Original Article

Yonsei Med J 2021 Dec;62(12):1125-1135
https://doi.org/10.3349/ymj.2021.62.12.1125

Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Effective End-to-End Deep Learning Process in Medical Imaging Using Independent Task Learning: Application for Diagnosis of Maxillary Sinusitis

Jang-Hoon Oh[1]*, Hyug-Gi Kim[2]*, Kyung Mi Lee[2], Chang-Woo Ryu[3],
Soonchan Park[3], Ji Hye Jang[4], Hyun Seok Choi[5], and Eui Jong Kim[2]

[1]Department of Biomedical Science and Technology, Graduate School, Kyung Hee University, Seoul;
[2]Department of Radiology, Kyung Hee University College of Medicine, Kyung Hee University Hospital, Seoul;
[3]Department of Radiology, Kyung Hee University College of Medicine, Kyung Hee University Hospital at Gangdong, Seoul;
[4]Department of Radiology, Korea Cancer Center Hospital, Seoul;
[5]Department of Radiology, Seoul Medical Center, Seoul, Korea.

Purpose: This study aimed to propose an effective end-to-end process in medical imaging using an independent task learning (ITL) algorithm and to evaluate its performance in maxillary sinusitis applications.
Materials and Methods: For the internal dataset, 2122 Waters' view X-ray images, which included 1376 normal and 746 sinusitis images, were divided into training (n=1824) and test (n=298) datasets. For external validation, 700 images, including 379 normal and 321 sinusitis images, from three different institutions were evaluated. To develop the automatic diagnosis system algorithm, four processing steps were performed: 1) preprocessing for ITL, 2) facial patch detection, 3) maxillary sinusitis detection, and 4) a localization report with the sinusitis detector.
Results: The accuracy of facial patch detection, which was the first step in the end-to-end algorithm, was 100%, 100%, 99.5%, and 97.5% for the internal set and external validation sets #1, #2, and #3, respectively. The accuracy and area under the receiver operating characteristic curve (AUC) of maxillary sinusitis detection were 88.93% (0.89), 91.67% (0.90), 90.45% (0.86), and 85.13% (0.85) for the internal set and external validation sets #1, #2, and #3, respectively. The accuracy and AUC of the fully automatic sinusitis diagnosis system, including site localization, were 79.87% (0.80), 84.67% (0.82), 83.92% (0.82), and 73.85% (0.74) for the internal set and external validation sets #1, #2, and #3, respectively.
Conclusion: ITL application for maxillary sinusitis showed reasonable performance in internal and external validation tests, compared with applications used in previous studies.

Key Words: Machine learning, deep learning, artificial intelligence, neural networks, computer, sinusitis

INTRODUCTION

The efficiency of deep learning based on feature recognition has been proven, and its performance in various medical applications, especially in medical imaging of the breast,[1-3] chest,[4,5] brain,[6,7] and other parts of the body,[8-11] has improved. Accordingly, the use of deep learning in medical imaging is progressing not only in research but also in the medical industry. Food and Drug Administrations in several countries have released guidelines for the regulation of artificial intelligence (AI)-based software as medical devices.[12] The primary prerequisite is that AI-based medical devices should provide information on how

Received: June 3, 2021 **Revised:** August 25, 2021
Accepted: September 27, 2021
Corresponding author: Kyung Mi Lee, MD, PhD, Department of Radiology, Kyung Hee University College of Medicine, Kyung Hee University Hospital, 23 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, Korea.
Tel: 82-2-958-8623, Fax: 82-2-968-0787, E-mail: bandilee@khu.ac.kr

*Jang-Hoon Oh and Hyug-Gi Kim contributed equally to this work.
•The authors have no potential conflicts of interest to disclose.

to help medical doctors or patients in diagnosis, and studies[13,14] on chest radiographs support the performance of such software in disease classification and detection.

Paranasal sinusitis is an inflammation of the mucosal lining of the paranasal sinuses (PNSs) and is a common clinical problem among the general population. Recently published studies[15-17] have highlighted the performance of deep learning algorithms in the classification of maxillary sinusitis, comparing performance between deep learning models and radiologists and generating an activation map to explain how deep learning models evaluate a result. However, practical use of AI models from previous studies has been limited due to the following reasons: First, most studies involved time-consuming preprocessing steps, such as handcraft-based patch process for images. Second, developed algorithms have shown a high level of reliability on internal datasets, but low accuracy and sensitivity in external validation sets, which limits their clinical application. In addition, the major tasks involve minimizing overfitting for generalization and explainability to make reasonable decisions, such as the confidence score and localization information.

To bolster the clinical impact of deep learning models in the medical field, inclusion of an end-to-end (E2E) process that proceeds a fully automatic process from input to output is becoming increasingly important[18] to obtaining simpler and more explainable models. To increase high performance of deep learning model, published models was used complex model such as ensemble approach with several deep learning models,[19,20] two or three phases approach,[21-23] and various pre- or post-processing steps.[24] These increase in the number of steps added to the evaluation model that classification, detection, and segmentation for specific disease is difficult because it is necessary to consider not only the optimization of each step but also harmony with other steps. In addition, there is a problem that it is difficult to explain about the final result due to complicated step-by-step process. Therefore, the E2E process with as simple model as possible is effective when considering the needs for explainable model, recently.

In this study, we propose an E2E-based process for medical imaging using the independent task learning (ITL) algorithm that is applicable in real clinics for diagnosing maxillary sinusitis on conventional X-ray images. The ITL algorithm is concatenated with multiple steps, such as role-based processing or deep learning, as detailed later, that can overcome issues with manpower and time-consuming processes by using fully automatic processes, generalization by minimizing bias of images from different institutions and uninterpretable problems by generating a conclusion that is concatenated reasonable results of each step. Furthermore, to overcome limitations with classifying maxillary sinusitis based on deep learning in previous studies,[15-17] our proposed method was designed to be applied for maxillary sinusitis using X-ray imaging for reliable validation of its effectiveness.

MATERIALS AND METHODS

Independent task learning

The purposed ITL algorithm was designed "end-to-end" based on similar processes as in a radiologist's review process of medical images. Fig. 1 shows the four ITL processes in the E2E workflow. The first step involves pre-processing, such as normalization and other processes suitable for lesion characteristics. The second step is to extract a region containing a lesion using an object detector model (patch detector model). In this step, the input is a pre-processed image, and the output is a cropped patch image and the coordinate axes of the cropped image. The third step involves recognizing ambiguous features using a lesion-focused detector (lesion detector model): the input is the cropped patch image from the second step, and the output is a lesion-focused box-based area. The coordinate axes are set from the extracted patch by the lesion detector model. The final step comprises post-processing with the recognized lesion localization and provides a report with the results of the model. The input data are two patch images and two coordinate axes from the patch and lesion detector models. The output is the lesion detection result of the input image from the third step and clinical reports.

Application with medical imaging

Workflow for the application of maxillary sinusitis

The entire workflow of sinusitis diagnosis by a radiologist is a chain process of finding the facial region, adjusting the window level and width, lesion detection, disease diagnosis, localization, and writing a clinical report. The application for maxillary sinusitis based on ITL algorithm was implemented in four steps (Fig. 2). Inverted digital imaging and communication in medicine (DICOM) image correction, intensity normalization (0–1), and resizing (224 by 224) are performed as a

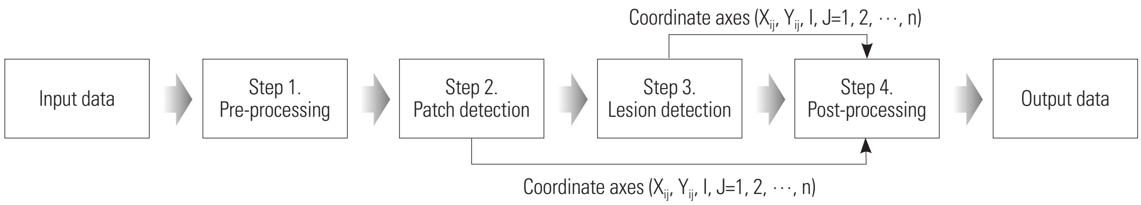

Fig. 1. Overview of the independent task learning process in the end-to-end workflow.

https://doi.org/10.3349/ymj.2021.62.12.1125

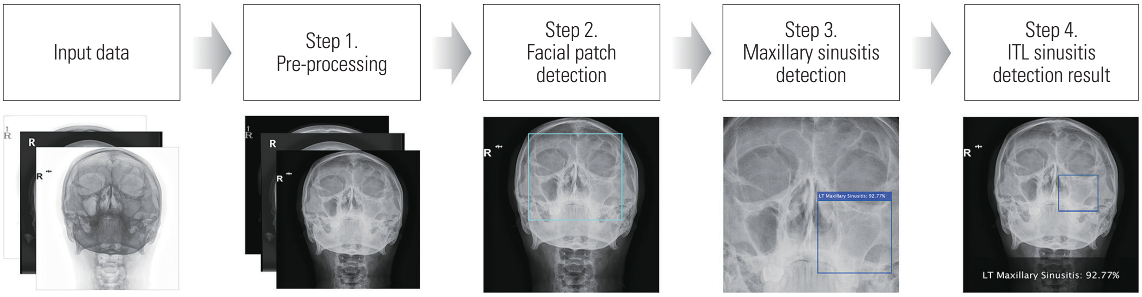

Fig. 2. Overview of the independent task learning (ITL) steps of the sinusitis detection process.

Table 1. Characteristics of the Training and Internal and External Validation Datasets

Characteristic	Training set	Internal validation set	External validation set		
			Dataset #1	Dataset #2	Dataset #3
No. of images	1824	298	300	200	200
Age (yr) (range)	34.84±24.94 (1–105)	34.21±25.45 (1–90)	35.18±26.69 (1–90)	46.04±25.14 (1–90)	18.46±24.64 (0–89)
Sex					
Female	944	140	164	102	78
Male	880	158	136	98	122
Label					
Normal	1227	149	193	104	82
Sinusitis	597	149	107	96	118
Right	188	51	33	33	32
Left	177	36	33	24	25
Bilateral	232	62	41	39	61
Manufacturer					
Philips medical systems	1027	208	300	200	
SIEMENS	657	49			
AGFA	124	41			
LISTEM					124
DongKang					54
INFINITT	16				
Samsung Electronics					15
GE Healthcare					7

Data are presented as mean±standard deviation or n.

pre-processing step for ITL (step 1). For the patch detection step of the ITL, facial patch detection (step 2) based on deep learning is performed. After the facial patch detection step, a facial patch selected by the facial patch detector is cropped to remove the background. For the lesion detection step, maxillary sinusitis detection (step 3) based on deep learning is performed. Before maxillary sinusitis detection, image intensity normalization with a minimum to maximum of 0–1 is applied. Finally, the result of the maxillary sinusitis detection is displayed as a report in the form of boundary boxes on the original image space (step 4).

Data collection

Table 1 shows the characteristics of internal and external validation sets. With the approval of the Institutional Review Board of Kyung Hee University Hospital (IRB number : KHU-IRB 2019-10-010), our retrospective study collected X-ray imaging data from anonymized patients. A total of 2122 Water's view images, consisting of 746 sinusitis (35.16%) and 1376 normal images (64.84%), were collected between July 2016 and September 2019 for the internal dataset. The mean and standard deviation (SD) of age was 34.84 and 24.94 years, with 1084 female individuals, which is 51.08% of the internal dataset. For external validation, 700 Waters' view images (300, 200, and 200) were collected between July 2019 and September 2019 from three different institutions (#1, Kyung Hee University Hospital at Gangdong; #2, Korea Cancer Center; and #3, Severance Hospital, respectively) as external validation sets #1, #2, and #3. The number of images for sinusitis and normal were 107 (35.67%) and 193 (64.33%), 96 (48.00%) and 104 (52.00%), 118 (59.00%) and 82 (41.00%) for external validation sets #1, #2, and #3, respectively. The mean±SD age was 35.18±26.69, 46.04±25.14, and 18.46±24.64, with 164 (54.67%), 102 (51%), and 78 (39%) female individuals in external validation sets #1, #2, and #3, respectively.

Subject labeling and dataset

To train an automatic facial patch detector, the internal dataset was randomly divided into a training set (1227 and 597 images labeled as normal and sinusitis, respectively) and internal validation set (149 and 149 images labeled as normal and sinusitis, respectively). To train a maxillary sinusitis detector, 597 images were labeled as sinusitis from the training set of images used in training the facial patch detector. For the training and internal validation sets and external validation sets #1, #2 and #3, all images were labeled twice (once for facial patch detection and once for maxillary sinusitis detection) by two radiologists. After the primary screening of clinical records from the picture archiving and communication system (PACS; Infinitt Gx PACS, INFINITT Healthcare), a board-certified radiologist with 10 years of experience rechecked and labeled all images in the internal set and three external validation sets. To train and evaluate the performance of the facial patch detector and maxillary sinusitis detector, every image was labeled with two different labels, that is, facial patch labels (Fig. 3A) and maxillary sinusitis labels (Fig. 3B). The facial patch label, which included the maxillary sinus, maxilla, nasal bone, and orbit, was selected as close to the square as possible to avoid distortion of the maxillary sinus shape when cropping and resizing the image. Maxillary sinusitis labeling was performed in the left and right maxillary sinusitis, and a bilateral sinusitis label was assigned when the left and right maxillary sinuses both had sinusitis.

Furthermore, we trained the network to classify normal and sinusitis individuals using localized information. Every sinusitis label was further checked for four subclasses of sinusitis forms to determine bias in maxillary sinusitis detection: "air/fluid," "full opacification," "cyst," and "mucosal thickening."

Training a detection network

The facial detection network was trained using a training dataset. The you only look once version 2 (YOLO v2) object detection network,[25] which is a state-of-the-art deep learning technique based on pre-trained ResNet-50,[26] was trained with the following parameters: number of anchor boxes=7, optimizer="Adam," size of mini-batch=256, initial learning rate=1e-3, factor for L2 regularization=1e-4, and max epochs=200. Similar to the facial patch detector, the YOLO v2 object detection network was trained for maxillary sinusitis detection with a subset of the training set, which was labeled as sinusitis. Bayesian hyperparameter optimization[27,28] was applied to enhance the performance of the detector with the following parameters and ranges: input image size for training (isotropic)=224–512, number of anchor boxes=13–20, max epoch=30–100, initial learning rate=1e-6 to 1e-3, factor for L2 regularization=1e-6 to 1e-3, and maximum time=12 h. The objective function used to evaluate the trained model during hyperparameter optimization was set as [1–accuracy], where the trained model evaluated the internal validation set as normal or sinusitis. The other training options were as follows: random horizontal reflection=true, factor for image scale=0.9–1.1, optimizer="Adam," and size of the mini-batch=64. Bayesian hyperpa-

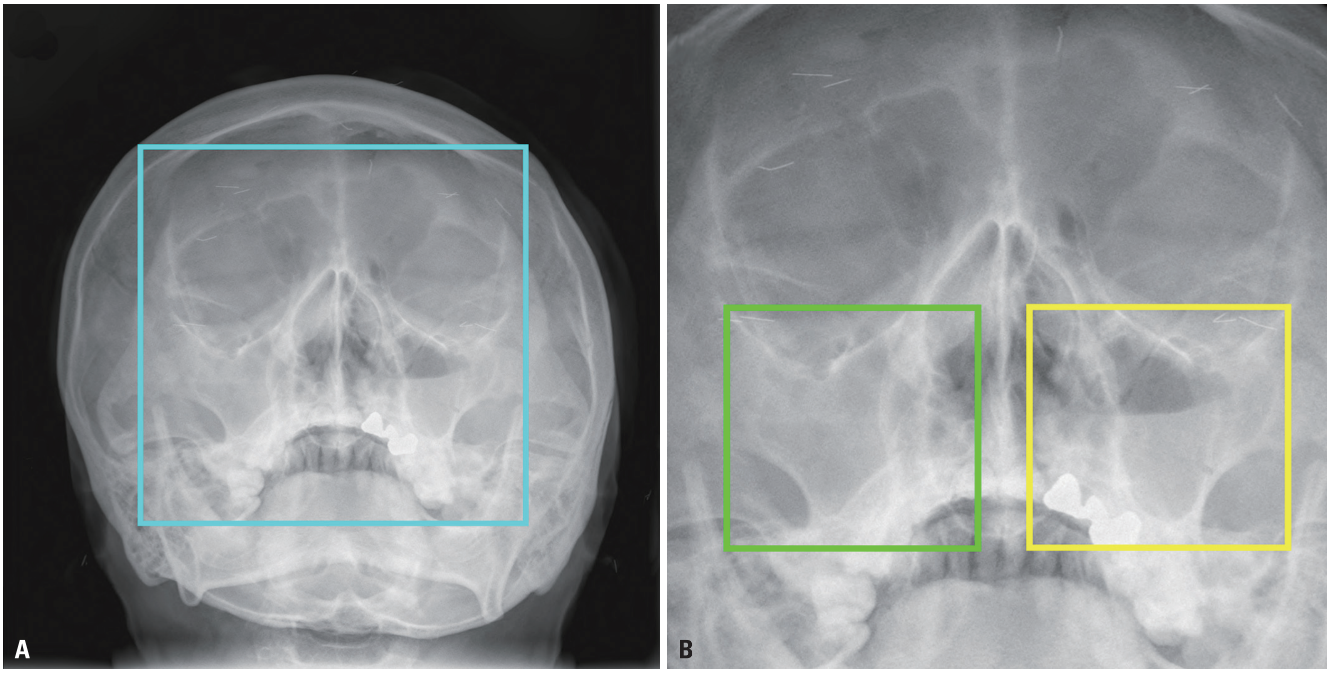

Fig. 3. Facial patch and maxillary sinusitis annotations. (A) Facial patch label is drawn as a cyan rectangle. (B) Maxillary sinusitis labels are drawn as green and yellow rectangles for the right and left maxillary sinusitis, respectively.

rameter optimization was performed for 37573 s with 30 iterations, and the performance of the best model was selected with the following parameters: the final value for the objective function=0.0738, input image size for training (isotropic)=270, number of anchor boxes=16, max epoch=30, initial learning rate=3.646×1e-4, and factor for L2 regularization=5.514×1e-5. To implement the flipping augmentation option that was randomly flipped left to right, we disregarded the left and right maxillary sinusitis information in the training of the network.

All processes were performed on a single-server computer on a Linux operating system (Ubuntu 18.04) with a single NVIDIA Quadro RTX 8000 with a 48-GB memory GPU. Image labeling, image processing, and training networks were performed using MATLAB (MathWorks, R2019b, Natrick, MA, USA).

Performance evaluation

First, the detection performances of the facial patch detector and maxillary sinusitis detector were evaluated using the mean intersection over union (mIoU). IoU, also known as the Jaccard index, is commonly used for evaluating the performance of object detection,[29] and it is defined as

$$\text{IoU}=\frac{\text{Area of overlap}}{\text{Area of union}} \qquad (1)$$

We calculated the mIoU using the average IoU of each detection box using MATLAB (MathWorks, R2019b).

Second, the performance of the facial patch detector was evaluated using accuracy. All DICOM images of Waters' view had a facial region that included the left and right maxillary sinuses, so we only evaluated accuracy because there were no negative samples. To calculate accuracy, a predicted patch was estimated as true positive if the predicted patch contained both left and right maxillary sinus.

Third, the performance of the maxillary sinusitis detector was evaluated using facial patch-cropped images after processing the facial patch detector. To estimate performance, a predicted box including the whole region of sinusitis was evaluated as true positive. All images were divided into left and right sides based on the bilateral maxillary sinuses and evaluated individually. The outcome of the maxillary sinusitis detection box does not have information about the left and right maxillary sinusitis; therefore, we assumed that the facial patch image had proper bilateral symmetry. Furthermore, the sinusitis boundary box was supposed to include the left and right sinusitis, but due to the center of gravity in the sinusitis boundary box, it was shifted to the left and right relative to the center of the facial patch image.

Fourth, to verify the capability of the maxillary sinusitis detector, subclasses labeled as air/fluid, full opacification, cyst, and mucosal thickening, representing the forms of sinusitis, were analyzed.

Finally, the performance of the ITL application for maxillary sinusitis was evaluated with outcome reports, which consisted of a normal sinus, right sinusitis, left sinusitis, and bilateral sinusitis. The outcome report of normal and bilateral sinusitis is supposed to include any and all of the left and right sinusitis boxes from the maxillary sinusitis detector, respectively.

Accuracy, area under the receiver operating characteristic curve (AUC), standard error, 95% confidence interval (CI), sensitivity, specificity, and results of statistical significance tests for AUC (i.e., significant differences in AUC based on 0.5) were measured to assess the maxillary sinusitis detector and ITL application for maxillary sinusitis. All statistical analyses were performed using MedCalc software (www.medcalc.org, Ostend, Belgium).

RESULTS

Performance evaluation for application to maxillary sinusitis

The mIoUs of facial patch and maxillary sinusitis detection were 0.86±0.06 and 0.73±0.12, 0.86±0.6 and 0.71±0.12, 0.79±0.10 and 0.68±0.14, and 0.82±0.07 and 0.69±0.11 in the internal set and external validation sets #1, #2, and #3, respectively. Table 2 shows the performances in facial patch detection, maxillary sinusitis detection, and ITL application for maxillary sinusitis in the internal set and external validation sets #1, #2, and #3. The accuracy of facial patch detection was 100%, 100%, 99.5%, and 97.5% for the internal set and external validation sets #1, #2, and #3, respectively. The accuracy (and AUC) in maxillary sinusitis detection was 88.93% (0.89), 91.67% (0.90), 90.45% (0.86), and 85.13% (0.85) for the internal set and external validation sets #1, #2, and #3, respectively. For all accuracy and AUC values, the performance in left maxillary sinusitis detection was slightly higher than that in right maxillary sinusitis detection, although the difference was not statistically significant. The sensitivity (and specificity) of maxillary sinusitis detection was 89.10 (88.83), 86.52 (93.25), 75.44 (96.48), and 85.37 (84.96) for the internal set and external validation sets #1, #2, and #3, respectively.

The accuracy (and AUC) of the ITL application for maxillary sinusitis was 79.87% (0.80), 84.67% (0.82), 83.92% (0.82), and 73.85% (0.74) for the internal set and external validation sets #1, #2, and #3, respectively. The sensitivity (and specificity) of the ITL application for maxillary sinusitis was 68.46 (91.28), 75.25 (89.45), 69.62 (93.33), and 74.53 (73.03) for the internal set and external validation sets #1, #2, and #3, respectively. The ITL application for maxillary sinusitis showed the lowest performance in regards to accuracy and AUC in external validation set #3 and in regards to sensitivity in external validation set #2. Figs. 4 and 5 show the outcomes and receiver operating characteristic curves of the ITL application for maxillary sinusitis.

Table 2. Performance Evaluation of the Internal and External Test Datasets for Each Deep Learning Model

	ACC	AUC	SE	95% CI	Sensitivity	Specificity	*p* value
Internal validation set							
Facial patch detection	100						
Sinusitis detection	88.93	0.89	0.013	0.862 to 0.914	89.10	88.83	<0.001
ITL application for maxillary sinusitis	79.87	0.80	0.022	0.749 to 0.843	68.46	91.28	<0.001
External validation set #1							
Facial patch detection	100						
Sinusitis detection	91.67	0.90	0.016	0.872 to 0.922	86.52	93.25	<0.001
ITL application for maxillary sinusitis	84.67	0.82	0.024	0.776 to 0.865	75.25	89.45	<0.001
External validation set #2							
Facial patch detection	99.50						
Sinusitis detection	90.45	0.86	0.021	0.822 to 0.892	75.44	96.48	<0.001
ITL application for maxillary sinusitis	83.92	0.82	0.028	0.754 to 0.866	69.62	93.33	<0.001
External validation set #3							
Facial patch detection	97.50						
Sinusitis detection	85.13	0.85	0.018	0.812 to 0.885	85.37	84.96	<0.001
ITL application for maxillary sinusitis	73.85	0.74	0.032	0.670 to 0.798	74.53	73.03	<0.001

ACC, accuracy; AUC, area under the curve; CI, confidence interval; ITL, independent task learning; SE, standard error.

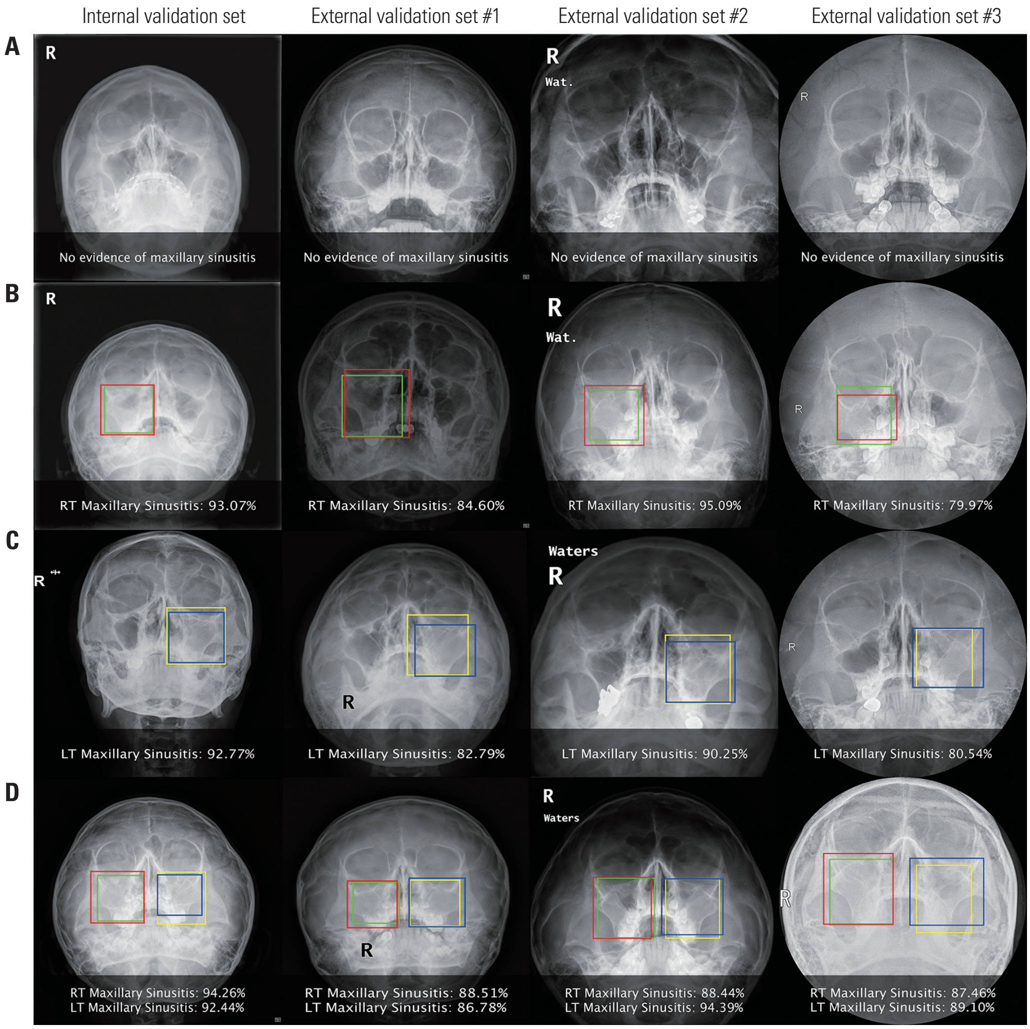
Internal validation set
External validation set #1
External validation set #2
External validation set #3
A
R
R
Wat.
R
No evidence of maxillary sinusitis
No evidence of maxillary sinusitis
No evidence of maxillary sinusitis
No evidence of maxillary sinusitis
B
R
R
Wat.
R
RT Maxillary Sinusitis: 93.07%
RT Maxillary Sinusitis: 84.60%
RT Maxillary Sinusitis: 95.09%
RT Maxillary Sinusitis: 79.97%
C
R
Waters
R
R
R
LT Maxillary Sinusitis: 92.77%
LT Maxillary Sinusitis: 82.79%
LT Maxillary Sinusitis: 90.25%
LT Maxillary Sinusitis: 80.54%
D
R
R
Waters
R
RT Maxillary Sinusitis: 94.26%
LT Maxillary Sinusitis: 92.44%
RT Maxillary Sinusitis: 88.51%
LT Maxillary Sinusitis: 86.78%
RT Maxillary Sinusitis: 88.44%
LT Maxillary Sinusitis: 94.39%
RT Maxillary Sinusitis: 87.46%
LT Maxillary Sinusitis: 89.10%

Fig. 4. Results of the independent task learning (ITL) application for maxillary sinusitis. (A) Normal, (B) right maxillary sinusitis, (C) left maxillary sinusitis, and (D) bilateral maxillary sinusitis for the internal validation set and external validation sets #1, #2, and #3, respectively. On the image, the right and left label information is marked with green and yellow rectangles, respectively; the right and left outcomes of the ITL application for maxillary sinusitis are marked with red and blue rectangles, respectively.

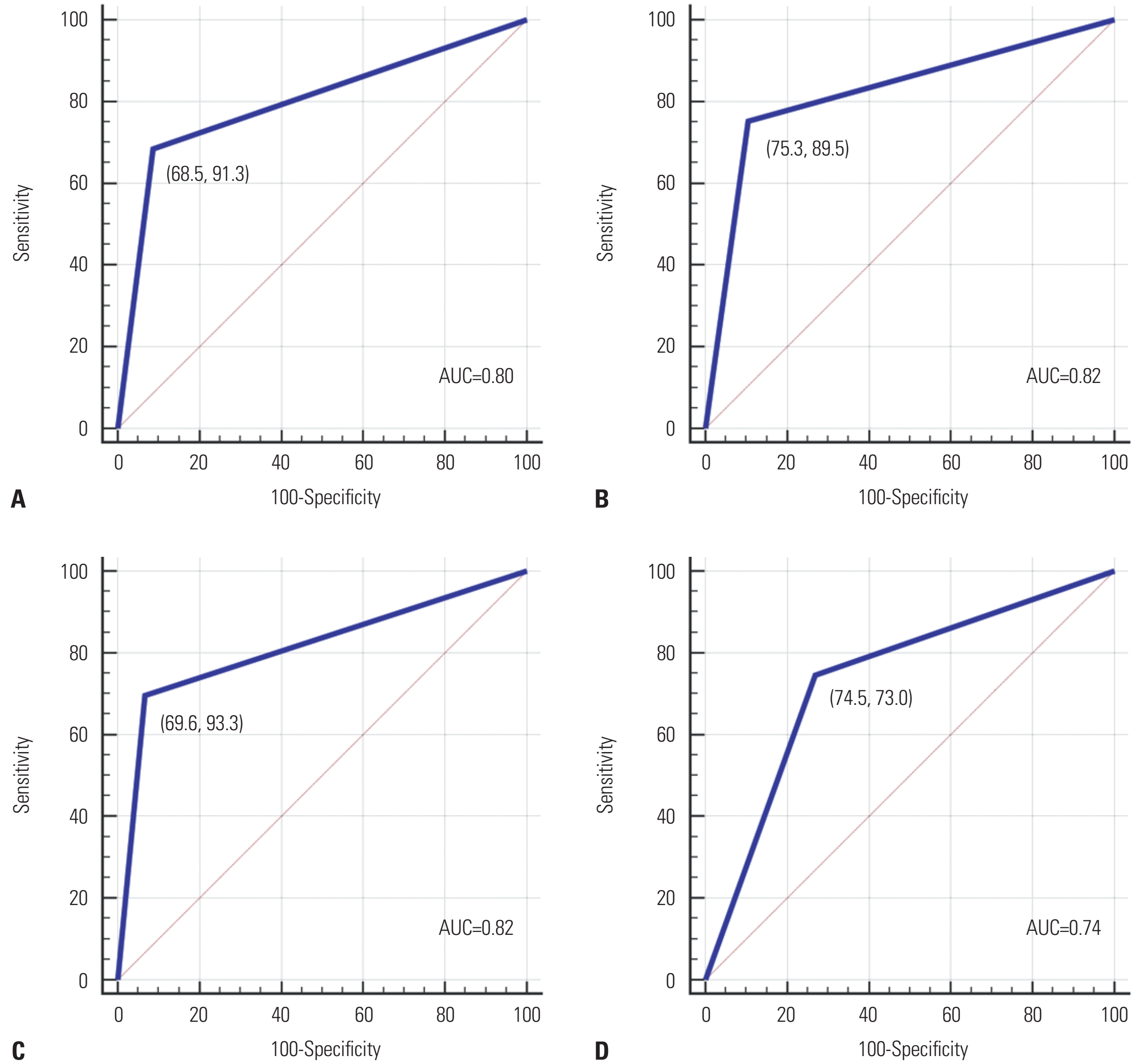

Fig. 5. Receiver operating characteristic curves of the independent task learning application for maxillary sinusitis for (A) the internal validation set and (B-D) external validation sets #1, #2, and #3, respectively. The cut-off point is presented with sensitivity and specificity in parenthesis. AUC, area under the receiver operating characteristic curve.

Interpretation for subclass of maxillary sinusitis and characteristics of trained model

Table 3 summarizes the accuracy and standard error of the maxillary sinusitis detector for diagnosing the subclasses of sinusitis. We investigated the sinusitis characteristics (Table 3) and sample cases of the individual subclasses of sinusitis presented in Fig. 6. The percentages of inaccurate detection for retention cysts were 66.7%, 40%, and 28.6% for the internal set and external validation sets #1 and #3, respectively. We reviewed cases with different results between the radiologists and AI. Although there was one normal PNS radiograph, most cases were misinterpreted: for instance, an underlying bone contour was mistaken as sinusitis. In these cases, radiologists and AI had high error rates. In cystic cases, the density of cysts is weaker than that of the surrounding bones, making them difficult to detect. In young patients, AI misinterpreted a baby tooth as sinusitis.

To understand the differences in the internal and external validation sets, the image characteristics incorrectly detected by the ITL application for maxillary sinusitis were investigated. Second, we investigated the age distribution, which was grouped by decades (Table 4), and found that the youngest group tended to be incorrectly detected by the ITL application for maxillary sinusitis, although no significant difference in age distribution between the validation datasets was found. Third, we investigated the scan parameters, which were kVp, mAs, exposure time, and source-to-detector distance (Table 4), and found no significant differences in scan parameters between the validation datasets.

DISCUSSION

Recent deep learning-based medical studies have been intro-

Table 3. Performance of the Maxillary Sinusitis Detector for Subclasses of Sinusitis

	Full opacification	Air/fluid	Cyst	Mucosal thickening
Temporal dataset				
Number of data	101	28	24	59
ACC (%)	97.0	85.7	66.7	86.4
SE	0.017	0.066	0.096	0.045
External validation set #1				
Number of data	70	21	15	35
ACC (%)	95.7	95.2	40.0	82.9
SE	0.024	0.046	0.126	0.064
External validation set #2				
Number of data	40	15	10	49
ACC (%)	90.0	53.3	80.0	69.4
SE	0.047	0.129	0.126	0.066
External validation set #3				
Number of data	113	7	7	37
ACC (%)	91.2	71.4	28.6	81.1
SE	0.027	0.171	0.171	0.064

ACC, accuracy; SE, standard error.

duced with improved clinical impact, explainability,[30] interpretability,[31] uncertainty,[32] etc. for better clinical applicability. In this study, we introduced an effective learning method applicable for medical images that are difficult to train due to ambiguous lesions.

Previous studies[15-17] have used classification-based methods to train features of PNS. Doing so, however, poses limitations in diagnostic assistance of sinusitis. To use a classification-based model, at least two categories should exist: normal and abnormal. This is because classification-based networks do not learn independent features. Dependent learning, such as classification, trains the relative features among categories. Therefore, a normal case that cannot visualize the lesion is indicated in a heatmap. For these reasons, lesion localization performance is low and inefficient. Since classification-based networks are not intended for target localization, it is difficult to determine the performance of feature recognition from a learned model. Explainability is also essential in diagnostic assistance software for medical imaging, and lesion localization is an efficient approach for explaining a learned model.

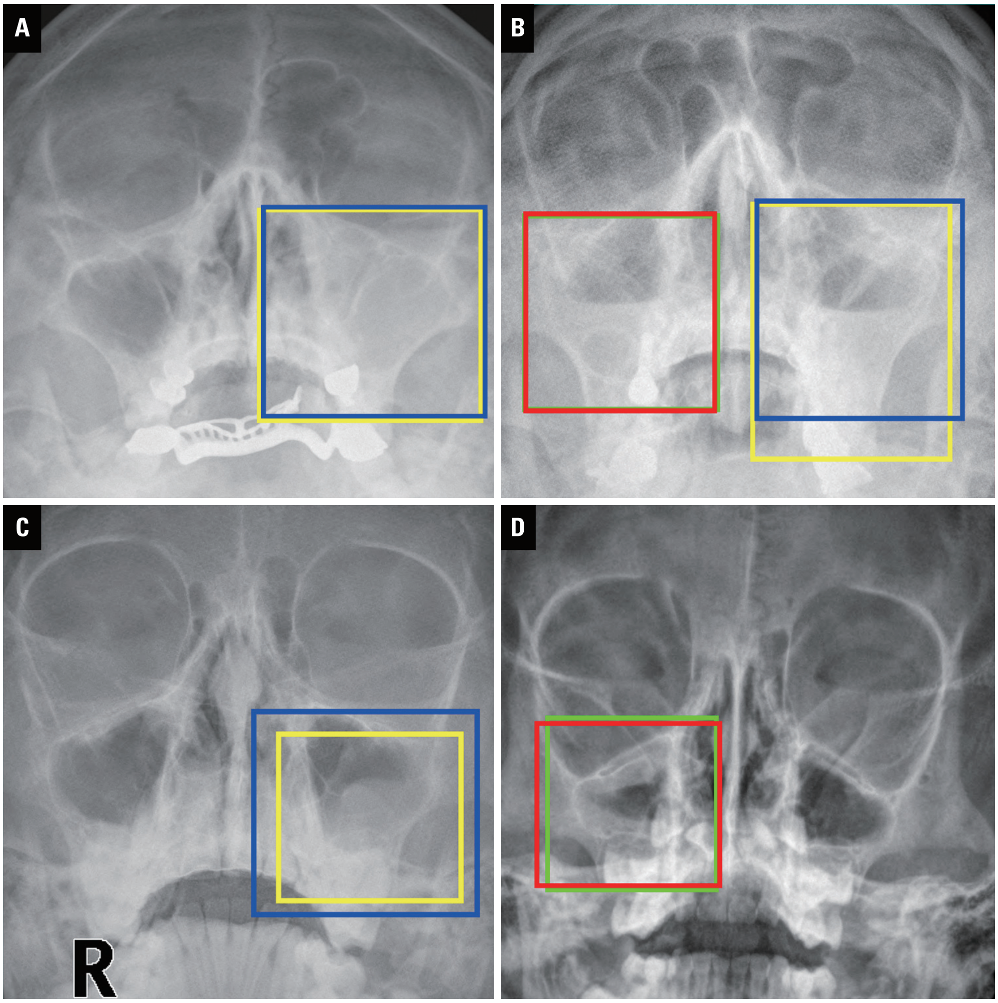

Fig. 6. Sample cases of each subclass of sinusitis. Sinusitis detection was performed after facial patch detection for all sample cases of (A) full opacification, (B) air/fluid, (C) cyst, and (D) mucosal thickening. Right and left label information is marked with green and yellow rectangles; the right and left outcomes of the independent task learning application for maxillary sinusitis are marked with red and blue rectangles, respectively.

Table 4. Specific Information of Incorrectly Detect Subjects

	Internal validation set		External validation set #1		External validation set #2		External validation set #3	
	Normal	Sinusitis	Normal	Sinusitis	Normal	Sinusitis	Normal	Sinusitis
Number of subjects	13	47	21	25	8	24	24	27
Sex (F/M)	7/6	20/27	9/12	12/13	3/5	14/10	10/14	11/16
Age (Incorrected/group total)								
0–10	4/26 (15.4)	18/66 (27.3)	10/70 (14.3)	7/40 (17.5)	5/15 (33.3)	3/21 (14.3)	17/56 (30.4)	21/83 (25.3)
10–20	1/16 (6.3)	2/7 (28.6)	1/5 (20)	2/9 (22.2)	0/2 (0)	0/3 (0)	0/1 (0)	1/3 (33.3)
20–30	2/25 (8)	3/6 (50)	1/13 (7.7)	0/5 (0)	1/11 (9.1)	1/4 (25)	1/6 (16.7)	0/1 (0)
30–40	0/17 (0)	2/9 (22.2)	0/13 (0)	2/6 (33.3)	0/14 (0)	0/1 (0)	0/4 (0)	2/4 (50)
40–50	3/16 (18.8)	8/13 (61.5)	2/17 (11.8)	5/10 (50)	0/12 (0)	4/6 (66.7)	1/4 (25)	1/2 (50)
50–60	0/19 (0)	4/18 (22.2)	2/25 (8)	1/9 (11.1)	2/26 (7.7)	4/14 (28.6)	1/5 (20)	3/8 (37.5)
60–70	2/17 (11.8)	5/20 (25)	4/37 (10.8)	3/12 (25)	0/23 (0)	4/15 (26.7)	5/7 (71.4)	1/4 (25)
70–80	0/9 (0)	3/7 (42.9)	1/17 (5.9)	5/8 (62.5)	1/14 (7.1)	4/11 (36.4)	0/4 (0)	2/5 (40)
80–90	1/4 (25)	2/3 (66.7)	0/2 (0)	0/2 (0)	0/4 (0)	4/4 (100)	0/3 (0)	0/0 (0)
Scan parameters								
kVp	76.30	73.45	70.76	72.52	77.00	77.00	78.19	79.39
mAs	20.50	20.98	20.38	17.92	13.25	19.67	17.62	16.50
Exposure time (ms)	100.00	62.45	49.90	45.64	39.25	57.50	55.33	57.17
SDD (mm)	1500.00	1506.74	1113.65	1126.40	1400.00	1400.00	1036.00	1153.30

SDD, source-to-detector distance.
Data are presented as n (%).

Although there is a heatmap-based localization approach, it lacks sufficient information with which to indicate lesion location because a heatmap from a classification network is not trained pixel-by-pixel ground truth. Several studies have been conducted to localize lesions in medical imaging using segmentation-based network.[33,34] However, ambiguous lesions are difficult to label as ground truth: full opacification and air/fluid in PNSs have clear margins on radiography, whereas cyst or mucosal thickening is difficult to label pixel-by-pixel by a radiologist.

To overcome the limitations of previous studies, we developed an approach with an ITL process that was shown to be reliable in diagnostic assistance of maxillary sinusitis with X-ray images. The proposed model was compared with those in previous studies. In previous studies[15,17] of conventional Waters' view radiographs and another study[16] of panoramic radiographs, a convolutional neural network based on deep learning was utilized, and handcrafted patched maxillary sinus images were classified as sinusitis or normal. These studies showed the relevance of a conventional neural network that shows superior or comparable accuracy, AUC, sensitivity, and specificity to results obtained by radiologists. These classification performances are highly comparable to ours in internal validation. The sensitivity and specificity of our model were 75.44% to 89.10% and from 84.96% to 96.48%, respectively, which are comparable to a previous study.[17] In external validation, our results showed that a gap between sensitivity and specificity 0.41 to 21.04 in external validation sets, which is lower than that in previous work.[17] ITL application for maxillary sinusitis provides not only the classification of normal or sinusitis, but also visual indication for localization of lesions, such as the right or left side. The use of our model minimizes time and manpower consumption for medical staff and provides more accurate diagnosis results to patients.

Although we compared performance with previous research, we utilized datasets for internal and external validation that were labeled by humans using conventional Waters' view radiography without verifying paranasal computed tomography, which is essential for ambiguous data, such as cystic or mucosal thickening subclasses of sinusitis, and this is a limitation of this study. It is unusual to obtain conventional radiography and paranasal computed tomography[17] simultaneously. According to previous studies[35,36] that evaluated the performance of radiologists using X-ray to detect maxillary sinusitis, the sensitivity was reported as 70% and 80%, and the specificity was reported as 100% and 92%. Another limitation is that the maxillary sinusitis detector was trained without normal maxillary sinus information. The purpose of this study was to efficiently detect maxillary sinusitis from the input data, so our study was designed to focus on sinusitis, not the normal maxillary sinus. To train a normal maxillary sinus label, an additional normal sinus label box is required, and the results of the detection method can overlap with one object. For example, for sinusitis and normal maxillary sinus label boxes on the cystic sinusitis (a well-trained network will detect small cystic lesions on the maxillary sinus), results of a normal maxillary sinus or sinusitis can be confused. In addition, an infection on other regions, such as the frontal, ethmoid, and sphenoid sinuses, that can diagnosed in conventional X-ray images was not targeted in this study, and this is another limitation of this study.

In conclusion, for an AI assistant software to be applied in clinical practice, an ITL approach must be applied to each step, and a system that includes the entire process of diagnosis of sinusitis and different diseases should be designed.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Government of South Korea (Ministry of Science and ICT, MSIT) (NRF-2020R1C1C1006623 and NRF-2021R1F1A1050515).

AUTHOR CONTRIBUTIONS

Conceptualization: Hyug-Gi Kim and Kyung Mi Lee. Data curation: Chang-Woo Ryu, Soonchan Park, Ji Hye Jang, and HyunSeok Choi. Formal analysis: Hyug-Gi Kim. Funding acquisition: Hyug-Gi Kim and Kyung Mi Lee. Investigation: Hyug-Gi Kim and Kyung Mi Lee. Methodology: Hyug-Gi Kim. Project administration: Eui Jong Kim. Resources: Hyug-Gi Kim and Kyung Mi Lee. Software: Jang-Hoon Oh. Supervision: Hyug-Gi Kim, Kyung Mi Lee, and Eui Jong Kim. Validation: Chang-Woo Ryu, Soonchan Park, Ji Hye Jang, and HyunSeok Choi. Visualization: Jang-Hoon Oh. Writing—original draft: Jang-Hoon Oh and Hyug-Gi Kim. Writing—review & editing: Kyung Mi Lee. Approval of final manuscript: all authors.

ORCID iDs

Jang-Hoon Oh	https://orcid.org/0000-0002-4251-5470
Hyug-Gi Kim	https://orcid.org/0000-0002-6786-9531
Kyung Mi Lee	https://orcid.org/0000-0003-3424-0208
Chang-Woo Ryu	https://orcid.org/0000-0002-4674-7295
Soonchan Park	https://orcid.org/0000-0001-6057-7117
Ji Hye Jang	https://orcid.org/0000-0003-3302-4343
Hyun Seok Choi	https://orcid.org/0000-0003-4999-8513
Eui Jong Kim	https://orcid.org/0000-0003-2183-8657

REFERENCES

1. Tsochatzidis L, Costaridou L, Pratikakis I. Deep learning for breast cancer diagnosis from mammograms-A comparative study. J Imaging 2019;5:37.
2. Al-Masni MA, Al-Antari MA, Park JM, Gi G, Kim TY, Rivera P, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Programs Biomed 2018;157:85-94.
3. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep 2019;9:12495.
4. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. Sci Rep 2019;9:6268.
5. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ. Identifying pneumonia in chest X-rays: a deep learning approach. Measurement 2019;145:511-8.
6. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017;19:277-85.
7. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30:449-59.
8. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30:427-41.
9. Das A, Acharya UR, Panda SS, Sabut S. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. Cogn Syst Res 2019;54:165-75.
10. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. Sci Rep 2018;8:15497.
11. Tanzi L, Vezzetti E, Moreno R, Moos S. X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach. Appl Sci 2020;10:1507.
12. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based Software as a Medical Device (SaMD). Silver Spring: U.S. Food and Drug Administration; 2019.
13. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290:218-28.
14. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2019;2:e191095.
15. Kim HG, Lee KM, Kim EJ, Lee JS. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. Quant Imaging Med Surg 2019;9:942-51.
16. Murata M, Ariji Y, Ohashi Y, Kawai T, Fukuda M, Funakoshi T, et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. Oral Radiol 2019;35:301-7.
17. Kim Y, Lee KJ, Sunwoo L, Choi D, Nam CM, Cho J, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest Radiol 2019;54:7-15.
18. Song L, Lin J, Wang ZJ, Wang H. An end-to-end multi-task deep learning framework for skin lesion analysis. IEEE J Biomed Health Inform 2020;24:2912-21.
19. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed 2018;153:1-9.
20. Qummar S, Khan FG, Shah S, Khan A, Shamshirband S, Rehman ZU, et al. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access 2019;7:150530-9.
21. Al-Antari MA, Al-Masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform 2018;117:44-54.
22. Uhm KH, Jung SW, Choi MH, Shin HK, Yoo JI, Oh SW, et al. Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. NPJ Precis Oncol 2021;5:54.
23. Si K, Xue Y, Yu X, Zhu X, Li Q, Gong W, et al. Fully end-to-end deep-learning-based diagnosis of pancreatic tumors. Theranostics 2021;11:1982-90.
24. Levy JJ, Salas LA, Christensen BC, Sriharan A, Vaickus LJ. PathFlowAI: a high-throughput workflow for preprocessing, deep learning and interpretation in digital pathology. Pac Symp Biocomput 2020;25:403-14.
25. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA: IEEE; 2017. p.7263-71.

26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA: IEEE; 2016. p.770-8.
27. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: a review of Bayesian optimization. Proceedings of the IEEE 2015;104:148-75.
28. Zhang Y, Sohn K, Villegas R, Pan G, Lee H. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7-12; Boston, MA, USA: IEEE; 2015. p.249-58.
29. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, CA, USA: IEEE; 2019. p.658-66.
30. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging 2020;6:52.
31. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov 2020;10:e1379.
32. Tanno R, Worrall D, Kaden E, Ghosh A, Grussu F, Bizzi A, et al. Uncertainty quantification in deep learning for safer neuroimage enhancement. arXiv 1907.13418 [Preprint]. 2019 [accessed on 2021 May 31]. Available at: https://arxiv.org/abs/1907.13418.
33. Narayanan BN, Hardie RC. A computationally efficient u-net architecture for lung segmentation in chest radiographs. Proceedings of the 2019 IEEE National Aerospace and Electronics Conference (NAECON); 2019 Jul 15-19; Dayton, OH, USA: IEEE; 2019. p.279-84.
34. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with Deep Neural Networks. Med Image Anal 2017;35:18-31.
35. Burke TF, Guertler AT, Timmons JH. Comparison of sinus x-rays with computed tomography scans in acute sinusitis. Acad Emerg Med 1994;1:235-9.
36. Aaløkken TM, Hagtvedt T, Dalen I, Kolbenstvedt A. Conventional sinus radiography compared with CT in the diagnosis of acute sinusitis. Dentomaxillofac Radiol 2003;32:60-2.