Original Article
Yonsei Med J 2026 Feb;67(2):112-121
https://doi.org/10.3349/ymj.2024.0404
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Endoscopic Diagnosis of Eosinophilic Esophagitis Using a Multi-Task U-Net: A Pilot Study

Ga Hee Kim[1]*, Jooyoung Park[2]*, Seungju Park[3], Jeongeun Hwang[4], Jisup Lim[5], Kanggil Park[2], Sunghwan Ji[6], Kwangbeom Park[7], Jun-young Seo[8], Jin Hee Noh[9], Ji Yong Ahn[10], Jeong-Sik Byeon[10], Do Hoon Kim[10], and Namkug Kim[2,5]

[1]Department of Internal Medicine, Severance Hospital, Yonsei University College of Medicine, Seoul;
[2]Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, Asan Medical Center, Seoul;
[3]CHA University School of Medicine, Seongnam;
[4]Department of Medical IT Engineering, College of Medical Sciences, Soonchunhyang University, Cheonan;
[5]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul;
[6]Division of Geriatrics, Department of Internal Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul;
[7]Department of Internal Medicine, Nowon Eulji Medical Center, Eulji University College of Medicine, Seoul;
[8]Department of Internal Medicine, Bundang Jesaeng General Hospital, Seongnam;
[9]Department of Internal Medicine, Hallym University College of Medicine, Hallym University Sacred Heart Hospital, Anyang;
[10]Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea.

Purpose: Endoscopically identifying eosinophilic esophagitis (EoE) is difficult due to its rare incidence and subtle morphology. We aimed to develop a robust and accurate convolutional neural network (CNN) model for EoE identification and classification in endoscopic images.

Materials and Methods: We collected 548 endoscopic images from 81 patients with EoE and 297 images from 37 normal patients. These datasets were labeled according to the four eosinophilic esophagitis endoscopic reference score (EREFS) features: edema, rings, exudates, and furrows. A multi-task U-Net with an auxiliary classifier on various levels of skip connections (*sca*U-Net) was proposed. Then, *sca*U-Net was compared with VGG19, ResNet50, EfficientNet-B3, and a typical multi-task U-Net CNN. The performances of each model were evaluated quantitatively and qualitatively based on accuracy (ACC), area under the receiver operating characteristics (AUROC), and gradient-weighted class activation map (Grad-CAM), and were also compared with those of 25 human endoscopists.

Results: Our *sca4*U-Net with 4th-level skip connection showed the best performances in ACC (86.9%), AUROC (0.93), and outstanding Grad-CAM results compared to other models, reflecting the importance of utilizing the deepest skip connection. Moreover, the *sca4*U-Net showed generally better performance when compared with endoscopists with various levels of experience.

Conclusion: Our method showed robust performance compared to expert endoscopists and could assist endoscopists of all experience levels in the early detection of EoE–a rare but clinically important condition.

Key Words: Eosinophilic esophagitis, endoscopy, diagnosis

Received: December 10, 2024 **Revised:** July 20, 2025 **Accepted:** August 7, 2025 **Published online:** November 12, 2025
Co-corresponding authors: Do Hoon Kim, MD, PhD, Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
E-mail: dohoon.md@gmail.com and
Namkug Kim, PhD, Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea.
E-mail: namkugkim@gmail.com

*Ga Hee Kim and Jooyoung Park contributed equally to this work.
•The authors have no potential conflicts of interest to disclose.

INTRODUCTION

Eosinophilic esophagitis (EoE) is one of the clinically significant chronic immune-mediated inflammatory diseases diagnosed during endoscopic assessment of dysphagia and food impaction.[1] The incidence and prevalence of EoE continue to increase worldwide, particularly in Asian countries.[2,3] Typical endoscopic findings of EoE include esophageal rings, white plaques or exudates, decreased vascularity or edema, and linear furrows.[4,5] However, these features require additional field training owing to their low incidence, and inconsistency in endoscopists' knowledge often leads to misinterpretations.[6]

Convolutional neural networks (CNNs) have emerged and started to show high performance in image classification tasks. Therefore, CNNs have now been actively studied to establish a computer-aided diagnosis system as a visual biopsy tool in medical imaging. There have also been such attempts in EoE to differentiate positive endoscopic images of EoE from normal images of the esophagus. A trained ResNet-50 model showed performance with an accuracy, sensitivity, and specificity of 94.7%, 90.8%, and 99.0%, respectively.[7] Trained DenseNet model demonstrated performance with an accuracy of 91.5%, sensitivity of 87.1%, and specificity of 93.6%. However, there are several limitations that could be addressed in the prior studies.[8]

1) Incorporating not only the binary classification of EoE but also other indicators of eosinophilic esophagitis endoscopic reference score (EREFS) visible on endoscopic images as prediction targets for the model; these can potentially aid in learning how to inspect endoscopic images.

2) Their gradient-weighted class activation map (Grad-CAM) results are not showing high correlation with fixed rings, plaques, or furrows.

3) The previous studies selected only one kind of CNN model for experimentation.

There have been a few studies to date in which endoscopic images of EoE were assessed using a CNN.[8-10] However, previous studies had limitations due to insufficient data quantity and diversity.

In this study, we proposed a skip connection auxiliary classifier U-Net (*sca4*U-Net) model which was built based on the U-Net model and trained with a multi-task learning method. The proposed model was trained to predict whether the given image contains edema, fixed rings, exudate, furrow, as well as EoE. For the performance comparison, Visual Geometry Group (VGG), ResNet, EfficientNet,[11-15] and auxiliary classifier U-Net (*aux*U-Net) models were adopted and trained in the same way as the *sca*U-Net.

MATERIALS AND METHODS

Patients

Adult patients (age >18 years) diagnosed with EoE via histology [15 eosinophils per high-power field (eos/hpf) in esophageal biopsies] between 2007 and 2021 at the Asan Medical Center were enrolled (Supplementary Fig. 1, only online). We only included cases with biopsies taken from at least 5 cm above the Z-line to exclude reflux esophagitis. Two experienced endoscopists (G.H.K and D.H.K) reviewed these endoscopic images. First, G.H.K labeled all images for EREFS scores, and all data were subsequently reviewed by the more experienced endoscopist (D.H.K.). Esophageal eosinophil counts were determined by a gastrointestinal pathologist using hematoxylin and eosin stains. Additional clinical data (e.g., age, sex, and presenting symptoms) were collected and analyzed. Controls were taken from normal persons who were screened via endoscopy and were confirmed with normal findings of five eos/hpf from esophageal biopsies. In the case of the normal control group, the reason for performing an esophageal biopsy was to diagnose systemic infiltrative diseases in the esophageal mucosa without symptoms of esophageal dysfunction. This study was approved by the Institutional Review Board of the Asan Medical Center (IRB approval no. 2021–1260), and all methods were performed in accordance with its guidelines.

Dataset and preprocessing

All endoscopic images were assessed for their EREFS based on edema, rings, exudates, furrows, and strictures. However, strictures, as an indicator, were excluded from our diagnostic judgment since the analysis of still images is unreliable for strictures. EREFS were retrospectively calculated for EoE patients by a board-certified endoscopist. Then, EREFS were one-hot encoded to build a classification model.

The dataset is composed of 548 endoscopic images from 81 EoE patients and 297 endoscopic images from 37 normal controls. For the image augmentations, 90°, 180°, and 270° rotations were applied to every training set, and Gaussian blur was applied to 30% of training sets to increase the diversity of the dataset. Our dataset was divided into 60% training, 20% validation, and 20% test sets in all experimental conditions without patient overlap.

Deep learning models for EoE classification

The models ResNet and DenseNet were adopted in previous works to establish an EoE visual biopsy tool.[7,8] To verify the effectiveness of the shortcut layer of ResNet and DenseNet, we adopted the VGG model (which is a model composed of CNN layers without any specialized layers). It was also important to demonstrate that CNN hyper-parameters found by AutoML techniques are effective for EoE classification. To discover the effectiveness of an AutoML technique, EfficientNet, which is a CNN model constructed with optimal hyper-parameters found by AutoML, was adopted for this study.

Not only were conventional CNN-based classification models adopted, but U-Net-based classification models were also used. To elaborate further, some auxiliary layers were attached

to U-Net's bottleneck, making U-Net able to perform additional classification tasks. Such approaches are called multi-task learning, and this training method makes certain assumptions: the optimal weights for task 1 (i.e., classification) and task 2 (i.e., segmentation) are approximately equal. As mentioned before, this idea can simply be implemented by attaching an auxiliary classifier called an *aux*U-Net to U-Net's bottleneck.[16,17] Notably, Zhou, et al.[18] used *aux*U-Net to enhance the efficacy of different segmentation and classification tasks by training them jointly, which was possible since *aux*U-Net can perform two tasks at the same time. Not only can *aux*U-Net exert synergy between image segmentation and classification tasks, but it can also exert synergy between image reconstruction and classification tasks as a pair.[19-21] Since this idea was inspired by the autoencoder concept and is easy to implement, these methods were widely used by simply changing the loss for segmentation to the loss for regression.[12,19] Haghighi, et al.[20] extracted four fixed-coordinate patches from chest X-rays to train a network to reconstruct each patch and classify their relative locations simultaneously. By doing this multi-task training, the model could understand the subtle diversity of each patch (i.e., its intensity, shape, and texture). Expecting a similar effect on the EoE dataset, *aux*U-Net was thus adopted for this study.

Since U-Net has such great versatility, various researchers' investigations have shown that U-Net's skip connection provides more useful information than its bottleneck.[22,23] In order to put this idea into practical application, we proposed a skip-connection *aux (sca)* U-Net, which uses an auxiliary classifier on the skip connection to classify EoE patients versus normal controls. Since our U-Net has four skip connections, the naming convention of the *sca*U-Net goes from *sca1*U-Net (whose auxiliary classifier was attached to the shallowest skip connection) to *sca4*U-Net (whose auxiliary classifier was attached to the deepest skip connection). Note that *sca1*U-Net and *sca2*U-Net were excluded in this study, since they required two GPUs for their computations with the large feature maps. Fig. 1 presents the main architectures of the competing models–*sca3*U-Net, *sca4*U-Net, and *aux*U-Net.

Multi-class classification

To achieve this study's goal of training a model to aid endoscopists in inspecting endoscopic images for features including edema, fixed rings, exudate, or furrow in diagnosing EoE, the conventional CNNs, *aux*U-Net, *sca3*U-Net, and *sca4*U-Net were trained. Every experiment was conducted using the same training hyper-parameters (Supplementary Table 1, only online). During this multi-class training, the feature vectors of the input images were extracted and converted to logit functions; the final layer then used sigmoid activation to compute the probability of each class.

Evaluation metrics and statistical comparison

We used a generalized estimation metric to compare models with the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) as parameters. The AU-

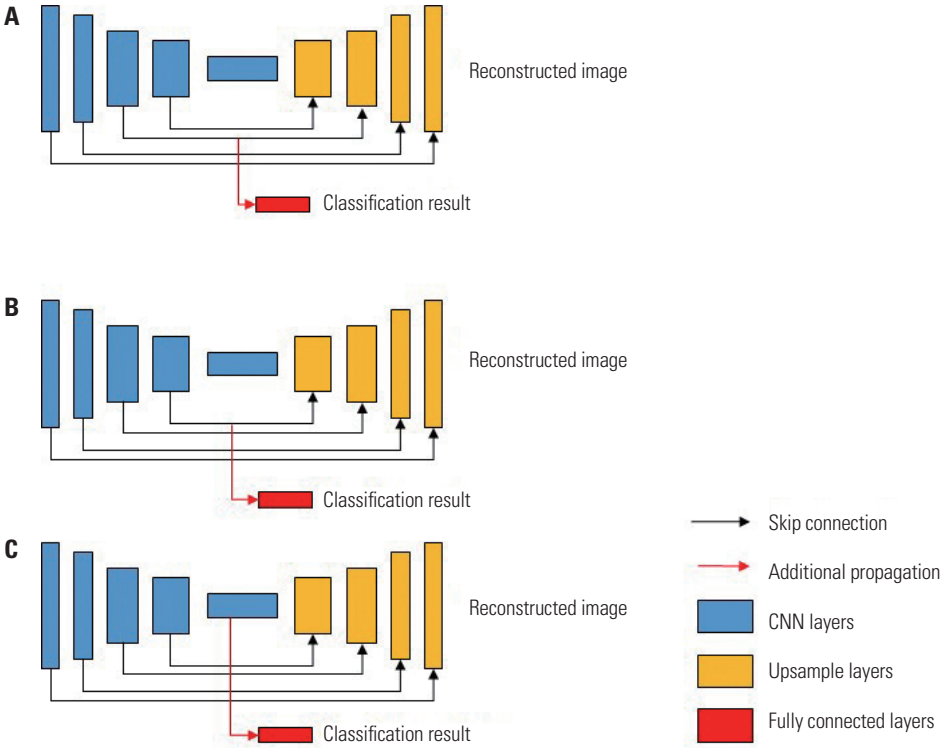

Fig. 1. The architecture of the models. (A) *sca3*U-Net. (B) *sca4*U-Net. (C) *aux*U-Net. CNN, convolutional neural network.

https://doi.org/10.3349/ymj.2024.0404

ROCs were compared using the DeLong test.[24]

For interpretability of the model, we used Grad-CAM[25] overlaid on the original image to visualize the model's performance. Grad-CAM is a tool for visual explanations of CNNs, by activating the mapping that the deep learning network is seeing. Thus, Grad-CAM is a tool for explainable AI.

A comparison of the diagnostic performances between the CNN models and the human experts was conducted. Twenty-five human endoscopists were asked to differentiate between 30 EoE and 40 normal images. The human endoscopists were placed into three groups: faculty members (FAC), fellows ≥1 year (FEL2), and fellows <1 year (FEL1). Groups FAC, FEL2, and FEL1 contained 7, 12, and 6 unique participants, respectively. Comparisons between groups were performed using the Mann–Whitney U Test.

RESULTS

Baseline characteristics of study patients

The clinicopathologic characteristics of the EoE patients are summarized in Table 1. The median age was 44.0 years [interquartile range (IQR): 35.0–56.0 years] and 64 were male. The clinical manifestations of these 81 cases were dysphagia/food impaction (21 cases, 25.9%), heartburn (20, 24.7%), dyspepsia (17, 21.0%), and epigastric pain (16, 19.8%). The median total EREFS score was 3.0 (IQR 2.0–4.0). The normal controls (n=37) were randomly selected from healthy individuals, 23 males and 14 females (median age, 56.0 years; IQR 48.0–63.0 years) with normal endoscopic findings and no esophageal abnormalities.

Table 1. Baseline Characteristics and Endoscopic Findings of the Study Participants

Variable	Value (n=81)
Age, yr	44.0 (35.0–56.0)
Male	64 (79.0)
Symptom profile	
Dysphagia/food impaction	21 (25.9)
Epigastric pain	16 (19.8)
Heartburn	20 (24.7)
Dyspepsia	17 (21.0)
Nausea/vomiting	4 (4.9)
Regurgitation	3 (3.7)
Peak eosinophil count*	36.0 (22.0–58.5)
Endoscopic findings based on EREFS	
Edema	
Grade 0: Absent	15 (18.5)
Grade 1: Present	66 (81.5)
Ring	
Grade 0: None	40 (49.4)
Grade 1: Mild	36 (44.4)
Grade 2: Moderate	5 (6.2)
Grade 3: Severe	0
Exudate	
Grade 0: None	25 (30.9)
Grade 1: Mild	51 (63.0)
Grade 2: Severe	5 (6.2)
Furrow	
Grade 0: None	7 (8.6)
Grade 1: Mild	74 (91.4)
Grade 2: Severe	0
Stricture	
Grade 0: Absent	81 (100)
Grade 1: Present	0
Total EREFS score	3.0 (2.0–4.0)

EREFS, eosinophilic esophagitis endoscopic reference score.
Data are presented as n (%) or median (interquartile range).
*Per high-power field.

Diagnostic ability

This section shows the results of each network's quantitative indicators and their comparisons. In Table 2, the accurate results of each network are presented. The classical CNNs showed marginal EREFS accuracies. However, the *sca4*U-Net showed generally better accuracies except for fixed rings. It also had the best AUROC results in EoE, edema, and exudate (0.93, 0.91, and 0.89, respectively). The *sca3* and *aux*U-Nets were used for ablation studies, showing significantly lower performance, even lower than those of the classical CNNs. Also, the *sca4*U-Net showed significantly different classification results compared with other models according to the DeLong test. Therefore, the last layer of the U-Net is the best location for the auxiliary classifier. In Fig. 2, the AUROC curves of each class from each network are plotted. The results show that the *sca4*U-Net is generally a better model for EoE screening.

Additionally, we conducted a 10-fold cross-validation using the original training data. Specifically, the originally designated validation set was retained as fold 0, while the remaining training data were partitioned into nine additional folds. The resulting classification accuracies are summarized in the Supplementary Table 2 (only online). As demonstrated, the model exhibited stable and reliable performance across folds, with an average EoE classification accuracy of 0.90 and a standard deviation of 0.03, thereby supporting the robustness and generalizability of the proposed approach.

Grad-CAM evaluation

Through Grad-CAM, one can visualize what a deep learning algorithm sees using gradients at the image level. Fig. 3 shows the randomly selected results of 10 positive and 10 negative EoE images from each network. The *sca4*U-Net Grad-CAM results are more clinically relevant than those of the other networks and capture the detailed disease structure of each image. Overall, the Grad-CAM results showed that the *sca4*U-Net successfully captured the pathologic regions, whereas the other networks failed.

Table 2. Model Performance Comparison with 95% CI

	VGG19	ResNet50	EfficientNet-B3	*aux*U-Net	*sca3*U-Net	*sca4*U-Net
Edema						
ACC	0.72 (0.67–0.76)	0.62 (0.56–0.67)	0.76 (0.72–0.81)	0.68 (0.63–0.73)	0.76 (0.71–0.80)	0.80 (0.75–0.84)
SEN	0.45 (0.37–0.53)	0.14 (0.09–0.20)	0.36 (0.41–0.57)	0.38 (0.30–0.45)	0.52 (0.44–0.60)	0.61 (0.53–0.69)
SPE	0.94 (0.90–0.97)	1.00 (0.98–1.00)	0.88 (0.95–0.99)	0.93 (0.89–0.96)	0.95 (0.91–0.98)	0.95 (0.90–0.97)
AUROC	0.90 (0.87–0.94)	0.73 (0.68–0.78)*	0.86 (0.83–0.90)*	0.81 (0.77–0.85)***	0.88 (0.84–0.91)*	0.91 (0.89–0.94)
Fixed rings						
ACC	0.71 (0.66–0.76)	0.73 (0.68–0.78)	0.74 (0.69–0.78)	0.70 (0.65–0.75)	0.73 (0.68–0.77)	0.69 (0.64–0.74)
SEN	0.25 (0.17–0.35)	0.14 (0.08–0.23)	0.36 (0.27–0.47)	0.03 (0.01–0.09)	0.15 (0.09–0.24)	0.21 (0.14–0.31)
SPE	0.89 (0.84–0.92)	0.96 (0.93–0.98)	0.88 (0.83–0.91)	0.96 (0.93–0.98)	0.95 (0.91–0.97)	0.88 (0.83–0.91)
AUROC	0.79 (0.75–0.83)	0.66 (0.61–0.71)*	0.73 (0.68–0.77)	0.70 (0.65–0.75)**	0.74 (0.69–0.78)	0.76 (0.72–0.81)
Exudate						
ACC	0.74 (0.70–0.79)	0.67 (0.62–0.72)	0.68 (0.62–0.72)	0.63 (0.58–0.68)	0.62 (0.56–0.67)	0.77 (0.73–0.81)
SEN	0.45 (0.37–0.53)	0.24 (0.17–0.31)	0.27 (0.20–0.35)	0.13 (0.08–0.19)	0.14 (0.09–0.21)	0.55 (0.46–0.63)
SPE	0.97 (0.93–0.99)	0.99 (0.97–1.00)	0.97 (0.94–0.99)	1.00 (0.98–1.00)	0.96 (0.93–0.98)	0.94 (0.90–0.97)
AUROC	0.88 (0.85–0.91)	0.86 (0.83–0.90)*	0.78 (0.74–0.83)	0.83 (0.79–0.87)**	0.89 (0.86–0.93)	0.89 (0.86–0.93)
Furrows						
ACC	0.71 (0.66–0.75)	0.70 (0.65–0.75)	0.77 (0.72–0.81)	0.67 (0.62–0.72)	0.69 (0.64–0.74)	0.80 (0.75–0.84)
SEN	0.45 (0.38–0.53)	0.38 (0.30–0.45)	0.57 (0.49–0.65)	0.36 (0.29–0.44)	0.41 (0.33–0.48)	0.64 (0.56–0.71)
SPE	0.94 (0.89–0.97)	0.99 (0.97–1.00)	0.95 (0.91–0.97)	0.94 (0.89–0.97)	0.95 (0.91–0.97)	0.94 (0.89–0.97)
AUROC	0.87 (0.84–0.91)*	0.91 (0.88–0.94)	0.82 (0.78–0.86)**	0.83 (0.79–0.86)***	0.87 (0.83–0.90)*	0.90 (0.87–0.93)
EoE						
ACC	0.76 (0.71–0.80)	0.66 (0.61–0.71)	0.81 (0.76–0.85)	0.74 (0.69–0.78)	0.78 (0.74–0.82)	0.87 (0.83–0.90)
SEN	0.53 (0.45–0.60)	0.32 (0.25–0.40)	0.62 (0.54–0.69)	0.53 (0.45–0.60)	0.57 (0.49–0.64)	0.76 (0.69–0.82)
SPE	0.99 (0.96–1.00)	0.99 (0.97–1.00)	1.00 (0.98–1.00)	0.95 (0.91–0.98)	1.00 (0.98–1.00)	0.98 (0.95–1.00)
AUROC	0.91 (0.88–0.94)*	0.90 (0.87–0.93)*	0.86 (0.82–0.89)**	0.87 (0.83–0.91)***	0.91 (0.89–0.94)	0.94 (0.91–0.96)

CI, confidence interval; EoE, eosinophilic esophagitis; ACC, accuracy; SEN, sensitivity; SPE, specificity; AUROC, area under the receiver operating characteristic curve; CNN, convolutional neural network.
Values are presented as mean (95% CI). Metrics include ACC, SEN, SPE, and AUROC for each model in detecting edema across different CNN architectures. Highest metric score; *p<0.05, **p<0.001, ***p<0.0001 by DeLong test comparing *sca4*U-Net with other models.

Comparison of performances between endoscopists and *sca4*U-Net

Table 3 shows that the accuracy (0.94) and NPV (0.93) of the *sca4*U-Net were greater than those of the FAC, FEL2, and FEL1 groups (accuracy: 0.92, 0.89, 0.85; NPV: 0.9, 0.88, 0.63, respectively).The specificity (0.95) and PPV (0.93) of the *sca4*U-Net were greater than those of FEL2 (specificity: 0.94; PPV: 0.91) but lesser than those in FAC (specificity: 0.98; PPV: 0.97). The AUROC curve of the *sca4*U-Net and the performance of the three groups are described in Fig. 4. Moreover, we additionally analyzed performance by symptom severity (total EREFS score >3 vs. total EREFS score ≤3). There was no significant difference between the main categories in the average prediction probability values according to the Mann-Whitney U Test (Supplementary Material and Supplementary Table 3, only online).

DISCUSSION

Recently, CNN-based image classification networks have been used to diagnose gastrointestinal malignancies from endoscopic images, demonstrating the rapid prediction capability of computer-vision methods prior to receiving pathological results. Notably, such technology-enabled capabilities provide great assistance to endoscopists of all skill levels. Therefore, we developed and validated our novel EoE screening tool using *sca4*U-Net trained with a small dataset of endoscopic images, assessed after augmentation with the EREFS scoring system.

Development and evolution of endoscopic imaging, particularly imaged-enhanced endoscopy, show great potential in diagnosing gastrointestinal lesions. However, endoscopists often have difficulty detecting and diagnosing EoE due to its rarity. For these reasons, the disease may be delayed in diagnosis or misdiagnosed. Therefore, early diagnosis is crucial to follow-up care and treatment. Our method improves overall EoE diagnosis with deep machine learning.

https://doi.org/10.3349/ymj.2024.0404

Our proposed *sca4*U-Net reliably outperformed the classical CNNs[11,13,14] in identifying EoE positive vs. EoE negative (accuracy: 86.9%; AUROC: 0.93). Additionally, we determined that the deepest skip connection of the U-Net[26] offers the best feature extraction. The classical CNNs sometimes show good (but mixed) performance, but the Grad-CAM[25] showed that

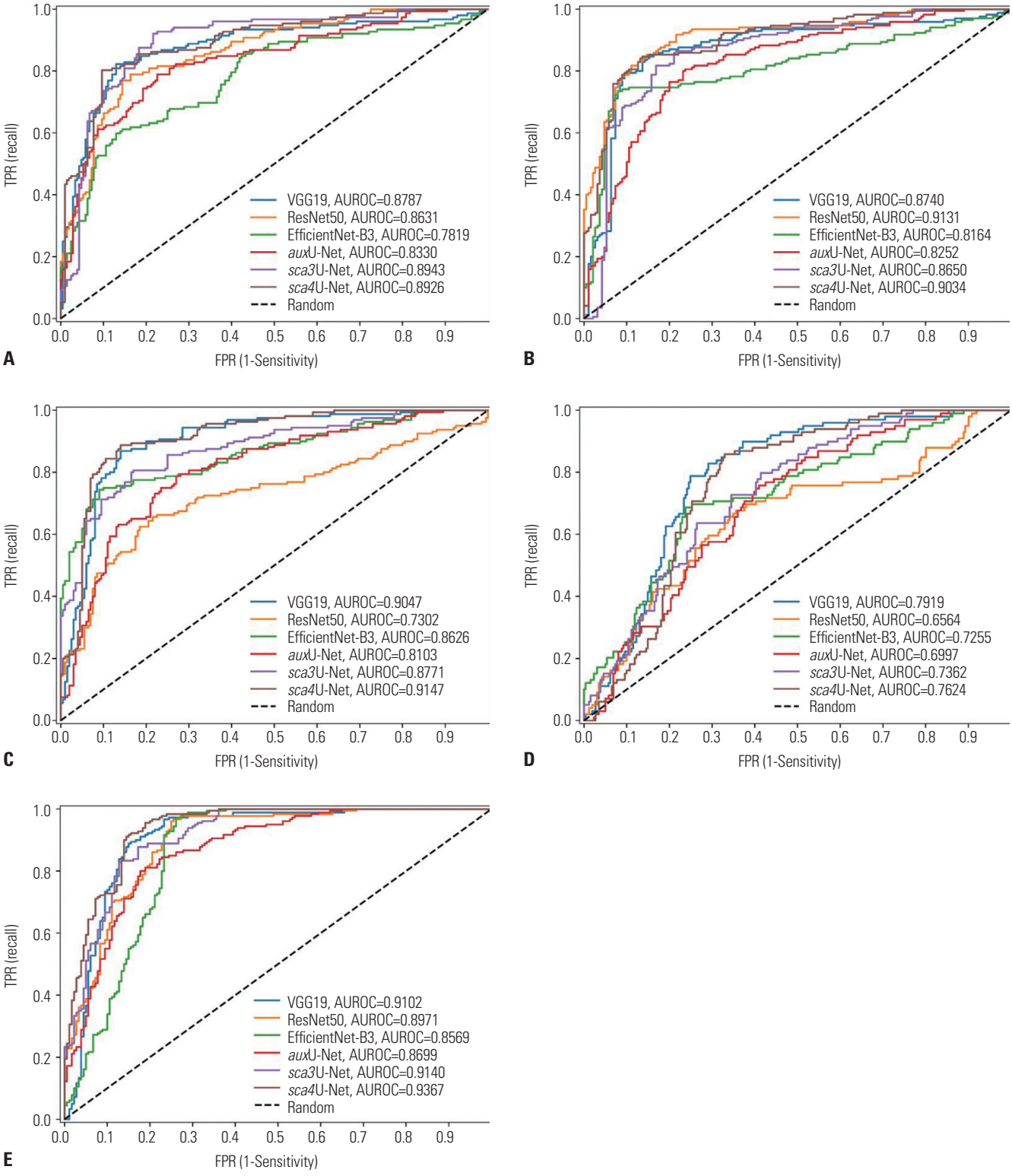

Fig. 2. AUROC comparisons by class. Blue denotes VGG19; orange denotes ResNet-50; green denotes EfficientNet-B3. Red, purple, brown denote auxU-Net, *sca3*U-Net, and *sca4*U-Net, respectively. (A) Edema. (B) Fixed rings. (C) Exudates. (D) Furrows. (E) EoE. AUROC, area under the receiver operating characteristic curve; TRP, true positive rate; FPR, false positive rate.

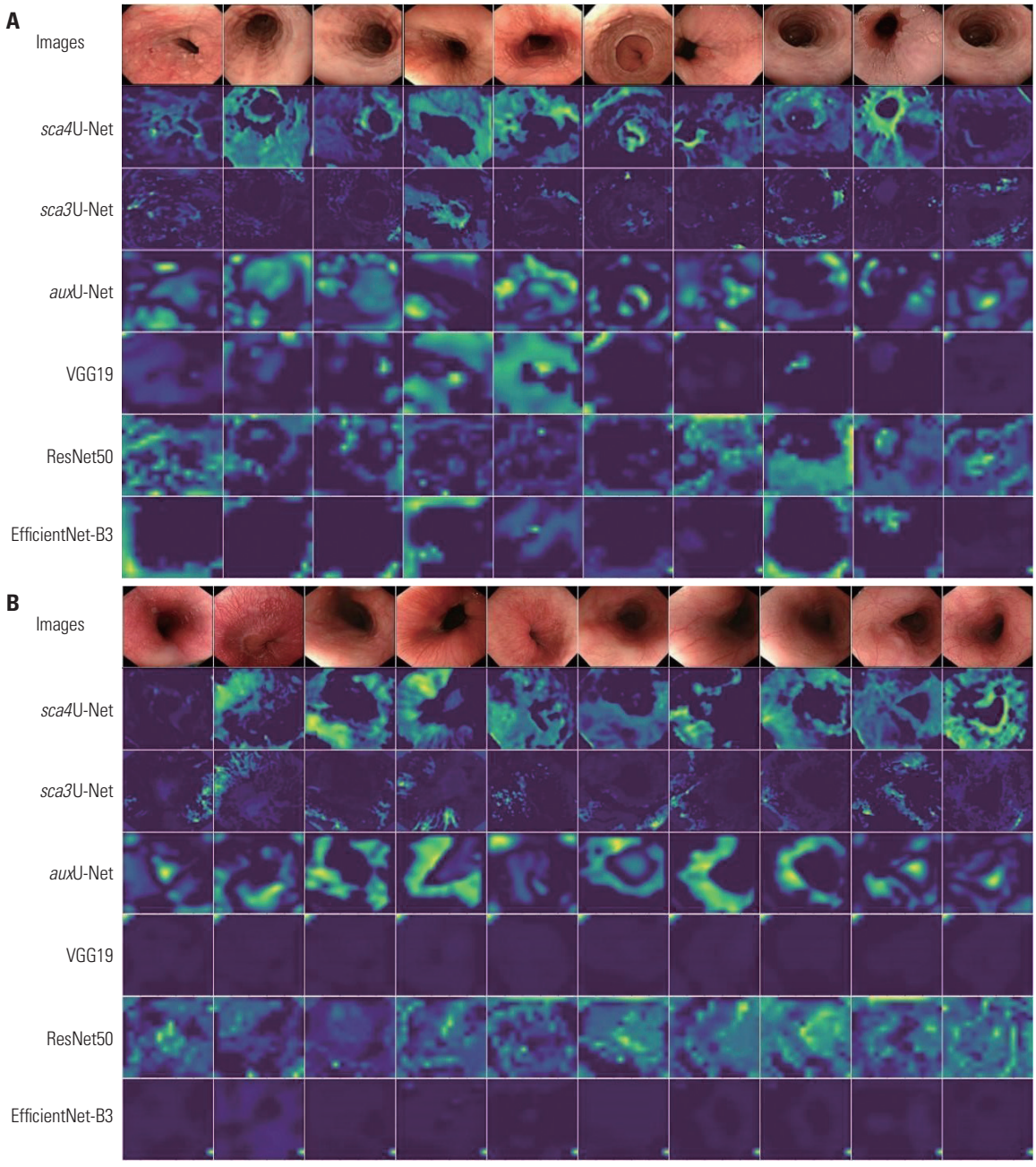

Fig. 3. Grad-CAM results for endoscopic images. The highlighted regions provide a visual explanation of the model's decision. (A) Grad-CAM results of EoE-positive images. (B) The results of EoE-negative images. EoE, eosinophilic esophagitis.

they are inappropriate from the viewpoint of clinical usage. In Fig. 3, clinically, Grad-CAM visualizations can help clinicians increase the accuracy of their diagnosis of EoE by visually showing the characteristic findings of EoE, such as mucosal furrows and circular rings in brightly colored areas, and the direction of the lumen. Table 2 shows the accuracy results of *sca4*U-Net in detecting the EoE features using the EREFS scoring system. We also compared the diagnostic performances of human endoscopists and *sca4*U-Net, finding that the sensitivity and NPV of *sca4*U-Net were significantly greater than all three groups of endoscopists, and the specificity and PPV of *sca4*U-Net were between the FAC and FEL2 groups. Moreover, the predicted probabilities from our model (*sca*4U-Net) between groups with low and high EREFS scores showed a significant difference, which may imply (but not strictly) that our model can detect EoE regardless of its severity (Supplementary Material, only online).

In the study, we divided the endoscopists into FAC, FEL2, and FEL1 groups based on their experience levels. FEL1 comprised medical fellows who lacked experience in endoscopic diagnosis of EoE; FEL2 consisted of fellows with some experience (fellows ≥1 year), whereas those in FAC had abundant

https://doi.org/10.3349/ymj.2024.0404

Table 3. Comparison of Diagnostic Performance between Three Groups of Endoscopists and *sca4*U-Net on 70 Images from the Test Dataset with 95% CIs (40 Images from Normal Controls and 30 Images from EoE Patients)

	ACC	SEN	SPE	PPV	NPV
*sca4*U-Net	0.94 (0.86–0.98)	0.98 (0.87–0.99)	0.93 (0.78–0.98)	0.95 (0.83–0.98)	0.93 (0.82–0.99)
		ACC		PPV	NPV
Group FAC		0.92 (0.88–0.93)		0.97 (0.92–0.98)	0.90 (0.84–0.91)
Group FEL2		0.89		0.92	0.88
Group FEL1		0.85		0.88	0.86

CI, confidence interval; EoE, eosinophilic esophagitis; ACC, accuracy; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; FAC, faculty members; FEL2, fellows ≥1 year; FEL1, fellows <1 year.

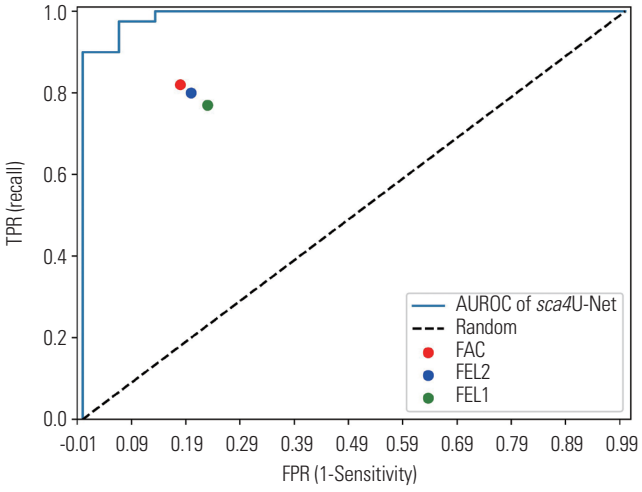

Fig. 4. Comparison of classification performance between physicians and *sca4*U-Net. We sampled 70 images from the test set (eosinophilic esophagitis: 30; normal controls: 40) and asked physicians to classify them. FAC is shown as a red point (sensitivity, 82%; specificity, 98%); FEL2 as a blue point (sensitivity, 80%; specificity, 94%); and FEL1 as a green point (sensitivity, 77%; specificity, 89%). *sca4*U-Net shows an AUROC of 0.991. FAC, faculty members; FEL2, fellows ≥1 year; FEL1, fellows <1 year; AUROC, area under the receiver operating characteristic curve; TRP, true positive rate; FPR, false positive rate.

experience. The sensitivity and NPV of the *sca4*U-Net were significantly greater than those of the other groups, suggesting that *sca4*U-Net can be used by all endoscopist groups as a screening tool to aid in EoE diagnosis. However, it might not be a fair comparison, since endoscopists usually diagnose patients using video, not still images. The strength of this study is that it is the first to validate a novel EoE screening tool using a U-Net CNN in the real world, with high accuracy (86.9%). In addition, there were a relatively large number of normal controls and endoscopic images of patients with EoE compared with those in previous studies. The dataset comprised 548 endoscopic images from 81 patients with EoE and 297 endoscopic images from 37 normal controls. Second, we used endoscopic images from healthy controls, and we selected them for histological confirmation by biopsy (<5 eosinophils per high-power field). EoE is diagnosed histologically based on eosinophilia in a biopsy sample. Previous studies have also confirmed this histologically in normal control groups, which is a great strength of our data. Finally, the dataset used in the present study included relatively diverse endoscopic findings. In our study, we selected and analyzed patients with typical endoscopic findings of EoE (median total EREFS score: 3.0), which may have contributed to the increased value of AI analysis for the diagnosis of EoE.

Our study has several limitations. First, it is a single-center study with a retrospective design based on observational data, which needs to be re-evaluated in a multi-center study. Second, the collected dataset was relatively small, which increases the risk of overfitting. To overcome this problem, we applied both hard and soft augmentations when considering the rarity of EoE. However, the models could not be trained further, since their learning curves showed signs of overfitting after epochs in the mid-40s. Third, because our study collected data retrospectively, we were unable to analyze video footage and only analyzed still photographs. We will study video clips in future studies. For more clinically meaningful applications, this work needs to be applied to video data with a 3D CNN[27,28] and multicenter or multinational studies are required to further expand the validity of this tool. However, it is the first study that conducted multitask experiments with multiple CNN models with robust accuracy. Additionally, in our study, analysis without the EREFS score was not conducted. EoE is an immune-mediated inflammatory disorder; therefore, it is not easy to identify without specific criteria. EoE does not show changes in volume or characteristic mucosal changes from the local area to the lumen of the gastrointestinal tract, such as colon polyps and stomach cancer, which are the main lesions in AI analyses. Please note that the proposed model, *sca4*U-Net, used EREFS information only during training and does not require EREFS information for inference, since it is the target to be predicted by our model based on our study design. EoE exhibits a low prevalence, which makes it difficult to acquire a sufficient dataset and leads to overfitting in our model. For this reason, EREFS information was adopted to enhance information density during training, which can reduce overfitting in the model.

In conclusion, our method showed robust performance compared to expert endoscopists, which could assist endoscopists of all experience levels in the early detection of the infrequent but clinically significant EoE. Multi-center prospective studies with mature datasets are needed for clinical application.

ACKNOWLEDGEMENTS

This study was supported by a KSNM grant of the Korean Society of Neurogastroenterology and Motility for KSNM-22-03 and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI18C2383).

AUTHOR CONTRIBUTIONS

Conceptualization: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Data curation: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Formal analysis: Seungju Park, Jeongeun Hwang, Jisup Lim, Sunghwan Ji, Kanggil Park, Kwangbeom Park, Jun-young Seo, Jin Hee Noh, Ji Yong Ahn, and Jeong-Sik Byeon. Funding acquisition: Ga Hee Kim and Namkug Kim. Investigation: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Methodology: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Project administration: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Resources: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Software: Jooyoung Park and Namkug Kim. Supervision: Do Hoon Kim and Namkug Kim. Validation: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Visualization: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Writing—original draft: Ga Hee Kim, Jooyoung Park, Do Hoon Kim, and Namkug Kim. Writing—review & editing: Ga Hee Kim and Jooyoung Park. Approval of final manuscript: all authors.

ORCID iDs

Ga Hee Kim	https://orcid.org/0000-0002-7652-2580
Jooyoung Park	https://orcid.org/0000-0002-7641-3827
Seungju Park	https://orcid.org/0000-0002-5944-6344
Jeongeun Hwang	https://orcid.org/0000-0002-5821-5405
Jisup Lim	https://orcid.org/0000-0001-8938-233X
Kanggil Park	https://orcid.org/0009-0009-5525-3869
Sunghwan Ji	https://orcid.org/0000-0002-8150-933X
Kwangbeom Park	https://orcid.org/0000-0002-3826-3274
Jun-young Seo	https://orcid.org/0000-0002-4987-0686
Jin Hee Noh	https://orcid.org/0000-0001-6720-9528
Ji Yong Ahn	https://orcid.org/0000-0002-0030-3744
Jeong-Sik Byeon	https://orcid.org/0000-0002-9793-6379
Do Hoon Kim	https://orcid.org/0000-0002-4250-4683
Namkug Kim	https://orcid.org/0000-0002-3438-2217

REFERENCES

1. Dellon ES, Liacouras CA, Molina-Infante J, Furuta GT, Spergel JM, Zevit N, et al. Updated international consensus diagnostic criteria for eosinophilic esophagitis: proceedings of the AGREE conference. Gastroenterology 2018;155:1022-33.e10.
2. Kim GH, Park YS, Jung KW, Kim M, Na HK, Ahn JY, et al. An increasing trend of eosinophilic esophagitis in Korea and the clinical implication of the biomarkers to determine disease activity and treatment response in eosinophilic esophagitis. J Neurogastroenterol Motil 2019;25:525-33.
3. Kim GH, Jung KW, Jung HY, Choi KD, Lee J, Park YS, et al. Diagnostic trends and clinical characteristics of eosinophilic esophagitis: a Korean, single-center database study. J Neurogastroenterol Motil 2018;24:248-54.
4. Hirano I, Moy N, Heckman MG, Thomas CS, Gonsalves N, Achem SR. Endoscopic assessment of the oesophageal features of eosinophilic oesophagitis: validation of a novel classification and grading system. Gut 2013;62:489-95.
5. Yang EJ, Jung KW. Role of endoscopy in eosinophilic esophagitis. Clin Endosc 2025;58:1-9.
6. Lenti MV, Savarino E, Mauro A, Penagini R, Racca F, Ghisa M, et al. Diagnostic delay and misdiagnosis in eosinophilic oesophagitis. Dig Liver Dis 2021;53:1632-9.
7. Okimoto E, Ishimura N, Ishihara S. Clinical characteristics and treatment outcomes of patients with eosinophilic esophagitis and eosinophilic gastroenteritis. Digestion 2021;102:33-40.
8. Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M. Deep learning-based detection of eosinophilic esophagitis. Endoscopy 2022;54:299-304.
9. Okimoto E, Ishimura N, Adachi K, Kinoshita Y, Ishihara S, Tada T. Application of convolutional neural networks for diagnosis of eosinophilic esophagitis based on endoscopic imaging. J Clin Med 2022;11:2529.
10. Römmele C, Mendel R, Barrett C, Kiesl H, Rauber D, Rückert T, et al. An artificial intelligence algorithm is highly accurate for detecting endoscopic features of eosinophilic esophagitis. Sci Rep 2022;12:11115.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [accessed on 2024 March 1]. Available at: http://doi.org/10.1109/CVPR.2016.90.
12. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks [accessed on 2024 March 1]. Available at: http://doi.org/10.1109/TCI.2016.2644865.
13. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [Preprint]. 2014 [accessed on 2024 March 1]. Available at: https://doi.org/10.48550/arXiv.1409.1556.
14. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks [accessed on 2024 March 1]. Available at: https://proceedings.mlr.press/v97/tan19a.html.
15. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst 2022;33:6999-7019.
16. Yang X, Zeng Z, Yeo SY, Tan C, Tey HL, Su Y. A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv [Preprint]. 2017 [accessed on 2024 March 1]. Available at: https://doi.org/10.48550/arXiv.1703.01025.
17. Le TLT, Thome N, Bernard S, Bismuth V, Patoureaux F. Multitask classification and segmentation for cancer diagnosis in mammography. arXiv [Preprint]. 2019 [accessed on 2024 March 1]. Available at: https://doi.org/10.48550/arXiv.1909.05397.
18. Zhou Y, Chen H, Li Y, Liu Q, Xu X, Wang S, et al. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med Image Anal 2021;70:101918.
19. Baldi P. Autoencoders, unsupervised learning, and deep architectures [accessed on 2024 March 1]. Available at: https://proceedings.mlr.press/v27/baldi12a.html.
20. Haghighi F, Taher MRH, Zhou Z, Gotway MB, Liang J. Transferable visual words: exploiting the semantics of anatomical patterns for self-supervised learning. IEEE Trans Med Imaging 2021;40:2857-68.
21. Amyar A, Modzelewski R, Li H, Ruan S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: classification and segmentation. Comput Biol Med 2020;126:104037.
22. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: redesigning skip connections to exploit multiscale features in image seg-

mentation. IEEE Trans Med Imaging 2020;39:1856-67.
23. Feng J, Deng J, Li Z, Sun Z, Dou H, Jia K. End-to-end Res-Unet based reconstruction algorithm for photoacoustic imaging. Biomed Opt Express 2020;11:5321-40.
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization [accessed on 2024 March 1]. Available at: http://doi.org/10.1109/ICCV.2017.74.
26. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention-MICCAI 2015. Cham: Springer; 2015. p.234-41.
27. Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset [accessed on 2024 March 1]. Available at: http://doi.org/10.1109/CVPR.2017.502.
28. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks [accessed on 2024 March 1]. Available at: http://doi.org/10.1109/ICCV.2015.510.