83 (0.75−0.91); VGG-16 first model 0.75 (0.65−0.84), second model 0.75 (0.66−0.85). The parentheses denote 95% confidence interval.

11.6 GB, and 1 hour and 62.4 GB for AlexNet, GoogLeNet, and VGG-16, respectively.

Diagnostic performance

The testing results are summarized in Table I, and the ROC curves are presented in Figures 4 and 5. In comparison of the AUCs between the 70 × 70 and 140 × 140 pixel patches, the smaller patches had relatively high AUCs, ranging from 0.88 to 0.93, whereas the values for the larger patches ranged from 0.75 to 0.93. There were no significant differences between the performance of AlexNet and GoogLeNet or GoogLeNet and VGG-16. The differences between AlexNet and VGG-16 were significant in both the first and second models only in the 140 × 140 pixel patches. The first 70 × 70 pixel GoogLeNet model produced the largest AUC of 0.93, but there were no significant differences compared with the other models created by 70 × 70 pixel patches (see Figure 4 and Table I). Although all models achieved high performance with use of the suggested cutoff values, some patients were misdiagnosed by all CNNs (Figure 6).

Table II. Intra-CNN consistency

CNN	*Patch size*	*Kappa value*
AlexNet	70 × 70 pixel	0.98
	140 × 140 pixel	0.91
GoogLeNet	70 × 70 pixel	0.86
	140 × 140 pixel	0.90
VGG-16	70 × 70 pixel	0.99
	140 × 140 pixel	0.80

CNN: convolutional neural network

Consistency

The intra-CNN consistency is summarized in Table II. All 3 CNNs provided very good consistency for both image patch sizes. The inter-CNN consistency is

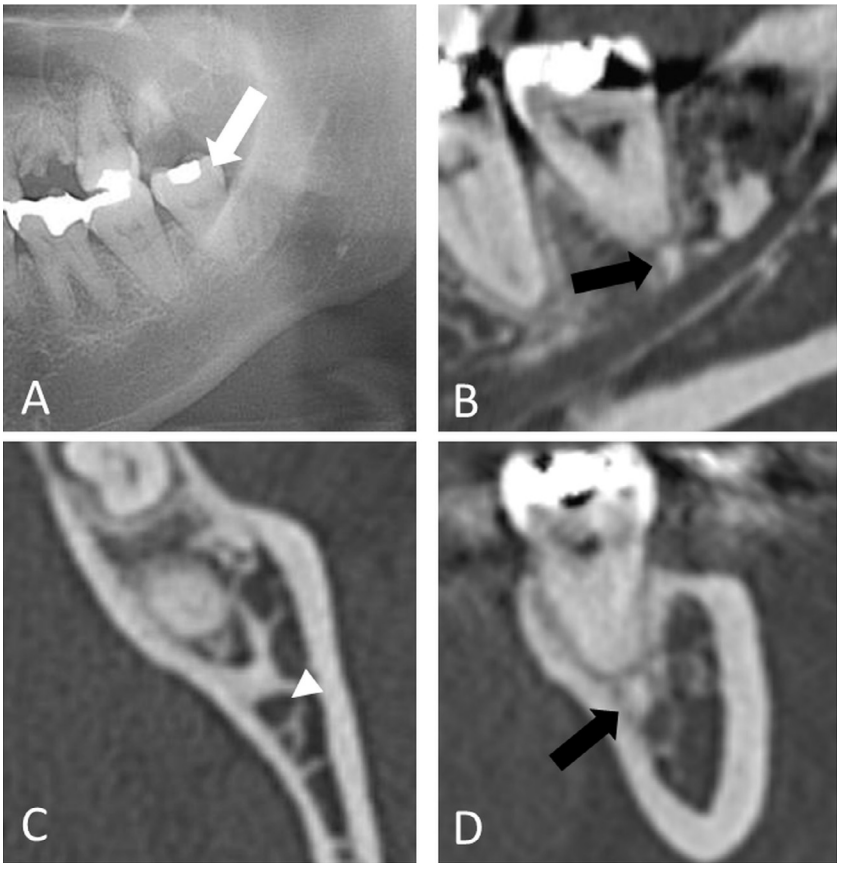

Fig. 6. A noncontact case that all networks misdiagnosed. **A,** Panoramic radiograph, with a white arrow indicating a mandibular third molar that was mistakenly interpreted as the roots contacting the canal. **B,** Sagittal computed tomography (CT) image, with a black arrow indicating the sclerotic area just above the mandibular canal. The sclerotic bone was misinterpreted as part of the third molar root. **C,** Axial CT image denoting no contact between the mandibular third molar and the mandibular canal. A white arrow head indicates the mandibular canal. **D,** Coronal CT image, with a black arrow indicating the sclerotic area adjacent to the mandibular third molar root apex and the mandibular canal.

Table III. Inter-CNN consistency

CNN	*Patch size*	*Kappa value*		
		First model	*Second model*	*Average*
AlexNet vs GoogLeNet	70 × 70 pixel	0.94	0.86	0.90
	140 × 140 pixel	0.86	0.76	0.81
AlexNet vs VGG-16	70 × 70 pixel	0.92	0.89	0.91
	140 × 140 pixel	0.72	0.71	0.72
GoogLeNet vs VGG-16	70 × 70 pixel	0.88	0.83	0.86
	140 × 140 pixel	0.65	0.61	0.63

CNN, convolution neural network.

presented in Table III. The CNN models created by using 70 × 70 pixel image patches exhibited very good consistency among all 3 pairs of 2 CNNs, whereas the consistency was good for the pairs including the model created by VGG-16 with 140 × 140 pixel image patches.

DISCUSSION

The diagnostic performance of CNNs in previous studies[24-32] has been confirmed to be equivalent to that of experienced radiologists. In the present study, some CNN models achieved AUCs of greater than 0.90 with the use of only 600 panoramic images. In the future, it may be possible to automatically diagnose the relationship between the mandibular molar and the mandibular canal on panoramic images by using CNN models. This level of performance, which should approach 100% accuracy for automatic diagnosis, has not yet been achieved, and even if it were, the essential question of who should take responsibility for the diagnostic results would still remain.

The time and storage space required for the learning process depended on the depth of the CNN layers and the total number of parameters learned in the process, respectively.[33,34] The time and capacity required in the present study reflected these opinions. The time required for AlexNet (30 minutes), VGG-16 (1 hour), and GoogLeNet (2 hours) increased according to the 8, 16, and 20 layers, respectively, in these 3 CNNs. GoogLeNet required 11.6 GB of storage space for 5,975,602 parameters, less than the other 2 CNNs. AlexNet and VGG-16 required 63.8 GB and 62.4 GB with their 56,876,418 and 165,746,503 parameters, respectively. In GoogLeNet, the dimension of the parameters was reduced by using the 1 × 1 convolutional layer, thus achieving a relatively small capacity and high accuracy.[17]

A smaller size image patch is generally better for the training process.[21] Our results support this hypothesis; the smaller patch size generally produced better diagnostic results (Table I, Figures 4 and 5). This is probably attributable to the unnecessary information that is included in larger-sized patches. AlexNet had the simplest and shallowest layers among the 3 CNNs evaluated, and was the least influenced by patch size difference, indicating that it is the most versatile CNN. In contrast, VGG-16 was the CNN most strongly influenced by the difference in patch size. This suggests that more attention should be paid to the cropped patch size when a learning model is created by using VGG-16.

In the present study, the training was performed according to the method proposed by Krizhevsky et al.[16] for creating learning models. The image patches were automatically and randomly separated into several mini-batches and assigned as training and validation data sets in every training process. Therefore, there were slight differences in performance between the first and second models, even when using the same data sets for the learning process (see Table I). However, all of the models created in this study showed very good intra-CNN consistency, with kappa values of 0.80 or greater and a maximum value of 0.99 for the VGG-16 70 × 70 pixel patch size (see Table II). This may indicate that a DL CAD system can provide highly reproducible diagnoses. When comparing image patch sizes, the inter-CNN consistency was higher with the smaller-size patches than with the larger-size patches (see Table III). Although a considerable difference was found between the AlexNet and GoogLeNet structures, the inter-CNN consistency was very good regardless of patch size.[23] The inter-CNN consistency, including the VGG-16 model created with 140 × 140 pixel patches, was relatively low, probably because VGG-16 was adversely influenced by the patch size (see Table III). Figure 6 shows a case misdiagnosed by all models, possibly due to a sclerotic area resembling the root between the root apex and the mandibular canal. However, it is generally difficult to determine the cause of differences in diagnoses because DL systems do not easily reveal the reasons behind their judgments.

The cutoff values determined by the ROC curves varied widely in the present study. These values might be important for evaluating the performance of CNN systems because they strongly affect the performance. When sensitivity and specificity are equally important, the values are usually determined by the method used in the present study or by the Youden index for

calculating the maximum value of (sensitivity + specificity − 1).[35] Although the value should be adequate for every model when using a small number of training images, as in the present study, a learning model developed by the accumulation of massive amounts of data might solve this problem by providing steady cutoff values.

There are some limitations to the present investigation. First, cases were abstracted regardless of the status of the mandibular third molar, such as erupted versus impacted, and regardless of the direction, numbers, and morphology of roots. These factors might influence the training and testing results. Second, hyperparameters, such as the learning rate and batch size, could not be sufficiently optimized in the training process, although all of them were set as equally as possible among the CNNs. Third, the panoramic images were obtained from only 1 institution. Panoramic images should be gathered from many institutions to create more accurate CNN models. Fourth, sensitivity and specificity are inversely related. If high sensitivities are desired, the clinician has to expect relatively low specificities, and vice versa. The use of CNN models should be verified taking this relationship into account. Fifth, the relationship between the mandibular third molar and the mandibular canal was not evaluated 3-dimensionally in the present study but was determined by the consensus opinion of experienced radiologists. Such relationships, which could be obtained with CT or CBCT, would be suitable for more detailed and accurate analysis. Therefore, a future DL study should be conducted using such modalities as gold standards. Sixth, CNNs are developed as part of machine learning in the quest for an intelligent machine. They are not readily interpretable in that they are based on hidden functions and transformations that are exceedingly difficult to explain. Therefore, CNN CAD systems have a critical disadvantage in that they are unable to show how to make a diagnosis.

CONCLUSIONS

The diagnostic performance and consistency of learning models created by 3 CNNs were compared with regard to evaluation of the relationship between the mandibular third molar and the mandibular canal on panoramic images. The CAD system created with DL appears to be useful for evaluating this relationship on panoramic images. For the learning process, the size of image patches should be carefully determined to ensure high diagnostic performance and consistency. No significant differences in diagnostic performance were found among the 3 CNNs created by smaller-size image patches, but some differences were verified in the time and storage capacity required for the learning process.

ACKNOWLEDGMENTS

We thank Stuart Jenkinson, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

REFERENCES

1. Rood JP, Sheehab BA. The radiological prediction of inferior alveolar nerve injury during third molar surgery. *Br J Oral Maxillofac Surg*. 1990;28:20-25.
2. Szalma J, Lempel E, Jeges S, Szabo G, Olasz L. The prognostic value of panoramic radiography of inferior alveolar nerve damage after mandibular third molar removal: retrospective study of 400 cases. *Oral Surg Oral Med Oral Pathol Oral Endod*. 2010;109:294-302.
3. Szalma J, Lempel E, Jeges S, Olasz L. Darkening of third molar roots: panoramic radiographic associations with inferior alveolar nerve exposure. *J Oral Maxillofac Surg*. 2011;69:1544-1549.
4. Neves FS, Souza TC, Almeida SM, Haiter-Neto F, Freitas DQ, Boscolo FN. Correlation of panoramic radiography and cone beam CT findings in the assessment of the relationship between impacted mandibular third molars and the mandibular canal. *Dentomaxillofac Radiol*. 2012;41:553-557.
5. Umar G, Obisesan O, Bryant C, Rood JP. Elimination of permanent injuries to the inferior alveolar nerve following surgical intervention of the "high risk" third molar. *Br J Oral Maxillofac Surg*. 2013;51:353-357.
6. Nakamori K, Fujiwara K, Miyazaki A, et al. Clinical assessment of the relationship between the third molar and the inferior alveolar canal using panoramic images and computed tomography. *J Oral Maxillofac Surg*. 2008;66:2308-2313.
7. Nakagawa Y, Ishii H, Nomura Y, et al. Third molar position: reliability of panoramic radiography. *J Oral Maxillofac Surg*. 2007;65:1303-1308.
8. Tanaka T, Murakami K, Kishida T, Itoh T, Morita Y, Noikura T. Relation between mandibular third molar and mandibular canal as assessed by three-dimensional computed tomographic reconstruction. *Jpn J Oral Maxillofac Surg*. 2000;46:251-261.
9. Harada N, Subash BV, Matsuda Y, et al. Characteristic findings on panoramic radiography and cone-beam CT to predict paresthesia after extraction of impacted third molar. *Bull Tokyo Dent Coll*. 2015;56:1-8.
10. Fujita H, Uchiyama Y, Nakagawa T, et al. Computer-aided diagnosis: the emerging of three CAD systems induced by Japanese health care needs. *Comput Methods Programs Biomed*. 2008;92:238-248.
11. Ohashi Y, Ariji Y, Katsumata A, et al. Utilization of computer-aided detection system in diagnosing unilateral maxillary sinusitis on panoramic radiographs. *Dentomaxillofac Radiol*. 2016;45:20150419.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.
13. Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach. In: *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA: Internation Machine Learning Society; 2011:513-520.
14. Erhan D, Szegedy C, Toshev A, Anguelov D. Scalable object detection using deep neural networks. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH: IEEE; 2014:2155-2162.
15. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago: IEEE; 2015:1520-1528.

16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Lake Tahoe, UT: NIPS; 2012:1097-1105.
17. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA: IEEE; 2015:1-9.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia: ICLR; 2015:730-734.
19. Murata M, Ariji Y, Ohashi Y, et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral Radiol*. 2019;35:301-307.
20. Hiraiwa T, Ariji Y, Fukuda M, et al. A deep-learning artificial intelligence system for assessment of root morphology of the mandibular first molar on panoramic radiography. *Dentomaxillofac Radiol*. 2019;48:20180218.
21. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. *AJR Am J Roentgenol*. 2019;212:513-519.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-845.
23. Altman DG. Practical Statistics for Medical Research. London, U.K.: Chapman and Hall; 1991:404.
24. Fukuda M, Inamoto K, Shibata N, et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol*. 2019. https://doi.org/10.1007/s11282-019-00409-x. [epub ahead of print].
25. Ariji Y, Yanashita Y, Kutsuna S, et al. Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2019;128:424-430.
26. Tuzoff DV, Tuzova LN, Bornstein MM, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofac Radiol*. 2019;48:20180051.
27. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. *Dentomaxillofac Radiol*. 2018: 20170344. https://doi.org/10.1259/dmfr.20170344. [epub ahead of print].
28. Ekert T, Krois J, Meinhold L, et al. Deep learning for the radiographic detection of apical lesions. *J Endod*. 2019;45:917-922.e5.
29. Vinayahalingam S, Xi T, Bergé S, Maal T, de Jong G. Automated detection of third molars and mandibular nerve by deep learning. *Sci Rep*. 2019;9:9007.
30. Krois J, Ekert T, Meinhold L, et al. Deep learning for the radiographic detection of periodontal bone loss. *Sci Rep*. 2019;9:8495.
31. Kats L, Vered M, Zlotogorski-Hurvitz A, Harpaz I. Atherosclerotic carotid plaque on panoramic radiographs: neural network detection. *Int J Comput Dent*. 2019;22:163-169.
32. Lee JH, Kim DH, Jeong SN. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral Dis*. 2020;26:152-158.
33. Justus D, Brennan J, Bonner S, McGough AS. Predicting the computational cost of deep learning models. In: *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA: IEEE; 2018:3873-3882.
34. Chien SWD, Markidis S, Sishtla CP, et al. Characterizing deep-learning I/O workloads in TensorFlow. In: *Proceedings of the 2018 IEEE International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PSQ-DISCS)*, Dallas, TX: IEEE; 2018:54-63.
35. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32-35.

Reprint requests:

Motoki Fukuda
Assistant Professor
Department of Oral and Maxillofacial Radiology
Aichi-Gakuin University School of Dentistry
2-11 Suemori-dori
Chikusa-ku
Nagoya 464-8651
Japan
halpop@dpc.agu.ac.jp