Evaluation of deep learning for detecting intraosseous jaw lesions in cone beam computed tomography volumes

Yiing-Shiuan Huang, DDS, MS,[a] Pavel Iakubovskii, MS,[b] Li Zhen Lim, BDS, MS,[a,c] André Mol, DDS, MS, PhD,[a] and Donald A. Tyndall, DDS, MSPH, PhD, FICD[a]

Objective. The study aim was to develop and assess the performance of a deep learning (DL) algorithm in the detection of radiolucent intraosseous jaw lesions in cone beam computed tomography (CBCT) volumes.

Study Design. A total of 290 CBCT volumes from more than 12 different scanners were acquired. Fields of view ranged from 6 × 6 × 6 cm to 18 × 18 × 16 cm. CBCT volumes contained either zero or at least one biopsy-confirmed intraosseous lesion. 80 volumes with no intraosseous lesions were included as controls and were not annotated. 210 volumes with intraosseous lesions were manually annotated using ITK-Snap 3.8.0. 150 volumes (10 control, 140 positive) were presented to the DL software for training. Validation was performed using 60 volumes (30 control, 30 positive). Testing was performed using the remaining 80 volumes (40 control, 40 positive).

Results. The DL algorithm obtained an adjusted sensitivity by case, specificity by case, positive predictive value by case, and negative predictive value by case of 0.975, 0.825, 0.848, and 0.971, respectively.

Conclusions. A DL algorithm showed moderate success at lesion detection in their correct locations, as well as recognition of lesion shape and extent. This study demonstrated the potential of DL methods for intraosseous lesion detection in CBCT volumes. (Oral Surg Oral Med Oral Pathol Oral Radiol 2024;138:173−183)

Artificial intelligence (AI) is defined as the theory and development of computer programs capable of performing complex tasks traditionally accomplished using human intelligence.[1] Although first introduced in the 1960s, early computer vision tasks were limited by computing power and the amount of data available until the development of deep learning and artificial neural networks in the 1980s. Deep learning (DL) is a branch of AI that uses multiple interconnected and layered networks to learn from data.[2] Since then, AI has shown promising results in the medical field in serving as an aid for diagnosis and treatment decision-making for physicians.

In dentistry, AI algorithms have focused on diagnostic tasks, including anatomic landmark localization,[3] automatic tooth identification,[4] assessment of root morphology,[5] temporomandibular joint assessment,[6] and detection of diseases such as dental caries,[7] periapical lesions,[8] periodontal bone loss,[9,10] and vertical root fractures.[11] Several studies have investigated the use of various AI methods to classify and diagnose maxillofacial cysts and tumors on radiographic images for surgical treatment planning.[12,13] At this time, such studies have focused on a narrow scope of diseases, namely cystic and periapical lesions.[14]

Furthermore, most of these studies investigate algorithms developed from 2-dimensional imaging modalities—intraoral and panoramic radiography. While 2D imaging provides useful information, image distortion and superimposition of anatomic structures may impact the diagnostic accuracy of the AI models. In contrast, cone beam computed tomography (CBCT) allows for visualization of osseous structures and pathoses in high resolution and in 3 dimensions. Therefore, the application of AI to 3D imaging in dentistry may overcome limitations of 2D image-based models.

Literature suggests that CBCT-trained models for cyst and tumor classification and segmentation are early in development. Yilmaz et al[15] used a support vector machine to classify 50 CBCT image datasets as either odontogenic keratocysts or periapical cysts. Lee et al[16] found that a deep convolutional neural network (CNN) could be effectively used to distinguish between various odontogenic cysts and showed better results with CBCT axial slices than with panoramic radiographs. Chai et al[17] developed a CNN to automatically classify lesions as either ameloblastoma or OKC using CBCT data. While these studies have shown moderate

Statement of Clinical Relevance

The development of an automatic lesion detection and segmentation model may eventually improve dental practitioners' ability to identify potentially clinically significant findings in cone beam computed tomography volumes requiring further follow-up or referral for management.

[a]Oral and Maxillofacial Radiology, Adams School of Dentistry, University of North Carolina, Chapel Hill, NC, USA.
[b]Denti.AI Technology Inc., Toronto, Ontario, Canada.
[c]Discipline of Oral and Maxillofacial Surgery, Faculty of Dentistry, National University of Singapore, Singapore.
Corresponding author: Yiing-Shiuan Huang, DDS, MS E-mail address: yiingshiuanhuang@gmail.com
Received for publication May 9, 2023; returned for revision Sep 6, 2023; accepted for publication Sep 15, 2023.

2212-4403/$-see front matter
https://doi.org/10.1016/j.oooo.2023.09.011

to excellent results, most use small training datasets with a particular set of conditions or images from a single institution, therefore limiting the generalizability. In addition, all of these models still require that the initial step of lesion detection be performed manually by a clinician,[12] and thus are dependent upon the clinician's ability to recognize and identify areas of pathosis. Currently, 2 studies have demonstrated excellent results using a deep CNN to automatically segment and detect periapical pathosis in CBCT volumes, detecting approximately 93% of periapical lesions.[18,19] These studies illustrated that AI methods can potentially be developed for lesion detection in CBCT volumes.

By providing an assessment purely based on data analysis, AI has been suggested to serve as a "second opinion" that is both objective and reliable, allowing healthcare professionals to form more accurate diagnoses and make appropriate referrals.[20] Although deep learning diagnostic software may benefit all dental practitioners, its potential applications to CBCT imaging may be of greater use to clinicians not specialized in radiology. CBCT imaging in dentistry has significantly increased in recent years, with an estimated 5.2 million volumes taken annually as of 2014-15.[21] However, CBCT interpretation requires more time and familiarity with navigating 3D anatomy as compared to 2-dimensional imaging. General practitioners are responsible for all information within an acquired image[22] but may have less experience in identifying lesions on CBCT and may take longer to review entire CBCT volumes compared to radiology specialists. Therefore, by serving as an adjunctive diagnostic tool, AI software has the potential to improve efficiency and reduce the workload of general dentists by identifying potentially significant lesions.

To our knowledge, no other published studies have investigated the performance of deep learning algorithms for intraosseous lesion detection and segmentation in CBCT volumes. Therefore, the aim of this study was to develop and assess the performance of a deep learning algorithm in the detection and segmentation of various radiolucent jaw lesions in CBCT volumes.

MATERIALS AND METHODS

CBCT dataset selection

Institutional review board approval and waivers of informed consent for research were provided by the University of North Carolina at Chapel Hill (#21-0534).

CBCT volumes were selected retrospectively from the UNC-CH Adams School of Dentistry (UNC ASoD) and the Peking University School and Hospital of Stomatology. Volumes were acquired with 16 different CBCT imaging models and included fields of view (FOV) ranging from 6 × 6 × 6 cm to 18 × 18 × 16 cm. The scanners included the Galileos (Sirona Dental Systems), Orthophos XG or SL (Dentsply Sirona), CS 9000 or CS 9300 (Carestream Dental Inc.), NewTom3G or 5G or VGi MK4 (NewTom Inc.), Scanora 3D (Soredex Co.), Promax 3D (Planmeca), i-CAT 17-19 or 17-19DX (Imaging Sciences International LLC), 3D Accuitomo 170 (J Morita Corp), and DCT PRO (Vatech Co., Ltd.).

Inclusion criteria for lesion-positive volumes. (i) Patients with radiolucent intraosseous jaw lesions in either the maxilla or mandible, including recurrent or multiple lesions and (ii) lesions with a definitive histopathologic diagnosis.

Exclusion criteria for lesion-positive volumes: (i) CBCT volumes of poor diagnostic quality or that did not include the entire intraosseous lesion, (ii) lesions with inconclusive histopathologic results, (iii) lesions of systemic/metabolic or developmental nature, and (iv) lesions with poorly defined borders.

Inclusion criteria for control volumes: Absence of intraosseous lesions as verified by a board-certified oral and maxillofacial radiologist (OMR) with over 40 years of experience.

Exclusion criteria for control volumes: (i) CBCT volumes of poor diagnostic quality due to artifacts, (ii) apical rarefying lesions of odontogenic origin, (iii) paranasal sinus inflammation with > 10 mm thickness, and (iv) patients with generalized, systemic/metabolic, or developmental conditions.

The final dataset consisted of 290 volumes. The dataset was divided into a training set (n = 150), validation set (n = 60), and testing set (n = 80). Together, the training and validation sets consisted of an equal distribution of positive volumes from the UNC ASoD and Peking University. The validation set consisted of 30 lesion-positive and 30 control volumes. The testing set consisted of 40 lesion-positive and 40 control volumes. Of the lesion-positive test volumes, 20 volumes were from UNC and 20 volumes were from Peking University. Tables I and II show the distribution of lesions and CBCT volume sources in the complete dataset.

CBCT volume preparation

All CBCT volumes were anonymized and exported in a multifile DICOM (Digital Imaging and Communications in Medicine) format.

The following characteristics were recorded for each control volume: CBCT unit, FOV, and reason for scan.

The following characteristics were recorded for each lesion-positive volume: CBCT unit, FOV, developmental stage of dentition (mixed or permanent dentition), and number of radiolucent intraosseous lesions present. For each lesion, the location (maxilla or mandible, anterior or posterior), internal architecture (unilocular or multilocular), disease category (cyst, benign tumor, lesion of bone, inflammation, malignancy) and diagnosis were noted. One malignant lesion

Table I. Distribution and allocation of CBCT dataset

Total volumes (n = 290)	*Training*	*Validation*	*Testing*
Pilot (n = 60)	Positive volumes UNC-CH (n = 20) Peking (n = 20)	Positive volumes UNC-CH (n = 5) Peking (n = 5) Control volumes UNC-CH (n = 10)	N/A
Extended (n = 290)	Positive volumes UNC-CH (n = 68) Peking (n = 72) Control volumes UNC-CH (n = 10)	Positive volumes UNC-CH (n = 17) Peking (n = 13) Control volumes UNC-CH (n = 30)	Positive volumes UNC-CH (n = 20) Peking (n = 20) Control volumes UNC-CH (n = 40)

(fibrosarcoma) with well-defined margins was included. The following characteristics were additionally recorded for the lesions in the test set volumes: lesion shape (round or non-round), border cortication (completely, partially or noncorticated), border expansion (expansion or no expansion), association with teeth (crown, root or nontooth associated), and the widest approximate dimension of the lesion. The widest approximate dimension of the lesions was reported to the closest millimeter. Supplemental Table SI shows the distribution of characteristics for the complete dataset and the training, validation and testing subsets.

Lesion-positive volumes were manually annotated on multiple slices in the axial, coronal and sagittal planes using ITK-Snap 3.8.0[23] (Figure 1). The brush annotation tool allowed for approximation of the lesion borders. Therefore, lesion borders and a narrow margin of the surrounding normal tissue were included within the region of annotation. Annotations were made by an OMR resident (Y.S.H.) and a board-certified OMR with 5 years of experience (L.Z.L.) and were exported as NRRD (Nearly Raw Raster Data) files. All annotations were reviewed by the OMR resident for consistency prior to presentation to the DL algorithm. Control volumes were not annotated.

Table II. Distribution of intraosseous lesions

210 positive volumes, 227 lesions		*UNC-CH*	*Peking*
Cyst	Dentigerous cyst	20	27
	Radicular or residual cyst	12	27
	Odontogenic keratocyst	26	22
	Nasopalatine duct cyst	10	10
	Calcifying odontogenic cyst	2	0
	Buccal bifurcation cyst	2	0
	Surgical ciliated cyst	1	0
	Lateral periodontal cyst or botryoid cyst	7	0
Benign tumors	Ameloblastoma	11	13
	Myxoma	4	1
	Adenomatoid odontogenic tumor	2	0
	Central odontogenic fibroma	2	0
Lesions of bone	Simple bone cavity	7	2
	Aneurysmal bone cyst	1	0
	Central giant cell lesion	1	0
Inflammation	Periapical granuloma	11	5
Malignancy	Fibrosarcoma	1	0

Pilot training and validation

A pilot trial was first conducted using a subset of 60 volumes (Table I) to develop a DL algorithm and verify methodology. A U-Net-like segmentation model based on a ResNet18 encoder initialized with ImageNet weights was used. 40 positive volumes with an equal distribution between the two institutional sources were used to develop and train the model using supervised and transfer learning. Training was performed with axial slices from the positive volumes, where positive slices (containing lesion) and negative slices (without lesion) were sampled with a 1:1 ratio. The pilot model was validated using 20 volumes (10 positive, 10 control). Model predictions were visually compared with the lesion annotations. The pilot model predicted lesions in the correct locations in 8 of the 10 positive volumes. All predictions showed a reduced size, especially in the superoinferior dimension, in comparison to the corresponding annotation size. False-positive predictions were present in all control and in 8 of the 10 positive volumes and presented as a linear, flat shape, consistent with training based on axial slices only. The pilot model demonstrated a preliminary ability to recognize patterns and detect intraosseous lesions in the correct locations and suggested that further training would improve the quality of lesion detection. As the pilot model was trained only on axial slices, the pilot trial also suggested the possibility of improved algorithm performance with training based on 3D (i.e., volumetric) data instead of 2D (i.e., axial slices) data alone. It was also hypothesized that training using 3D data could help to eliminate the characteristic flat-shaped false-positive predictions observed in the pilot trial.

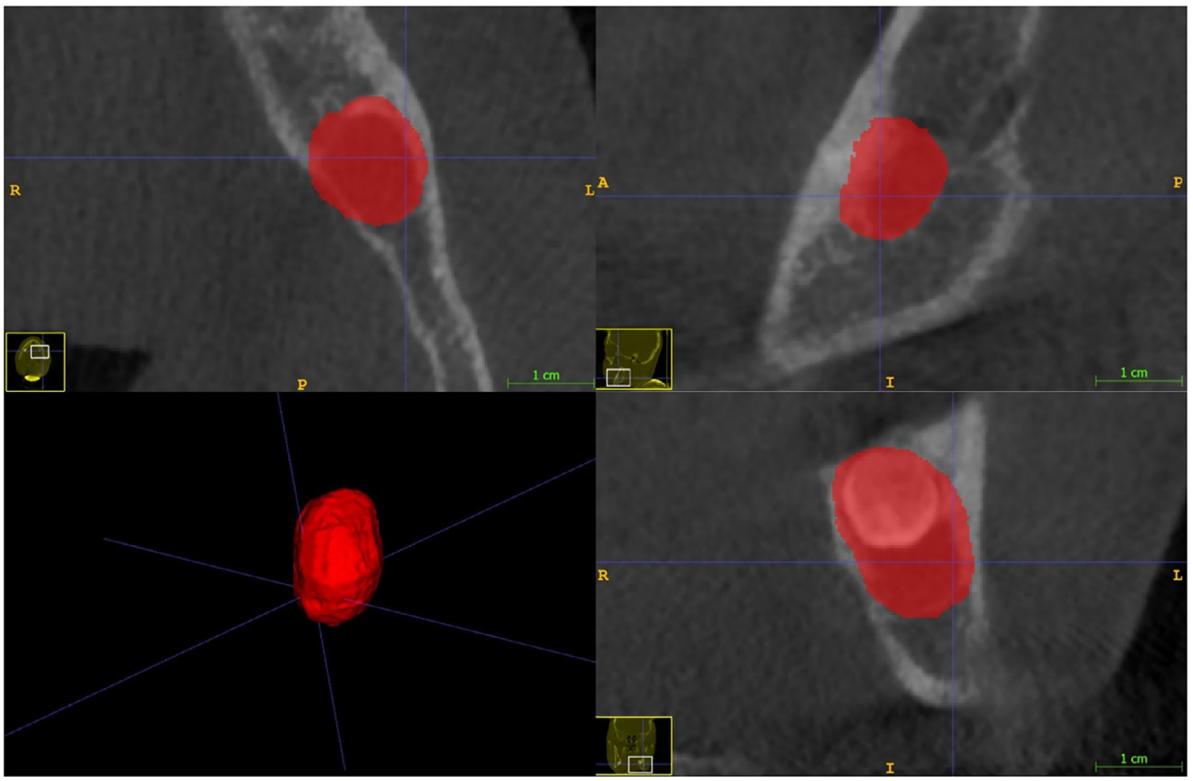

Fig. 1. Completed annotation on a positive CBCT volume using ITK-Snap 3.8.0.

Extended algorithm training and validation

All volumes were resized to a spatial resolution of 0.5 mm/pixel in all dimensions. Because some CBCT volumes were not oriented in a standard way (i.e., axial and sagittal views were interchanged or axial view displayed the front of patient facing down instead of up), all volumes adjusted to the same orientation. Due to the variety of CBCT scanners included in the dataset, a form of normalization was necessary to adjust for varying brightness and contrast across different CBCT scanners. In this study, normalization was performed during the training process via rescale intercept data augmentation to increase model tolerance to different pixel values. For rescale intercept data augmentation, a random number from −1000 to 1000 was added to slice values. Rescale intercept data augmentation was not performed during either validation or testing.

Based on the pilot trial findings, two DL models were trained and validated to assess and compare performance. The following definitions are used to describe the data sampled from each CBCT volume in the training process:

1) Positive: a slice or volume containing a manual annotation mask.
2) Negative: a slice or volume without any manual annotation mask.

In the first model (2D), training was performed using random positive and negative axial slices within each positive case. Positive and negative slices were sampled with a 1:1 proportion. For each slice, the 2 neighboring slices were also included. Validation was performed by passing all axial slices of each validation volume sequentially through the neural network and stacked to 3D tensor.

In the second model (3D), training was performed using random positive and negative volumetric crops with a size of 64 × 64 × 64. Positive and negative crops were sampled with a 5:1 proportion. Validation was performed by passing the full CBCT volume as a 3-dimensional image through the model.

Algorithm testing

The 3D model was selected for testing because it showed better performance than the 2D model during validation. This model used a 3D U-net architecture with a pretrained ResNet18 encoder. Testing was performed using 80 volumes (40 positive, 40 control). The full distribution of the test set characteristics is shown in Supplemental Tables SI and SII.

Performance metrics

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were used to assess the performance of the DL model. F1 score[i] is calculated as the harmonic mean of sensitivity and PPV and provides an overall measure of model performance while minimizing errors from both false-positive and false-negative predictions. Therefore, when either sensitivity or PPV is low, the F1 score will be low even if the other metric is high. Practically speaking, lesion detection software with a high F1 score will not miss many lesions and will not result in many false positives, thereby reducing the amount of time and effort required for clinician review of the AI results.

Intersection over Union (IoU)[ii] is a measure of the amount of overlap between two masks and was calculated to determine the voxel-matching accuracy of the prediction masks. IoU values were calculated using the volumetric sizes (number of voxels) of the 3-dimensional manual segmentation and model prediction masks. For validation, predictions were considered as true positive (TP) when the manual segmentation mask and the model prediction had an IoU > 0.5. All predictions in the validation phase had a volumetric size greater than 200 voxels. Therefore, a minimum volume threshold of 200 voxels was selected, such that all predicted regions with a volumetric size less than 200 voxels were not visualized during the testing phase.

During testing, predictions were considered TP for sensitivity and specificity calculations based on three IoU minimum thresholds: 0.5, 0.3, and 0.1. In other words, at an IoU threshold of 0.5, predictions are considered TP only if the IoU of the prediction volume and manual segmentation is greater than 0.5. Therefore, at lower IoU thresholds, a greater number of predictions are considered as "true positive" detections. Metrics were assessed both by lesion (based on the total number of lesions tested) and by volume (calculated per test CBCT volume and averaged across all test volumes) where appropriate. For instance, any normal location where no prediction was made should be considered TN and cannot be discretely counted. As such, there are an infinite number of locations that could be considered TN, so specificity cannot be meaningfully calculated on a by-lesion basis. Sensitivity by lesion was also calculated based on the following subgroups: volume source, lesion location in the maxilla or mandible, and in the anterior or posterior regions.

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

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

$$Specificity = \frac{TN}{TN + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$F1\ score = \frac{2 * TP}{2 *\ TP + FP + FN}$$

$$IoU = \frac{\left|Annotation \cap Prediction\right|}{\left|Annotation \cup Prediction\right|}$$

(*TP*, true positive; *TN*, true negative; *FP*, false positive; *FN*, false negative.)

RESULTS

Validation

In validation, the 3D model achieved a lesion sensitivity of 0.907, PPV of 0.625 and F1 score of 0.738. By volume, the 3D model achieved a sensitivity of 0.936 and specificity of 0.768. Table III shows the performance metrics of the 3D model by lesion and by volume after validation.

Testing

Figure 2 shows two examples of model prediction in comparison to the manual annotation, visualized on a single axial slice. Model performance analysis was performed using Python. Use of a minimum IoU threshold to define TP predictions resulted in potential underestimation of PPV because some model predictions were doubly counted as both FN and FP in the calculation. Therefore, adjusted by-volume metrics were manually calculated where a positive volume was considered true positive if at least one lesion was correctly

Table III. Performance metrics of 3D model validation

30 controls, 30 positive volumes, 31 lesions

	Sensitivity	*Specificity*	*PPV*	*F1*
By lesion	0.907		0.625	0.738
By case	0.936	0.768		

PPV, positive predictive value.

[i]F1 score ranges between 0 and 1.0, where 1.0 signifies perfect sensitivity and positive predictive values. A F1 score of 0 indicates that either sensitivity or positive predictive value is zero.

[ii]Intersection over Union ranges between 0 and 1.0, where 1.0 signifies perfect overlap of two masks. An IoU of 0 indicates that there is no overlap between the manual segmentation and predicted segmentation masks.

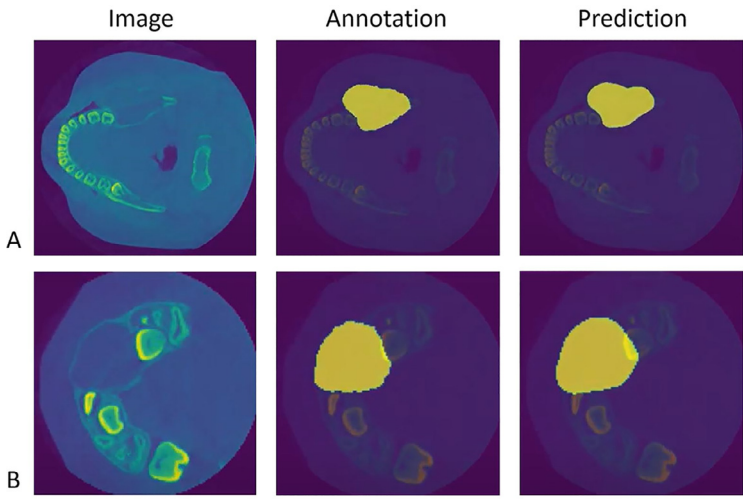

Fig. 2. Visualization of model prediction compared to manual annotation on a single axial slice. **A)** Unicystic ameloblastoma, IoU = 0.91. **B)** Inflamed dentigerous cyst, IoU = 0.746.

detected and false negative if no lesions were detected or no lesions met the IoU threshold. Table IV shows the adjusted by-case metrics. Nonadjusted metrics can be found in Supplemental Table SIII.

At an IoU minimum threshold of 0.1, model testing achieved an adjusted by-volume sensitivity and specificity of 0.975 and 0.825, respectively. Sensitivity by lesion and by volume showed minimal increases at lower IoU thresholds in both nonadjusted and adjusted calculations.

Table V shows the subgroup analysis by lesion location in the jaws, anteroposterior position and volume source. Sensitivity by lesion based on lesion shape and margin characteristics were not included due to the small and unbalanced sample sizes of each subgroup. Statistical differences between subgroups were not assessed due to the small sample sizes. The model showed a higher by-lesion sensitivity for maxillary lesions in comparison to mandibular lesions. The model also showed increased sensitivity for lesions located anterior to the canine compared to those located in posterior regions. Sensitivity was higher for the lesions in volumes from Peking University than those collected from the UNC database.

DISCUSSION

The findings demonstrated an overall successful development of a DL algorithm for automatic detection and segmentation of intraosseous radiolucent lesions in CBCT volumes with a sensitivity ranging between 0.8 and 0.9. On the other hand, nonadjusted PPV values ranged between 0.5 and 0.6 due to the relatively high number of false-positive predictions. However, these PPV values likely underestimate model performance due to three reasons. First, several volumes demonstrated multiple false-positive predictions, meaning that false positives were highly prevalent in only a few cases rather than throughout the entire test set. Second, each continuous collection of voxels was considered a separate prediction, even if the predictions were

Table IV. Adjusted lesion detection performance metrics

40 controls, 40 positive volumes, 47 lesions

	A case is considered TP if IoU:					
	>0.5		*>0.3*		*>0.1*	
	Positive	*Control*	*Positive*	*Control*	*Positive*	*Control*
At least 1 lesion is correctly detected	36	7	37	7	39	7
No lesions are detected OR No lesions meet the IoU threshold	4	33	3	33	1	33
Sensitivity by case	0.9		0.925		0.975	
Specificity by case	0.825		0.825		0.825	
PPV by case	0.837		0.841		0.848	
NPV by case	0.892		0.917		0.971	

TP, true positive; *IoU*, Intersection over Union; *PPV*, positive predictive value; *NPV*, negative predictive value.

Table V. Sensitivity by lesion based on subgroup

	A predicted lesion is considered TP if IoU:		
	>0.5	*>0.3*	*>0.1*
Maxilla (n = 22)	0.818	0.864	0.909
Mandible (n = 25)	0.76	0.76	0.8
Anterior (n = 21)	0.857	0.857	0.905
Posterior (n = 26)	0.731	0.769	0.808
UNC (n = 27)	0.704	0.741	0.778
Peking (n = 20)	0.9	0.9	0.95

TP, true positive; *IoU*, Intersection over Union.

visually located in the same area. For instance, in one case, 3 false-positive predictions were made but all 3 were immediately surrounding an impacted tooth and counted separately even though collectively they indicated the same area of interest. In another case, a false-positive prediction was caused by a small number of voxels disconnected from the true positive prediction. Third, the use of an IoU threshold to define TP detections meant that some model predictions were doubly counted as both FN and FP in the calculation of nonadjusted metrics. Although model predictions that did not meet the specified IoU thresholds were considered FN, these areas are still highlighted by the model and would be visualized by a user (Figure 3A). In other words, the number of FP used to calculate both specificity by case and PPV was misleadingly high. Therefore, the nonadjusted PPV by lesion values likely underestimated the ability of the model to detect areas that truly represent intraosseous radiolucent lesions and subsequently produced F1 scores ranging between 0.6 and 0.7. Although nonadjusted F1 score values were only fair to moderately good, these values most likely underestimate the model performance due to the low nonadjusted PPV values and double counting of detections as FN and FP. This is supported by the increase in PPV by case in the adjusted calculations.

Performance metrics showed minimal improvement as the IoU threshold was lowered from 0.5 to 0.3 and 0.1. It is important to note that using an IoU threshold to determine which predictions are considered true positive leads to some nuances in understanding the overall performance of the DL algorithm. At an IoU threshold of 0.5, 10 predictions were labeled as "false negative." However, 3 of these predictions were actually located in the area of the manual annotation and were considered "false negative" because they did not meet the IoU criteria for the true positive label (Figure 3A). Due to the nature of the IoU calculation, these 3 predictions were either much smaller or greater in volume when compared to size of the manual annotation. At the 0.1 IoU threshold, these 3 predictions are labeled as true positive predictions, thereby slightly increasing the by-lesion sensitivity. At the 0.1 threshold, there were no predictions that were considered false negatives due to the selected threshold.

The minimum IoU thresholds applied in this study have minimal significance when translated to the clinical setting, where the healthcare professional is responsible for verifying the results produced by AI software prior to making a clinical decision. Lesion detection software would present a prediction indicating a region of interest that requires further review by the clinician to confirm the presence or absence of a lesion. As such, it would be unnecessary for the size and shape of the prediction to exactly correspond with that of the actual lesion in order to be correctly detected and identified by a clinician. Given this perspective, there is flexibility in the degree of IoU required for a prediction to be considered "true positive," a lesion that has been correctly detected and identified in an accurate location. A lower IoU threshold provides performance metrics that better describe the algorithm's ability to purely detect lesions at the appropriate location while a higher IoU threshold produces metrics that account for the ability to predict an accurate lesion size and shape in addition to location (Figure 3).

In this study, lowering the IoU minimum threshold from 0.5 to 0.1 affected the classification of only the 3 predictions that were doubly counted as discussed above, shifting them to TP. This accounted for all predictions that were made in the correct locations, and at the IoU threshold of 0.1, all "false positive" detections truly represented predicted areas that did not represent an intraosseous lesion. Furthermore, excluding the 3 predictions whose classification changed depending on the IoU, all other 37 true positive predictions had an IoU ranging between 0.543 and 0.91. Of these, 30/37 (81.1%) predictions had an IoU over 0.7. The minimal improvement in performance metrics as IoU threshold was adjusted and overall high IoU values indicates that the DL algorithm already demonstrates moderate success at not only lesion localization, but also in recognizing lesion shape and extent as well.

At all IoU thresholds, the DL algorithm produced false-positive predictions for the following findings: areas of mucosal thickening within the maxillary sinuses, follicular spaces around unerupted or impacted teeth, areas of sparse trabecular pattern, and soft tissue calcifications. Most of these false-positive findings may not be clinically significant and may not require further management. However, they may be considered abnormalities or deviations from normal and can be reported as incidental findings in an interpretation report. Some systematic reviews have found that incidental findings are found in 77%-92% of CBCT volumes.[24] It is suggested that as many as 30% of

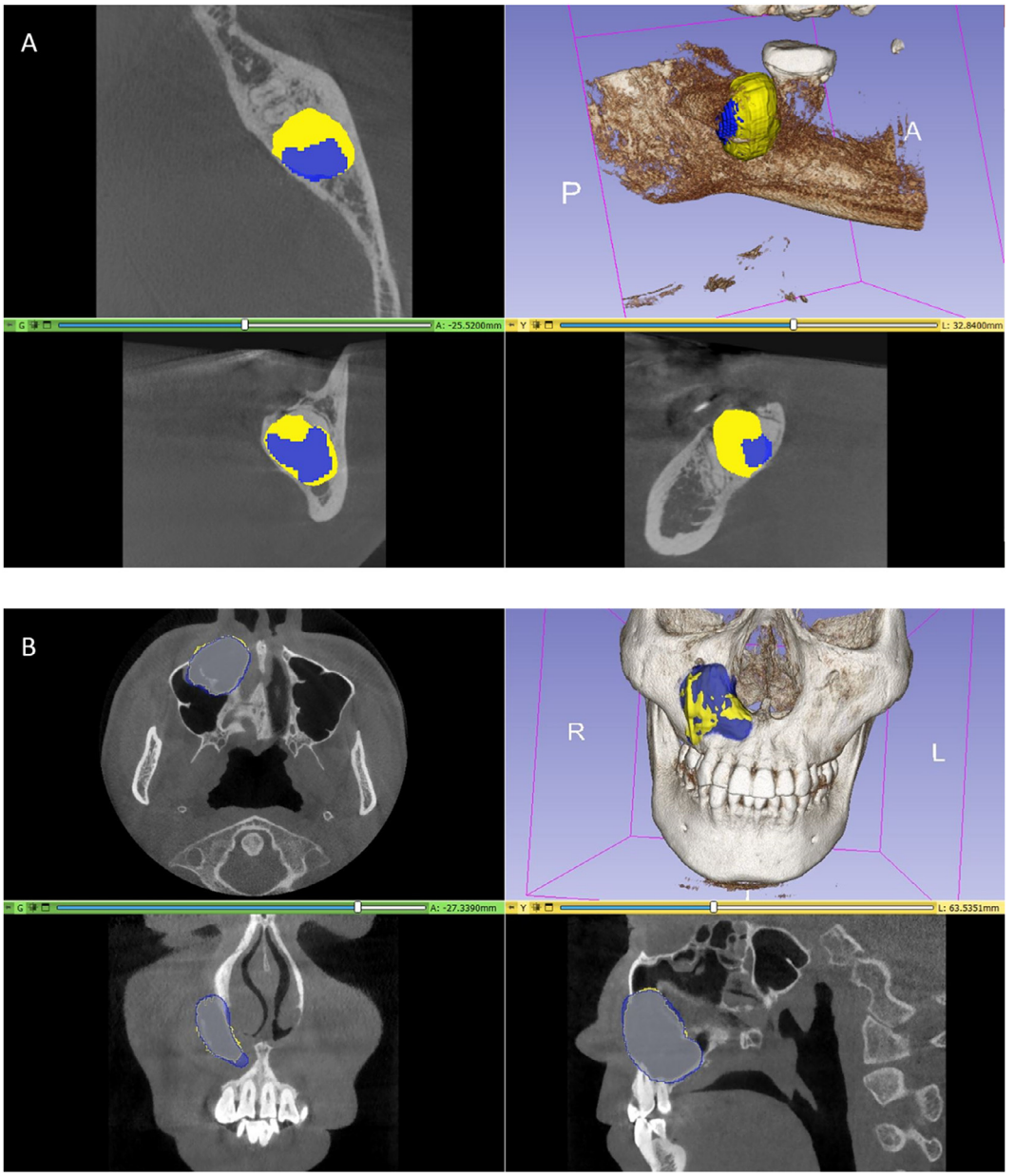

Fig. 3. Model prediction mask compared to annotation mask. Yellow = annotation. Blue = prediction. **A)** IoU = 0.253. Although IoU is relatively low, the prediction is located in the correct location. In the nonadjusted metrics, this prediction was considered both a FN and FP when the IoU threshold was set to >0.5 and 0.3, but was considered TP when IoU > 0.1. **B)** IoU = 0.859. IoU is relatively high and the prediction is both located in the correct location and more closely approximates the shape and size of the lesion.

incidental findings are of clinical significance, requiring further follow-up or referral for management. Therefore, although the current study focused on detection of intraosseous lesions, it is clinically appropriate for lesion detection software to identify all possible areas of concern and not be limited to a particular type of lesion presentation. Our study shows that the development of such DL software is promising with further training on a dataset containing lesions with a more diverse set of presentations.

Two false-positive predictions were located in areas of normal anatomic structures, specifically the mental foramen and the nasopalatine canal. Both of these structures present as well-defined hypodense areas, typically with a corticated border and therefore may mimic features of true radiolucent lesions, leading to a false-positive prediction. The follicular spaces around unerupted and impacted teeth that were detected as false positives also present as well-defined, corticated radiolucent regions. Similarly, bone marrow defects

are areas of cancellous bone that show a sparse trabecular pattern and visually appear as moderately well-defined radiolucent regions. This suggests that the DL algorithm has been trained to recognize areas of relative hypodensity to the surrounding regions as abnormalities and that training with additional data may help to further refine performance.

In the current study, almost all false-negative predictions involved lesions with an approximate widest dimension of 6 mm or less in both validation and testing phases. Furthermore, during validation, all false-negative predictions consisted of small inflammatory lesions located in the maxilla. In testing, only 2 of the 7 false-negative predictions were measured greater than 6 mm wide. These 2 lesions were located near the maxillary sinus floor without a corticated border and in the posterior mandible around an impacted tooth. The training dataset contained 149 total lesions of various sizes, with the approximate widest dimensions ranging from 6 to 93 mm. The majority of lesions ranged between 10 and 30 mm wide. However, only 7/149 (4.7%) lesions measured 10 mm or less in size. Clinically, small lesions are the most likely to be missed. Early detection of lesions, especially of aggressive benign tumors or malignant neoplasms, can greatly affect treatment outcomes. Further training with small lesions with a diameter less than 5 mm may help to improve the algorithm's ability to detect lesions of all sizes and thereby improve patient care.

In the literature, few studies have investigated the use of AI methods in automatic lesion detection and segmentation on CBCT. One study demonstrated successful detection of cystic lesions on CBCT using an automatic segmentation system based on asymmetry analysis instead of deep learning methods.[25] However, the symmetry detection system could not handle cysts and lesions with large asymmetric variations, cysts with poorly defined boundaries, and cases of bilateral cystic lesions. In addition, this study focused only on three types of cystic lesions, thus limiting the general application of this system to detection of other lesions. More recently, two studies have shown excellent results with neural networks in the detection of periapical pathosis. One detected 92.8% of the periapical lesions in 109 patients and found no statistically significant difference in the size of the segmentation volumes between the segmentations made by the radiologist and those predicted by the deep CNN.[18] The other study demonstrated a 93% detection accuracy but included only 20 CBCT scans with a total of 29 lesions in their test set.[19] Both studies used CBCT volumes acquired from a single scanner using a single FOV, and therefore have limited generalizability. The present study showed a detection rate of 40/47 (85.1%) lesions using volumes of varying FOV from 16 different CBCT scanners at 2 different institutions.

The current study has several limitations. First, our dataset consisted of volumes acquired from only two institutions. Although these volumes were acquired using at least 12 different CBCT scanner models, the algorithm demonstrated higher performance metrics for the Peking volumes as compared to the UNC volumes. The Peking volumes included only four scanners while the UNC volumes contained a greater variety of scanners, leading to increased variability in image quality. This suggests that the algorithm may have limited generalizability to CBCT volumes acquired using other scanners at other institutions. However, it is interesting to note that there were three control volumes in the test set acquired using a CBCT machine (AT ProVecta) that were not included in the training set. No false positives were detected in any of these three volumes, which suggests that the model is moderately generalizable. Testing with more volumes from unfamiliar scanners would be necessary to accurately evaluate the generalizability of the model.

Second, our dataset consisted almost entirely of benign and only radiolucent lesions, most of which were cystic in nature. These lesions are well-defined by nature and allow for easier segmentation, whereas ill-defined lesions such as malignancies would be difficult to annotate with confidence regarding the extent of disease. Furthermore, there are a greater number of clinically significant diseases that present as radiolucent lesions in comparison to mixed density or radiopaque lesions. Our dataset also focused only on lesions of the maxilla and mandible. However, it is clinically relevant for a lesion detection software to identify all possibly significant abnormalities in all anatomic areas included in the imaging volume, especially those that are aggressive or malignant in behavior as this directly impacts a patient's well-being. Clinicians may also be less confident and experienced in detecting lesions outside of the maxilla and mandible. A third possible limiting factor related to our dataset was the inclusion of recurrent lesions. In cases of recurrence, there is usually a history of previous surgical intervention which may alter the appearance of surrounding normal anatomy and therefore could have negatively affected model training and performance. However, as with the volumes containing small lesions, cases with recurrent lesions accounted for a small proportion of the dataset and may have had a minimal effect in our study. Further investigation is needed to train and develop an algorithm capable of detecting aggressive diseases and lesions of varying internal density in the entire oral and maxillofacial structures.

In addition, the positive volumes were manually annotated by two individuals with varying levels of

training, with most of the volumes being annotated by the OMR resident. Algorithm performance relies on the quality of input data and requires that the training data is accurately segmented. The segmentation masks identifying the size and shape of the lesions can vary depending on the individual and their level of experience. Therefore, the use of manual segmentation may not have provided the most ideal results. However, in this study, a board-certified OMR annotated 25 of the 170 positive volumes used in training and validation and 1 of 40 positive volumes used in testing. Because the vast majority of positive cases were annotated by a single individual and the same individual reviewed all annotations for consistency prior to algorithm training, it is unlikely that manual segmentation had any significant impacts on the resulting performance metrics. Furthermore, manual segmentation was highly time-consuming, requiring 15 minutes to more than 1 hour to annotate a single CBCT volume. For future studies, it is recommended to investigate the use of an automated segmentation method to eliminate these uncertainties and to reduce the time needed to prepare an adequate dataset.

Finally, this study required both normalization of brightness and contrast, as well as reorientation of the CBCT volumes prior to presentation to the deep learning model. Normalization via rescale intercept augmentation was performed during the training process only, while all validation and test volumes were presented to the model at their original brightness and contrast levels. Therefore, the results of the deep learning model represent the clinical setting in which CBCT volumes are presented to the software without normalization. Because the model was trained using volumes in the same orientation, the model may be sensitive to large changes in head position. However, in this study, reorientation was performed because a few volumes were oriented inappropriately. Minor head positioning errors during acquisition of approximately 10-15 degrees were still present within the CBCT volumes after reorientation and did not affect model performance. If DICOM tags for the CBCT volumes are properly indicated, orientation adjustments can be performed automatically rather than manually by a human operator.

CONCLUSION

This study demonstrates the successful initial development of a DL algorithm in the automatic detection and segmentation of intraosseous radiolucent lesions of the jaws. Further training with a more diverse dataset containing small lesions, mixed density and radiopaque lesions, lesions in areas beyond the maxilla and mandible and from multiple institutions is needed to improve performance and generalizability before such technology can be applied to clinical practice. In the future, observer studies investigating how the use of lesion detection software influences a clinician's ability to diagnose diseases with CBCT will help provide insight into the potential benefits and drawbacks of DL implementation in practice.

ACKNOWLEDGMENT

We would like to thank our colleagues Dr. Kai-Yuan Fu and Dr. Jie Lei at the Peking University School and Hospital of Stomatology for the contribution of CBCT volumes to this project.

PRESENTATION

A pilot study of this research was presented as a poster presentation at the 73rd Annual Session of the American Academy of Oral and Maxillofacial Radiology on September 9, 2022.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF INTERESTS

This study was conducted in collaboration with Denti. AI Technology Inc. (Toronto, CA). Model development and evaluation of performance was performed by a computer science engineer (P.I.) employed by Denti. AI. One author (D.A.T.) serves as an informal consultant for Denti.AI. No funding was provided by Denti. AI in relation to this study.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.oooo.2023.09.011.

REFERENCES

1. Mupparapu M, Wu CW, Chen YC. Artificial intelligence, machine learning, neural networks, and deep learning: futuristic concepts for new dental diagnosis. *Quintessence Int*. 2018;49(9):687-688. https://doi.org/10.3290/j.qi.a41107.
2. Mazurowski MA. Artificial intelligence in radiology: some ethical considerations for radiologists and algorithm developers. *Acad Radiol*. 2020;27(1):127-129. https://doi.org/10.1016/j.acra.2019.04.024.
3. Shahidi S, Bahrampour E, Soltanimehr E, et al. The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images. *BMC Med Imaging*. 2014;14:32. https://doi.org/10.1186/1471-2342-14-32.
4. Tuzoff DV, Tuzova LN, Bornstein MM, et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofac Radiol*. 2019;48(4):20180051. https://doi.org/10.1259/dmfr.20180051.
5. 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(3):20180218. https://doi.org/10.1259/dmfr.20180218.
6. Lee KS, Kwak HJ, Oh JM, et al. Automated detection of TMJ osteoarthritis based on artificial intelligence. *J Dent Res*. 2020;99(12):1363-1367. https://doi.org/10.1177/0022034520936950.
7. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J Dent*. 2018;77:106-111. https://doi.org/10.1016/j.jdent.2018.07.015.
8. Ekert T, Krois J, Meinhold L, et al. Deep learning for the radiographic detection of apical lesions. *J Endod*. 2019;45(7):917-922.e5. https://doi.org/10.1016/j.joen.2019.03.016.
9. Chang HJ, Lee SJ, Yong TH, et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Sci Rep*. 2020;10(1):7531. https://doi.org/10.1038/s41598-020-64509-z.
10. Krois J, Ekert T, Meinhold L, et al. Deep learning for the radiographic detection of periodontal bone loss. *Sci Rep*. 2019;9(1):8495. https://doi.org/10.1038/s41598-019-44839-3.
11. Fukuda M, Inamoto K, Shibata N, et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol*. 2020;36(4):337-343. https://doi.org/10.1007/s11282-019-00409-x.
12. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. *Dentomaxillofac Radiol*. 2020;49(1):20190107. https://doi.org/10.1259/dmfr.20190107.
13. Mureşanu S, Almăşan O, Hedeşiu M, Dioşan L, Dinu C, Jacobs R. Artificial intelligence models for clinical usage in dentistry with a focus on dentomaxillofacial CBCT: a systematic review. *Oral Radiol*. 2023;39(1):18-40. https://doi.org/10.1007/s11282-022-00660-9.
14. Hung K, Yeung AWK, Tanaka R, Bornstein MM. Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice. *Int J Environ Res Public Health*. 2020;17(12):4424. https://doi.org/10.3390/ijerph17124424.
15. Yilmaz E, Kayikcioglu T, Kayipmaz S. Computer-aided diagnosis of periapical cyst and keratocystic odontogenic tumor on cone beam computed tomography. *Comput Methods Programs Biomed*. 2017;146:91-100. https://doi.org/10.1016/j.cmpb.2017.05.012.
16. 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(1):152-158. https://doi.org/10.1111/odi.13223.
17. Chai ZK, Mao L, Chen H, et al. Improved diagnostic accuracy of ameloblastoma and odontogenic keratocyst on cone-beam CT by artificial intelligence. *Front Oncol*. 2022;11:793417. https://doi.org/10.3389/fonc.2021.793417.
18. Orhan K, Bayrakdar IS, Ezhov M, Kravtsov A, Özyürek T. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. *Int Endod J*. 2020;53(5):680-689. https://doi.org/10.1111/iej.13265.
19. Setzer FC, Shi KJ, Zhang Z, et al. Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images. *J Endod*. 2020;46(7):987-993. https://doi.org/10.1016/j.joen.2020.03.025.
20. Nagi R, Aravinda K, Rakesh N, Gupta R, Pal A, Mann AK. Clinical applications and performance of intelligent systems in dental and maxillofacial radiology: a review. *Imaging Sci Dent*. 2020;50(2):81-92. https://doi.org/10.5624/isd.2020.50.2.81.
21. Nationwide Evaluation of X-Ray Trends (NEXT), "Tabulation and graphical summary of the 2014-2015 dental survey." 2019. CRCPD Publication-E-16-2. https://cdn.ymaws.com/www.crcpd.org/resource/collection/81C6DB13-25B1-4118-8600-9615624818AA/E-19-2_2014-2015_Dental_NEXT_Summary_Report.pdf
22. Carter L, Farman AG, Geist J, et al. American Academy of Oral and Maxillofacial Radiology executive opinion statement on performing and interpreting diagnostic cone beam computed tomography. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 2008;106(4):561-562. https://doi.org/10.1016/j.tripleo.2008.07.007.
23. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116-1128. https://doi.org/10.1016/j.neuroimage.2006.01.015.
24. Khalifa HM, Felemban OM. Nature and clinical significance of incidental findings in maxillofacial cone-beam computed tomography: a systematic review. *Oral Radiol*. 2021;37(4):547-559. https://doi.org/10.1007/s11282-020-00499-y.
25. Abdolali F, Zoroofi RA, Otake Y, Sato Y. Automatic segmentation of maxillofacial cysts in cone beam CT images. *Comput Biol Med*. 2016;72:108-119. https://doi.org/10.1016/j.compbiomed.2016.03.014.