치의과학박사 학위논문

Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram

파노라마 영상에서

측두하악관절 골관절염 진단을 위한 인공지능 연구

2021년 8월

서울대학교 대학원

치의과학과 구강내과 · 진단학 전공

최 은 혜

Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram

파노라마 영상에서

측두하악관절 골관절염 진단을 위한 인공지능 연구

지도교수 박 희 경

이 논문을 치의과학박사 학위논문으로 제출함

2021 년 5 월

서울대학교 대학원

치의과학과 구강내과 · 진단학 전공

최 은 혜

최은혜의 치의과학박사 학위논문을 인준함

2021 년 7 월

위 원 장 _____

부위원장 _____

위 원 _____

위 원 _____

위 원 _____

ABSTRACT

Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram

Eunhye Choi, D.D.S., M.S.D.

Program in Oral Medicine and Oral Diagnosis, Dept. of Dental Science, Graduate School, Seoul National University

(Directed by Professor **Hee-Kyung Park**, D.D.S., M.S.D., Ph.D.)

Orthopantomogram (OPG) is important for primary diagnosis of temporomandibular joint osteoarthritis (TMJ OA), because of cost and the radiation associated with computed tomograms (CT). The aims of this study were to develop an artificial intelligence (AI) model and compare its TMJ OA diagnostic performance from OPGs with that of an oromaxillofacial radiology (OMFR) expert. An AI model was developed using Karas' ResNet model and trained to classify images into three categories: normal, indeterminate OA, and definite OA. This study included 1,189 OPG images confirmed by cone-beam CT (CBCT) and evaluated the results by model (accuracy, precision, recall, and F1 score) and diagnostic

performance (accuracy, sensitivity, and specificity). The model performance was unsatisfying when AI was developed with 3 categories. After the indeterminate OA images were reclassified as normal, OA, or omission, the AI diagnosed TMJ OA in a similar manner to an expert and was in most accord with CBCT when the indeterminate OA category was omitted (accuracy: 0.78, sensitivity: 0.73, and specificity: 0.82). Our deep learning model showed a sensitivity equivalent to that of an expert, with a better balance between sensitivity and specificity, which implies that AI can play an important role in primary diagnosis of TMJ OA from OPGs in most general practice clinics where OMFR experts or CT are not available.

Keywords: Temporomandibular joint; Osteoarthritis; Orthopantomogram; Artificial intelligence; Deep learning.

Student Number: 2015-31275

Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram

파노라마 영상에서 측두하악관절 골관절염 진단을 위한 인공지능 연구

Eunhye Choi, D.D.S., M.S.D.

Program in Oral Medicine and Oral Diagnosis, Dept. of Dental Science, Graduate School, Seoul National University

(Directed by Professor **Hee-Kyung Park**, D.D.S., M.S.D., Ph.D.)

최 은 혜

서울대학교 대학원 치의과학과 구강내과·진단학 전공

(지도교수 **박 희 경**)

- CONTENTS -

I. INTRODUCTION

Temporomandibular joint osteoarthritis (TMJ OA) is an important subtype of temporomandibular disorders (TMDs) and may lead to substantial joint pain, dysfunction, dental malocclusion, and reduced health-related quality of life[1]. Osteoarthritis (OA) is a disease of joints caused by a series of degenerative processes including gradual loss of joint cartilage, remodeling and hardening of subchondral bone, and formation of osteoproliferative bodies[2]. TMJ OA is confirmed by structural bony changes observed on radiographic examination. Because of the ability to show minute changes in the integrity of the cortical and subcortical bone of the TMJ, it is widely accepted that computed tomography (CT) is the reference standard for the diagnosis of TMJ OA[3]. Cone-beam CT (CBCT), which has the benefit of lower radiation exposure than conventional CT, is reportedly as accurate for the detection of TMJ OA[4]. However, CBCT is not the first choice for TMJ OA examination in the normal clinical setting yet because it still has a higher radiation dose and is costlier than plain radiographs. Normally, the orthopantomogram (OPG) is the most common examination method used for screening various lesions and conditions in the maxillofacial region, while it is less able to identify

bony changes in the TMJ structure that are small in size and overlapped by other skull structures[5]. This makes OPG useful for screening examinations that experienced experts such as oromaxillofacial radiology (OMFR) or orofacial pain specialists read and then recommend, if necessary, an additional CBCT to confirm a diagnosis. In a situation where such experts are not available, a patient's TMJ OA can be overlooked or misread. To address this possibility, an artificial intelligence (AI) algorithm was developed and trained to read TMJ OA on OPGs based on CBCT results already confirmed by experts. Various studies have applied AI algorithms to read OPG for clinical conditions such as tooth segmentation[6], age estimation[7], third molar and inferior alveolar nerve detection[8], cysts and tumors of the jaw[9], osteoporosis[10], and maxillary sinusitis[11]. However, there are few AI studies on TMJ OA diagnosis; one study that used CBCT and one study based on OPG have been reported[12,13].

This study aimed to investigate the clinical utility of an AI diagnostic tool developed for TMJ OA diagnosis from OPG using deep learning that compared the AI read with that of an expert.

II. REVIEW OF LITERATURE

1. Temporomandibular joint osteoarthritis

TMJ OA, a severe form of TMDs, is a degenerative disease of the joint that can cause persistent orofacial pain as well as functional and structural changes to its bone, cartilage, and ligaments. The spectrum of clinical and pathologic presentation of TMJ OA ranges from structural and functional failure of the joint with disc displacement and degeneration to erosions, osteophytes, loss of articular fibrocartilage, and synovitis. According to the Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) for clinical and research taxonomy, degenerative joint disease (DJD) is described as "a degenerative disorder involving the joint characterized by deterioration of articular tissue with concomitant osseous changes in the condyle and/or articular eminence" [14].

TMJ OA is a local inflammatory condition which occurs when the dynamic equilibrium between the breakdown and the repair of joint tissue is compromised. It could also be a consequence of a dysfunctional articular remodeling due to a decreased adaptive capacity of the articulating structures of the joint[15-17].

Symptoms and signs of TMJ OA are not very different from OA

occurring in other joints, including pain, movement limitations, click and crepitus sounds, and most critically joint deformity that can be identified through radiographic examinations.

Depending on the diagnostic method used, the prevalence of TMJ OA can vary from 1% to 84%[18]. Among the patients with TMDs, 8–12% get a diagnosis of DJD[19]. In studies based on autopsy, degenerative bone changes were observed on the condylar surface of the TMJ in 22–40% of subjects[20–22]. More recently, bone changes in the TMJ were observed in 71% of TMD patients referred to take CBCT examination[23]. The wide discrepancy illustrates that diagnosis is frequently guided by the presence or absence of rather non-specific signs and symptoms[24]. However, if the present or absence of clinical symptom and the cause of bone changes (whether damage or reactive changes) are not taken into account, it can be understood that about 30–60% of general TMJ undergoes some degree of bone change. Among them, the number of patients with clinical symptoms or destructive bone changes is a part of them. If the target is limited to TMD patients with symptoms, the prevalence rate is estimated about 8–12%[19].

2. Radiological examination of TMJ OA

DJD is diagnosed radiographically, as the clinical signs and

symptoms have poor validity[14]. OPG is the most useful and widely used clinical screening tool for the TMJ[25]. However, it is rarely used to evaluate complex pathology due to its poor reliability and sensitivity for detecting osseous changes[26]. OPG provides little information that influences the diagnosis or management for TMJ OA at early stages[27], because dramatic changes are only seen in advanced disease. CT is the preferred diagnostic modality as the reference standard providing excellent resolution of TMJ OA. Although CT provides anatomical detail, it can only detect calcified tissues and does not provide information about soft tissues—the cartilage, disc, and ligaments—that are important in OA. Magnetic resonance images (MRI), which have the highest reliability in reading soft tissues and joint discs, are known to provide sufficient reliability for evaluating bone changes[26]. Additionally, CBCT has a lower radiation dose but has diagnostic accuracy similar to conventional CT[28,29].

According to image analysis criteria for diagnosis of TMD, three cardinal radiographic features that lead to a diagnosis of DJD are osteophyte, surface erosion, and subcortical pseudocyst[26]. These features are defined as follows:

(a) An osteophyte is a marginal hypertrophy with sclerotic borders and exophytic angular formation of osseous tissue

arising from the surface.

(b) Surface erosion is loss of continuity of articular cortex of the condyle or the fossa.

(c) A subcortical pseudocyst is defined as a cavity below the articular surface that deviates from normal marrow pattern. It is not a true cyst but rather the loss of trabeculation.

Other radiographic findings related to possible osseous remodeling are articular surface flattening and subcortical sclerosis. Flattening and subcortical sclerosis may be indeterminate for degenerative joint disease as they may represent aging, functional remodeling of the joints, or a precursor to DJD[4]. Longitudinally, flattening and sclerosis may progress to DJD[30].

(a) A surface flattening is defined as a loss of rounded contour of the surface of the condyle or the articular eminence; this can be present in normal joints and be a variation of normal.

(b) A subcortical sclerosis is defined as any increased thickness of the cortical plate in the load-bearing areas related to the adjacent non-load-bearing areas; this likely results from increased loading, or from normal loading when the disc is displaced.

3. Artificial intelligence and dentistry

Currently, AI is expanding its use in the medical field. Watson developed by IBM, for example, has been used to support doctors' clinical decisions[31]. The diagnostic accuracy of deep learning algorithms is approaching levels of human experts in medical field, changing the role of computer-assisted diagnosis from a 'second-opinion' tool to a more collaborative one[32].

The term "artificial intelligence" originated from the concept that machines can perform human tasks. A subdomain of AI is machine learning (ML), which learns intrinsic statistical patterns in data, eventually making predictions on unseen data. Deep learning is a ML technique using multi-layer mathematical operations to learn and infer complex data like imagery[33]. Deep learning algorithms can classify datasets automatically, and with the use of multilayer convolutional neural networks (CNN) they can deeply learn the features contained within data[34]. They have been used effectively for image-based automated diagnosis in dentistry, including, tooth segmentation[6], age estimation[7], third molar and inferior alveolar nerve detection[8], cysts and tumors of the jaw[9], osteoporosis[10], and maxillary sinusitis[11]. When training image datasets are input into the system, the learning procedures are repeated automatically, without requiring manual definition of the imaging characteristics of the

lesions[35]. In this way, deep learning methods can learn adaptive image features and simultaneously perform image classification[36].

Sherman recognition and classification of objects using CNN-based models are the most popular cases of technology use[37]. However, the simple use of CNNs typically requires a substantial amount of training data and is time-consuming. Nevertheless, the recent advent of an object detection technique that classifies objects through bounding boxes has significantly improved the accuracy of object recognition algorithms[6]. A region-based convolutional neural network (R-CNN), which forms the basis of object detection, is used first to extract the region-of-interest from the candidate group and then to determine the individual algorithm groups.

One advantage of R-CNNs is that they can quickly extract regions with relatively high accuracy, even with smaller datasets. Starting with the R-CNN algorithm in 2013, object detection techniques have developed into various algorithms, including Fast R-CNN, Faster R-CNN, YOLO (You Only Look Once), and RefineDet. Notably, Faster R-CNN is slightly slower than other existing algorithms; however, in terms of accuracy, models based on Faster R-CNN show superior performance, compared to most models[38].

AI algorithms for OPG images are used to obtain various properties in various applications, such as tooth shape, tooth age, dental plaque, landmark detection, orthodontics, and osteoporosis disease. de Tobel et al. proposed a novel algorithm to classify the tooth age of infants and toddlers with OPGs[7]. Their algorithm successfully analyzes images through a basic CNN by analyzing the teeth of infants and toddlers but suffers from manual placement of the bounding box. In a study in 2017, automated tooth detection and numbering was conducted by a CNN that used a heuristic method to detect teeth[39]. Advances in the teeth detection field have led to heuristic techniques, thereby significantly increasing the accuracy of tooth shape determination[40].

An algorithm for diagnosing the jaw region through OPG was first developed in 2018. Specifically, Poedjiastoeti and Suebnukarn proposed an algorithm for identifying and finding jaw tumors with OPG images and further validated the proposed algorithm through activation maps[32]. Further, Ariji et al. conducted a study in which they found and analyzed mandible lesions by using object detection algorithms, including YOLO and DetectNet[42,43].

However, the clinical accuracy of AI in the dental field must be validated with a variety of cases and imaging modalities in spite of the difficulty of standardizing dental radiological examination before

AI can take on a more important role in making diagnostic recommendations. Furthermore, deep learning algorithms have inherent uncertainty like a black box, making it difficult for humans to identify or adjust the criteria used for diagnoses[44].

III. MATERIALS AND METHODS

1. Subjects.

Patients who visited the orofacial pain clinic of Seoul National University Dental Hospital complaining of TMD-related symptoms from January, 2015 to October, 2019 were reviewed. Subjects under 18 years of age or with a history of orthognathic surgery, macro trauma, and systemic diseases that could cause joint deformity were excluded.

The written documentation of informed consent was waived and approved by the decision of the Institutional Review Board of School of Dentistry, Seoul National University (S-D20200004) and ethics committee approval for the study in the same institute was also obtained. All methods were performed in accordance with relevant guidelines and regulations.

2. Datasets.

Based on TMJ CBCT (Dinnova3 CBCT scanner, HDX will Inc., Seoul, Korea) read by OMFR experts, OPG image dataset (Orthopantomograph OP, 100D, Instumentarium Corporation, Tuusula, Finland) was acquired. OPG images with a temporal

difference of more than 3 months with CBCT images were excluded.

The AI algorithm was trained, validated, and tested with 1,189 OPGs, all of which had been confirmed by additional CBCT examination, selected randomly and classified by an orofacial pain specialist according to the OMFR expert's read: no TMJ OA (normal), indeterminate for TMJ OA (indeterminate), and TMJ OA (definite). When OMFR experts read degenerative bone changes on a CT images, they diagnosis TMJ OA based on the presence, severity, and the extent of the image findings such as erosion, subchondral sclerosis, generalized sclerosis, flattening, subchondral defect, osteophyte, and joint mouth. Among 2,378 joints, 800 were diagnosed as normal, 779 as indeterminate, and 799 as definite based on the CBCT reads (Table 1).

3. AI model development.

The AI model development was based on jpg files. First, an algorithm to extract regions of interest (ROI) including the mandibular condyle and surrounding structures from each OPG by object detection was developed (Fig. 2). The object detection technique was based on Faster Regions with Convolutional Neural Networks (RCNN) using the Inception V3 model as the categorizing algorithm in which Region Proposal Network (RPN) and Image

Classification Network (ICN) work simultaneously and make it faster. The RCNN used an algorithm called selective search to extract approximately 2,000 areas where there were likely objects. This technique is called region proposals. For each region, 4,096-dimensional feature vectors were derived using CNN for image classification (Inception ResNet V2 was used in this paper). The CNN model took a 227x227 color image as the input and derived the included characteristics through 5 convolutional layers and 2 fully-connected layers. Therefore, region proposals must be warped to a size of 227 x 227 before putting them into the CNN. Then, a support vector machine was used to predict the class associated with the feature vector. Finally, Bounding-box (BB) regression was performed to determine the location of the objects more accurately. In the next step, extracted ROI images were classified as normal, indeterminate, and definite OA by means of the Keras' ResNet model based on a Convolutional Neural Network (CNN). The ROI images of 2,378 joints were divided randomly into training (1,478 images), validation (450 images), and test sets (450 images), with which the AI algorithm was developed using Keras' ResNet model. The test set consisted of 150 normal, 150 indeterminate, and 150 definite OA images (Fig. 3). Data augmentation was done by image rotation ± 5 degrees, image shift

± 10%, brightness ± 10%, and contrast ± 10% to compensate for the disadvantage of the small number of the data points to increase model robustness. Training and validation were repeated 35,000 times (700 epochs) with augmented data. The learning rate of the model was 1.0×10^{-6} and an Adam optimizer was used. After 700 training epochs, the validation loss of the model decreased from 12.2 to 0.1 (Fig. 4). In order to find the most suitable model for screening TMJ OA, Indeterminate OA was treated as follows during AI model development.

1. Initial trial: Indeterminate OA was treated independently.
2. Trial 1: Indeterminate OA was considered normal.
3. Trial 2: Indeterminate OA was considered TMJ OA.
4. Trial 3: Indeterminate OA was omitted.

In Trial 3, the remaining 1,599 joints, except for indeterminate OA, were divided randomly into training (1,027 images), validation (300 images), and test sets (272 images). The test set consisted of 145 normal and 127 definite OA images.

After selecting the optimal trial, 5-fold cross validation was performed to evaluate model training, while avoiding overfitting or bias. The whole 1,599 images consisting of 800 normal and 799 OA selected as the training dataset were randomly divided into five folds. Within each fold, the dataset was partitioned into independent

training and validation sets, using an 80 to 20 percentage split. The selected validation set was a completely independent fold from the other training folds, and it was used to evaluate model training status during the training. After one model training step was completed, the other independent fold was used as a validation set and the previous validation set was reused, as part of the training set, to evaluate the model training.

4. Model and statistical analysis.

Accuracy, precision, recall, and F1 score were calculated for model performance. Accuracy is defined as the ratio of correct predictions. Precision is the ratio of true positives to true positives and false positives. Recall is the ratio of true positives to true positives and false negatives. Finally, the F1 score is a harmonic mean of precision and recall: (2 × precision × recall)/(precision + recall). Accuracy, specificity, and sensitivity were calculated for diagnostic performance, Cohen's kappa was calculated to estimate the agreement of TMJ OA diagnosis between OPG and CBCT reads, and McNemar's test was done to evaluate the significance of difference. For evaluation of AI clinical usability, the results between OPG reads by the AI and the expert were compared. All p−values < 0.05 were considered to be statistically significant. The

Python programming language (v. 3.6), Tensorflow (v. 2.0) and a graphics card (GeForce GTX 2080) were used for analysis.

III. RESULTS

The results of the AI algorithm that was developed based on 3 categories, normal, indeterminate, and definite OA, are shown in Table 1. The total dataset consisted of 2,378 joint images, including 800 normal (mean age 34.06 ± 14.43 years), 779 indeterminate (mean age 33.11 ± 14.28 years), and 799 definite OA images (mean age 40.23 ± 16.89 years). Because overall accuracy was not satisfying (Table 2), the model development process was modified and reevaluated because indeterminate TMJ OA could have compromised AI training because of its vagueness on radiographic reading[14,26]. The AI was trained, validated, and tested again in 3 ways after modification. Indeterminate OA was considered as normal in Trial 1, OA in Trial 2, and omitted from the whole development process in Trial 3. The accuracy of the model performance was best in Trial 1 (0.80) followed by Trial 3 (0.78) and Trial 2 (0.73), but precision and recall were evenest in Trial 3 (Table 3). The recall value of TMJ OA was 0. 51, which means that the model predicted TMJ OA in patients with actual TMJ OA about half the time in Trial 1. In screening tests, it is important to suspect the presence of the disease so that additional tests can be

performed if necessary. For this reason, Trial 3 was chosen as a more suitable model in spite of the higher overall accuracy seen in Trial 1. Five-fold cross validation was performed on Trial 3. The average accuracy, precision, recall, and F 1 score were 0.76, 0.80, 0.71, and 0.75, respectively (Table 4).

The diagnostic performance of the AI and expert and agreement between their OPG read and CBCT reads is shown in Table 5. The comparison of sensitivities and specificities in Trials 1, 2, and 3 are shown in Figure 1. The AI in Trial 1 (0.80) and the expert in Trial 3 (0.85) were the most accurate, respectively. However, taking sensitivity and specificity into consideration collectively, it can be said that Trial 3 was the most accurate (0.85 for the expert and 0.78 for the AI). Cohen's kappa was highest as well in Trial 3 and showed a substantial level of agreement for the expert (0.69) and moderate agreement for the AI (0.56). In all 3 trials, the expert read more accurately than the AI. However, the result of McNemar's test showed that AI reads were more in accord with CBCT (p=0.366) in Trial 3 where TMJ OA was diagnosed dichotomously.

IV. DISCUSSIONS

OPG is the most widely used plain radiograph method for the primary diagnosis of TMJ OA. However, OMFR experts or CBCT are not always available in most general practice clinics and screening for TMJ OA is easily compromised. As reported previously, age, pain, and TMJ noise were not correlated with TMJ OA on CBCT, while a high incidence rate for OA changes in TMD patients was observed (27.3%)[45]. It was also reported that 24% of patients who did not show significant condyle bone changes on OPG had degenerative bone changes on CBCT[46]. Moreover, it is also very well known that the accuracy, sensitivity, and specificity of OPG is not good for the diagnosis of TMJ OA, even when an expert OMFR radiologist reads OPG images. OPG showed lower sensitivity (0.26) and higher specificity (0.99) compared to CT in TMJ OA patients[26]. The diagnostic accuracy of OPG for detecting cortical erosion of the mandibular condyle is less (0.55–0.64) compared to CBCT (0.77–0.95)[47]. Because the bone tissue must be demineralized sufficiently before becoming noticeable on OPG, which usually takes more than 6 months, the mandibular condyle may appear normal on OPG in the early stages of TMJ OA[19,48].

According to the position paper from the American Academy of Oral and Maxillofacial Radiology (AAOMR), OPG is only useful for diagnosing advanced TMJ OA due to its low sensitivity[49]. Image distortion and overlap on OPG are always concerns as well[50]. However, considering the prominent role of OPG in primary examinations, any supplementary diagnostic tool to screen for TMJ OA on OPG would be very helpful in clinics.

The AI algorithm developed in this study showed sufficient sensitivity compared to OMFR experts for the primary diagnosis of TMJ OA on OPG. The diagnostic performance (accuracy, 0.51, and F1 score, 0.53) was not satisfying in the initial trial when AI was trained with 3 categories of TMJ OA: normal, indeterminate, and definite. The Cohen's kappa value for the AI diagnosis on OPG was 0.27, which was less than the that of the expert, 0.38, while both represent fair agreement. This implies that it is still difficult to distinguish subtle changes in TMJ OA on OPG even for an expert. Besides, multi-label image classification is more challenging than single-label classification[51] as shown in a previous study that reported low performance (mean accuracy, 0.51) for AI classification of lower third molar development into multiple stages for age estimations[7]. Based on this initial result, classification of images was done in different ways.

When indeterminate OA was taken as normal in Trial 1, the AI model performance was best, and showed the highest accuracy, precision, recall, and F1 score (0.80, 0.81, 0.80, and 0.80, respectively). But, Trial 1' s sensitivity in diagnostic performance was lowest (0.52). In Trial 2 when indeterminate OA was defined as definite OA, the model performance (accuracy, 0.73, precision, 0.75, recall, 0.73, and F1 score, 0.74) and specificity (0.23) in diagnostic performance were lowest, while sensitivity was highest (0.97). In Trial 3 when indeterminate OA was omitted from the development process, the model performance (accuracy, 0.78, precision, 0.78, recall, 0.78, and F1 score, 0.78) came close to that in Trial 1, while Trial 3' s sensitivity (0.73) and specificity (0.82) were good and more balanced in terms of diagnostic performance. Based on this result, AI performed best for TMJ OA diagnosis on OPG when indeterminate OA was omitted during training and verification. The pilot study showed similar accuracy (0.77–0.84) with various AI models when indeterminate OA was omitted[13].

Indeterminate OA was recategorized as normal in previous studies[12,26] as in Trial 1 of this study. The sensitivity of the AI (0.52) and expert (0.61) was higher and specificity was lower (0.94 for AI and 0.91 for expert) than those of experts in the previous study (sensitivity, 0.26 and specificity, 0.99)[26]. The model

performance (accuracy, 0.86 and F1 score, 0.84) in a recent study[12] was better than in Trial 1 of this study (accuracy, 0.80 and F1 score, 0.80) but the materials for AI development were CBCT sagittal images in that previous study. CBCT images have higher detail and fewer artifacts at the anatomical boundaries of the ROIs and background than OPGs[52]. The superiority of CBCT to OPG in the performance of a DCNN model has already been reported[53].

Consequently, the AI algorithm in Trial 3 is the most appropriate for TMJ OA diagnosis on OPG, and showed the best balance between sensitivity and specificity among our 3 trials and equivalent diagnostic performance to the expert. Moreover, no statistical difference was observed between the AI diagnosis on OPGs and the expert diagnosis on CBCT only in Trial 3 (McNemar' s test, p=0.336), which must be because of a better balance between sensitivity and specificity. This implies that the AI model is more likely to accurately diagnose TMJ OA on OPG in accord with expert diagnosis using CBCT when indeterminate TMJ OA is excluded from AI training. When the AI in Trial 3 read 115 untested OPG images of indeterminate TMJ OA (data not shown), AI was more likely to read indeterminate TMJ OA as TMJ OA (41 OPGs, 35.7%) than the expert (23 OPGs, 20.0%). Indeterminate TMJ OA may be considered a normal variation, aging, physiologic

remodeling, or a precursor to TMJ OA[14], which means it can be diagnosed as normal and OA at the same time. Because the benefit of high sensitivity may exceed loss of low specificity in the diagnosis of TMJ OA where early detection is important, it would be clinically more beneficial for a TMJ OA screening tool to read indeterminate OA as OA. In a randomized controlled study, osseous condylar changes in adolescents/young adults with early-stage TMJ OA showed repair and even regeneration after conservative and splint therapy[54], which may emphasize the need for early detection and management.

Additional statistics were performed to explain Trial 3' s model prediction—which included factors correlated with the expert or AI' s diagnosis of TMJ OA based on the OMFR expert' s CBCT reads. Location and types of degenerative bone changes based on CBCT reads by the expert were evaluated by Mann-Whitney test (Table 6). The expert' s diagnosis of TMJ OA was correlated with surface erosion (p=0.001) and generalized sclerosis (p=0.040), while AI' s diagnosis was correlated with surface irregularity (p=0.033). In this study, AI was trained to learn the whole image with the extracted ROIs, and as a result, it might be said that image overall appearance could influence TMJ OA diagnosis rather than specific OA changes. This may explain why AI was closer to CBCT

than the expert in diagnosing TMJ OA with OPGs.

AI is an up-and-coming method for use in radiology, where a large amount of homogeneous image data are available, and this includes AI's strengths in segmenting, detecting, or classifying organs and lesions. Despite the lack of quality and number of studies, AI has been constantly shown to be equivalent to health-care professionals with respect to diagnostic performance[55]. On the other hand, AI is more likely to show better results when it is trained with an ROI image than with a whole image[56]. However, ROIs are extracted manually in most studies, which can compromise AI's practical usability. In this study, an AI algorithm was designed to extract ROIs from OPG images based on an object detection technique. The average precision of mandibular condyle detection was reported to be 99.4% on right side and 100% on left side in a previous study[13]. The ROI extracted automatically for this study covered the mandibular condyle, and the articular fossa and eminence, which is very important because TMJ OA is not defined only by bony changes in the condyle. It is obvious that OA changes occur in other structures of the TMJ such as the fossa and eminence[57], so that the ROI for AI diagnosis should cover all structures of the joint, rather than just the condyle. But, most TMJ OA studies usually focus on the condyle and only one AI study

based on TMJ CBCT also used the mandibular condyle as extracted from sagittal sections[12].

This study has several limitations. First, the absolute size of the training dataset was rather small. The use of a larger dataset for training and/or external validation may render different findings. In order to overcome the limitation of a small size dataset, data augmentation by image modification and a transfer learning technique with fine-tuning based on an imagenet database (http://image-net.org) were used. It has been reported that the accuracy of transfer learning increases to some extent for medical images rather than general objects[41]. Although AI did not exceed an expert with respect to accuracy in this study, this may improve in further studies with larger datasets, learning OPG and CBCT images together, elaborating deep learning models using landmarks, and an ensemble composed of multiple AI models. Second, interexaminer reliability among OMFR experts could not be estimated in this retrospective study. Third, it is well known that an external test dataset drawn from multiple medical centers secures high AI performance for reproducibility in general use. However, this study was based on an internal dataset, but, recently, there is no noticeable difference in performance between OPG devices from various manufacturers. Last, deep learning algorithms have inherent

uncertainty[58]. Compared to machine learning methods based on handcrafted features, the trained deep learning model is like a "black-box" so the results are not explainable[59].

V. CONCLUSIONS

Collectively, an AI may read OPGs to diagnose TMJ OA as well as OMFR experts can with respect to sensitivity and has greater accordance with CBCT interpretations than do OMFR experts, which implies that AI can play an important role in diagnosing TMJ OA primarily from OPGs in most general practice clinics, where OMFR experts or CBCT are not available.

REFRENCES

1. Wang, X., Zhang, J., Gan, Y. & Zhou, Y. Current understanding of pathogenesis and treatment of TMJ osteoarthritis. *J Dent Res.* **94**, 666-673 (2015).
2. Buckwalter, J. A., Mankin, H. J. & Grodzinsky, A. Articular cartilage: degeneration and osteoarthritis. *Instr Course Lect.* **54**, 465 (2005).
3. Schiffman, E. L. *et al.* The research diagnostic criteria for temporomandibular disorders. I: overview and methodology for assessment of validity. *J. Oral Facial Pain Headache*. **24**, 7 (2010).
4. Larheim, T., Abrahamsson, A., Kristensen, M. & Arvidsson, L. Temporomandibular joint diagnostics using CBCT. *Dentomaxillofacial Radiol.* **44**, 20140235 (2015).
5. White SC, Pharaoah MJ. Oral radiology: principles and interpretation(4th edition) 493-498 (Mosby, 2009).
6. Wirtz, A., Mirashi, S. G. & Wesarg, S. Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network. *Med Image Comput Comput Assist Interv.* **11073**, 712-719 (2018).
7. De Tobel, J., Radesh, P., Vandermeulen, D. & Thevissen, P. W.

An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *J Forensic odontostomatol.* **35**, 42 (2017).

8. Vinayahalingam, S., Xi, T., Bergé, S., Maal, T. & de Jong, G. Automated detection of third molars and mandibular nerve by deep learning. *Sci Rep.* **9**, 1–7 (2019).
9. Yang, H. *et al.* Deep learning for automated detection of cyst and tumors of the jaw in panoramic radiographs. *J Clin Med.* **9**, 1839 (2020).
10. Lee, K.–S., Jung, S.–K., Ryu, J.–J., Shin, S.–W. & Choi, J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. *J Clin Med.* **9**, 392 (2020).
11. Murata, M. *et al.* Deep–learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral radiol.* **35**, 301–307 (2019).
12. Lee, K. *et al.* Automated detection of TMJ osteoarthritis based on artificial intelligence. *J Dent Res.* **99**, 1363–1367 (2020).
13. Kim, D., Choi, E., Jeong, H. G., Chang, J. & Youm, S. J. Expert system for mandibular condylar detection and osteoarthritis classification in panoramic imaging using R–CNN and CNN. *App Sci.* **10**, 7464 (2020).

14. Schiffman, E. *et al.* Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: recommendations of the International RDC/TMD consortium network and orofacial pain special interest group. *J Oral Facial Pain Headache.* **28**, 6 (2014).

15. Arnett, G., Milam, S. & Gottesman, L. Progressive mandibular retrusion—idiopathic condylar resorption. Part I. *Am J Orthod Dentofacial Orthop.* **110**, 8–15 (1996).

16. Nitzan, D. W. The process of lubrication impairment and its involvement in temporomandibular joint disc displacement: a theoretical concept. *J Oral Maxillofac Surg.* **59**, 36–45 (2001).

17. Stegenga, B., de Bont, L. G. & Boering, G. Osteoarthrosis as the cause of craniomandibular pain and dysfunction: a unifying concept. *J Oral Maxillofac Surg.* **47**, 249–256 (1989).

18. de Souza, R. F., Da Silva, C. H. L., Nasser, M., Fedorowicz, Z. & Al-Muharraqi, M. A. Interventions for managing temporomandibular joint osteoarthritis. *Cochrane Database Syst Rev.* **4** (2012).

19. Kamelchuk, L. S. & Major, P. W. Degenerative disease of the temporomandibular joint. *J Orofac Pain.* **9** (1995).

20. Axelsson, S. Human and experimental osteoarthrosis of the

temporomandibular joint. Morphological and biochemical studies. *Swed Dent J Suppl.* **92**, 1–45 (1993).

21. Widmalm, S. E. *et al.* Temporomandibular joint pathosis related to sex, age, and dentition in autopsy material. *Oral Surg Oral Med Oral Pathol.* **78**, 416–425 (1994).

22. Magnusson, C., Ernberg, M. & Magnusson, T. A description of a contemporary human skull material in respect of age, gender, temporomandibular joint changes, and some dental variables. *Swed Dent J.* **32**, 69 (2008).

23. dos Anjos Pontual, M. *et al.* Evaluation of bone changes in the temporomandibular joint using cone beam CT. *Dentomaxillofac Radiol.* **41**, 24–29 (2012).

24. Haskin, C. L., Milam, S. B. & Cameron, I. L. Pathogenesis of degenerative joint disease in the human temporomandibular joint. *Crit Rev Oral Biol Med.* **6**, 248–277 (1995).

25. Scrivani, S. J., Keith, D. A. & Kaban, L. B. Temporomandibular disorders. *N Engl J Med.* **359**, 2693–2705 (2008).

26. Ahmad, M. *et al.* Research diagnostic criteria for temporomandibular disorders (RDC/TMD): development of image analysis criteria and examiner reliability for image

analysis. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod.* 107, 844–860 (2009).

27. Huumonen, S., Sipilä, K., Zitting, P. & Raustia, A. Panoramic findings in 34-year-old subjects with facial pain and pain-free controls. *J Oral Rehabil.* 34, 456–462 (2007).

28. Honda, K., Larheim, T., Maruhashi, K., Matsumoto, K. & Iwai, K. Osseous abnormalities of the mandibular condyle: diagnostic reliability of cone beam computed tomography compared with helical computed tomography based on an autopsy material. *Dentomaxillofac Radiol.* 35, 152–157 (2006).

29. Hintze, H., Wiese, M. & Wenzel, A. J. D. R. Cone beam CT and conventional tomography for the detection of morphological temporomandibular joint changes. *Dentomaxillofac Radiol.* 36, 192–197 (2007).

30. Ahmad, M. & Schiffman, E. L. Temporomandibular joint disorders and orofacial pain. *Dent Clin North Am.* 60, 105 (2016).

31. Chen, Y., Argentinis, J. E. & Weber, G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clin Ther.* 38, 688–701 (2016).

32. Burt, J. R. *et al.* Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol.* **91**, 20170545 (2018).

33. Schwendicke, F., Samek, W. & Krois, J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res.* **99**, 769-774 (2020).

34. Xue, Y., Zhang, R., Deng, Y., Chen, K. & Jiang, T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PLos One.* **12**, e0178992 (2017).

35. Krizhevsky, A., Sutskever, I. & Hinton, G. E. J. A. i. n. i. p. s. Imagenet classification with deep convolutional neural networks. **25**, 1097-1105 (2012).

36. Wang, X. *et al.* Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Sci Rep.* **7**, 1-8 (2017).

37. Guo, Y. *et al.* Deep learning for visual understanding: A review. *Neurocomputing.* **187**, 27-48 (2016).

38. Gupta, A., Puri, R., Verma, M., Gunjyal, S. & Kumar, A. Performance comparison of object detection algorithms with different feature extractors. *IEEE.* In: *2019 6th International*

Conference on Signal Processing and Integrated Networks (SPIN). 472–477 (2019)

39. Oktay, A. B. Tooth detection with convolutional neural networks. *IEEE.* In: *2017 Medical Technologies National Congress (TIPTEKNO).* 1–4 (2017).

40. Mahoor, M. H. & Abdel-Mottaleb, M. Classification and numbering of teeth in dental bitewing images. *Pattern Recognition.* **38**, 577–586 (2005).

41. Poedjiastoeti, W. & Suebnukarn, S. Application of convolutional neural network in the diagnosis of jaw tumors. *Health Inform Res.* **24**, 236–241 (2018).

42. Ariji, Y. *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.* **128**, 424–430 (2019).

43. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 779–788 (2016).

44. Anifowose, F. A. Artificial intelligence application in reservoir

characterization and modeling: whitening the black Box. In: *SPE Saudi Arabia section Young Professionals Technical Symposium.* Society of Petroleum Engineers. (2011).

45. Kim, K., Wojczyńska, A. & Lee, J.-Y. The incidence of osteoarthritic change on computed tomography of Korean temporomandibular disorder patients diagnosed by RDC/TMD; a retrospective study. *Acta Odontol Scand.* **74**, 337-342 (2016).
46. Jeon, Y.-M. *et al.* The validity of computed tomography in diagnosis of temporomandibular joint osteoarthritis. *J Oral Med.* **33**, 195-204 (2008).
47. Honey, O. B. *et al.* Accuracy of cone-beam computed tomography imaging of the temporomandibular joint: comparisons with panoramic radiology and linear tomography. *Am J Orthod Dentofacial Orthop.* **132**, 429-438 (2007).
48. Cibere, J. Do we need radiographs to diagnose osteoarthritis? *Best Pract Res Clin Rheumatol.* **20**, 27-38 (2006).
49. Brooks, S. L. *et al.* Imaging of the temporomandibular joint: a position paper of the American Academy of Oral and Maxillofacial Radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod.* **83**, 609-618 (1997).
50. Mawani, F. *et al.* Condylar shape analysis using panoramic radiography units and conventional tomography. *Oral Surg Oral*

Med Oral Pathol Oral Radiol. **99**, 341-348 (2005).

51. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248-255 (2009).
52. Scarfe, W. C., Farman, A. G. & Sukovic, P. Clinical applications of cone-beam computed tomography in dental practice. *J Can Dent Assoc.* **72**, 75 (2006).
53. Lee, J. H., Kim, D. H. & Jeong, S. N. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral Dis.* **26**, 152-158 (2020).
54. Lei, J., Yap, A. U. J., Liu, M. Q. & Fu, K. Y. Condylar repair and regeneration in adolescents/young adults with early-stage degenerative temporomandibular joint disease: A randomised controlled study. *J Oral Rehab.* **46**, 704-714 (2019).
55. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* **1**, e271-e297 (2019).
56. Do, S., Song, K. D. & Chung, J. W. Basics of Deep Learning: A Radiologist's Guide to Understanding Published Radiology Articles on Deep Learning. *Korean J Radiol.* **21**, 33-41 (2020).

57. O'Ryan, F. & Epker, B. N. Temporomandibular joint fuction and morphology: observations on the spectra of normalcy. *Oral Surg Oral Med Oral Pathol.* 58, 272-279 (1984).

58. Gal, Y. Uncertainty in deep learning. *Los Altos: IEEE/ACM Transactions on Audio, Speech, and Language Precessing.* (2017).

59. Holzinger, A. From machine learning to explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA). Piscataway: IEEE.* 55-66 (2018).

Table 1. Clinical and demographic characteristics of the OPG dataset.

		Normal	Indeterminate	Definite
Female	Number of joints	619	656	683
	Mean age (±SD)	34.70 (±14.43)	34.03 (±14.54)	41.48 (±16.66)
	95% CI	33.56–35.83	32.92–35.14	40.23–42.73
Male	Number of joints	181	123	116
	Mean age (±SD)	31.86 (±14.27)	28.20 (±11.78)	32.88 (±16.50)
	95% CI	29.78–33.94	26.11–30.28	29.88–35.88
Total	Number of joints	800	779	799
	Mean age (±SD)	34.06 (±14.43)	33.11 (±14.28)	40.23 (±16.89)
	95% CI	33.06–35.05	32.11–34.11	39.06–41.40

Abbreviations: OPG, orthopantomogram; Indeterminate, indeterminate temporomandibular joint osteoarthritis; Definite, temporomandibular joint osteoarthritis.

Table 2. Confusion matrix and model performance for the initial AI.

Confusion matrix				Model performance					
		Predicted		Precision	Recall	Accuracy	Weighted average precision	Weighted average recall	F1 score
Actual	Normal	Indeterminate	Definite						
Normal	57	46	47	0.72	0.38	0.51	0.55	0.51	0.53
Indeterminate	14	53	83	0.44	0.35				
Definite	8	22	120	0.48	0.80				

Abbreviations: AI, artificial intelligence; Indeterminate, indeterminate temporomandibular joint osteoarthritis; definite, temporomandibular joint osteoarthritis.

Table 3. Confusion matrix and model performance in each Trial.

Confusion matrix				Model performance					
		Predicted		Precision	Recall	Accuracy	Weighted average precision	Weighted average recall	F1 score
	Actual	Normal	OA						
Trial 1	Normal	283	17	0.80	0.94	0.80	0.81	0.80	0.80
	OA	72	78	0.82	0.52				
Trial 2	Normal	35	115	0.81	0.23	0.73	0.75	0.73	0.74
	OA	8	292	0.72	0.97				
Trial 3	Normal	119	26	0.78	0.82	0.78	0.78	0.78	0.78
	OA	34	93	0.78	0.73				

Abbreviations: OA, temporomandibular joint osteoarthritis

Table 4. Five-fold cross-validation in Trial 3.

Work	Precision	Recall	F1 score	Accuracy	AUC (95% CI)
1	0.82	0.59	0.69	0.74	0.83 (0.79-0.88)
2	0.80	0.71	0.75	0.77	0.86 (0.82-0.90)
3	0.83	0.76	0.79	0.81	0.87 (0.83-0.91)
4	0.75	0.76	0.75	0.75	0.83 (0.79-0.88)
5	0.77	0.74	0.76	0.76	0.84 (0.80-0.89)
Average	0.80	0.71	0.75	0.76	0.85 (0.81-0.89)

Table 5. Diagnostic performance and level of agreement in each Trial.

		Diagnostic performance			Cohen's kappa	Kappa index	McNemar's test
		Accuracy	Sensitivity	Specificity			
Trial 1	Expert	0.81	0.61	0.91	0.54	moderate	.001
	AI	0.80	0.52	0.94	0.51	moderate	.000
Trial 2	Expert	0.69	0.57	0.93	0.42	moderate	.000
	AI	0.73	0.97	0.23	0.25	fair	.000
Trial 3	Expert	0.85	0.72	0.97	0.69	substantial	.000
	AI	0.78	0.73	0.82	0.56	moderate	.366

Abbreviations: AI, artificial intelligence.

Table 6. Mann–Whitney test to explain Trial 3’ s model prediction between expert and AI.

Location	Factors	*p* value of expert	*p* value of AI
Condylar head	surface irregularity	.824	.033
	unclear cortical line	.904	.554
	articular surface flattening	.254	.161
	subcortical sclerosis	.747	.976
	subcortical cyst	.546	.504
	surface erosion	.001	.569
	osteophyte	.314	.174
	generalized sclerosis	.040	.336
Fossa	articular surface flattening subcortical sclerosis surface erosion	.716	.730
Eminence	articular surface flattening subcortical sclerosis surface erosion	.689	.461

Abbreviations: AI, artificial intelligence.

Figure legends

Figure 1. Comparison of the sensitivities and specificities in Trial 1, 2, and 3.

Figure 2. Result of ROI extraction, 300x300 pixels.

Figure 3. Clinical datasets used for training, validation, and test.

Figure 4. Change in validation loss as a result of model training. Validation loss decreased as the epochs increased.

Figure 1.

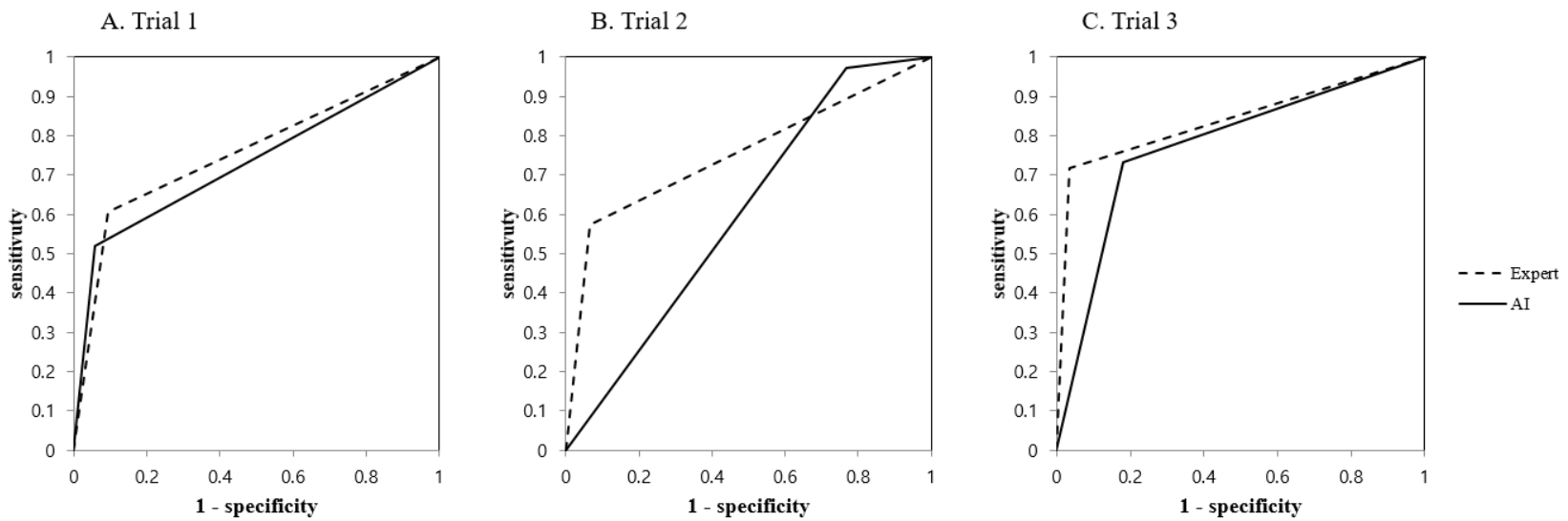

A. Trial 1
B. Trial 2
C. Trial 3
sensitivuty
1 - specificity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Expert
AI

Figure 2.

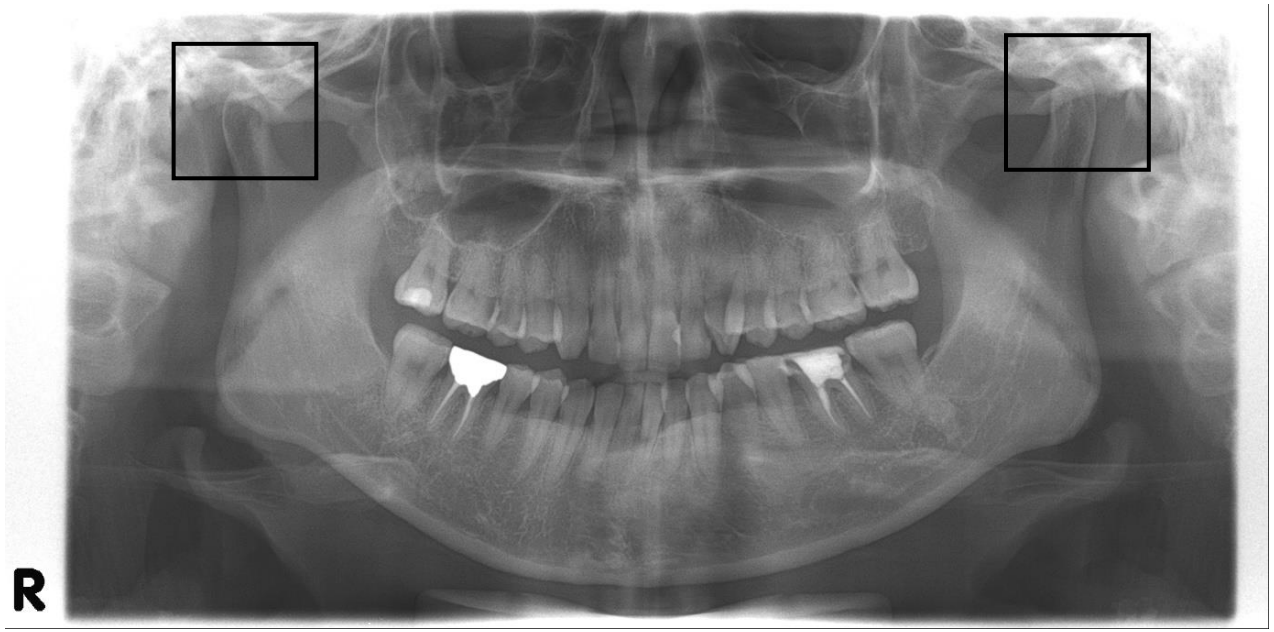


Figure 3.

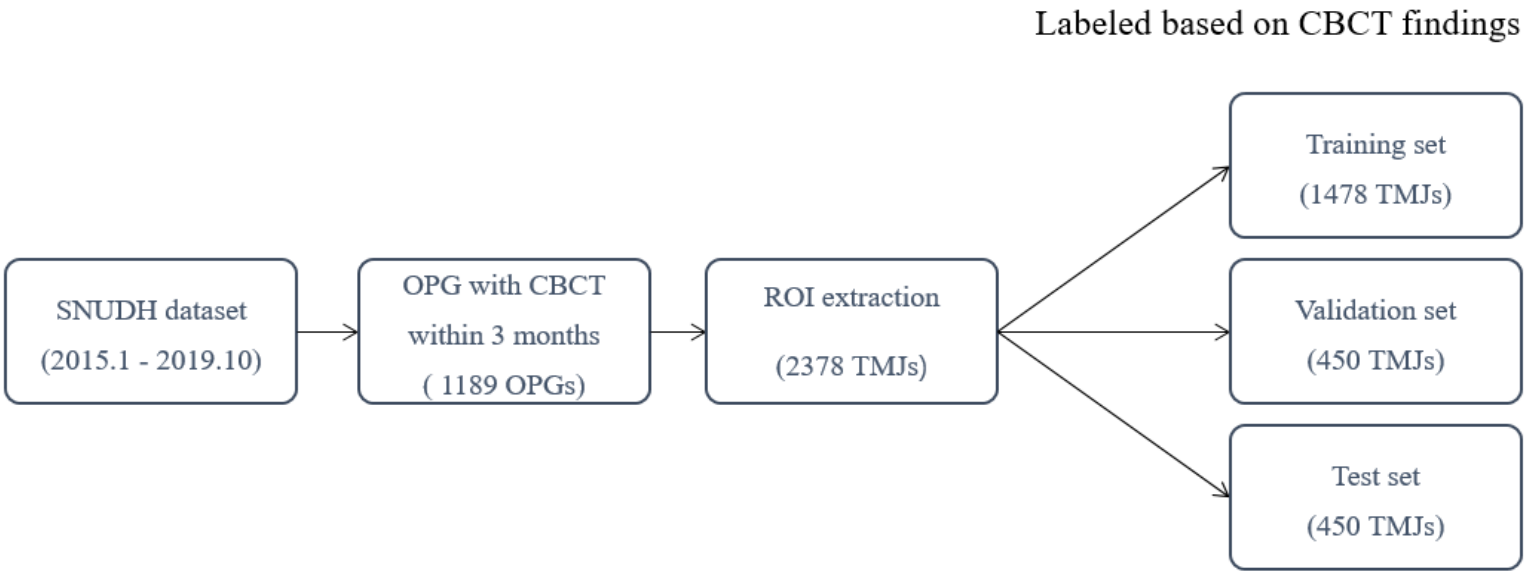
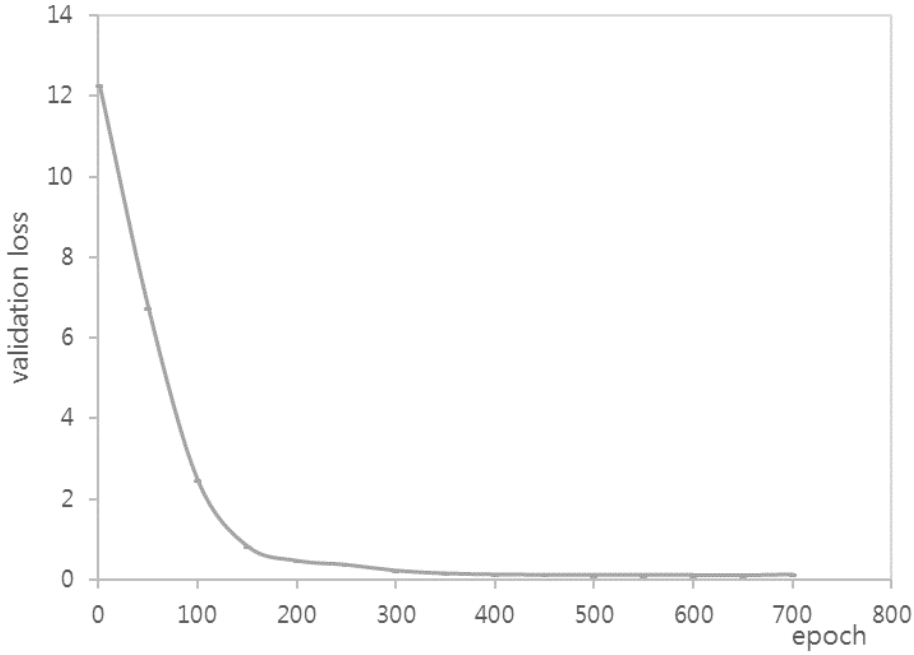

Labeled based on CBCT findings
SNUDH dataset
(2015.1 - 2019.10)
OPG with CBCT
within 3 months
(1189 OPGs)
ROI extraction
(2378 TMJs)
Training set
(1478 TMJs)
Validation set
(450 TMJs)
Test set
(450 TMJs)

Figure 4.


14
12
10
8
6
4
2
0
validation loss
0
100
200
300
400
500
600
700
800
epoch

국문초록

파노라마 영상에서 측두하악 골관절염 진단을 위한 인공지능 연구

최 은 혜

서울대학교 대학원 치의과학과 구강내과·진단학 전공

(지도교수 **박 희 경**)

파노라마 영상은 측두하악관절 골관절염의 일차 진단에서 일반적으로 사용된다. 본 연구의 목적은 파노라마 영상을 이용하여 측두하악관절 골관절염을 진단하는 인공지능 모델을 개발하고, 그 진단 성능을 영상치의학 전문가와 비교하는 것이다.

Karas' ResNet 모델을 사용하여 인공지능 모델을 개발하였다. 파노라마 영상을 no osteoarthritis, indeterminate osteoarthritis 및 definite osteoarthritis의 세 가지 범주로 분류하도록 훈련했다. 본 연구는 cone beam CT로 확인된 1,189개의 파노라마 영상을 포함하고 있으며, 모델 성능(정확도, 정밀도, 재현율 및 F1 score) 및 진단 성능(정확도, 민감도 및 특이도)을 평가했다.

인공지능 모델이 세 가지 범주를 가지고 개발되었을 때, 측두하악관절 골관절염 진단 모델의 정확도, 정밀도, 재현율, F1 score는 각각 0.51, 0.55, 0.51, 0.53의 결과값을 나타냈다. 따라서 Indeterminate for osteoarthritis 영상을 no osteoarthritis(Trial 1), osteoarthritis(Trial 2) 및 생략(Trial 3)으로 재분류하여 평가하였다. Trial 3의 경우 인공지능 모델의 정확도, 민감도, 특이도는 각각 0.78, 0.73, 0.82 였으며, 영

상치의학 전문가는 각각 0.85, 0.72, 0.97이었다. 또한 McNemar's test를 이용하여 인공지능 모델과 CBCT 결과를 비교한 결과 통계적으로 유의한 차이를 보이지 않았다(p value = 0.366).

본 연구의 인공지능 모델은 측두하악관절 골관절염 진단에 있어서 전문가에 상응하는 민감도와 특이도를 보인다. 이는 영상 치의학전문의가 부재하거나 CT 촬영이 불가능한 상황에서 파노라마 영상을 기반으로 측두하악관절 골관절염의 일차 진단이 이뤄질 때, 인공지능이 중요한 역할을 할 수 있다는 것을 보여준다.

주요어: temporomandibular joint, arthritis, orthopantomogram, artificial intelligence, deep learning.

학 번: 2015-31275