Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-026-01994-0

ORIGINAL PAPER

# Transfer learning for Multi-institutional Classification of Intussusception and Splenomegaly in Pediatric Abdominal Radiographs

**Minsoo Shin[1] · Sungwon Ham[2] · Yoon Lee[3] · Jung Ok Shim[4]**

Received: 16 November 2025 / Revised: 28 April 2026 / Accepted: 29 April 2026
© The Author(s) 2026

## Abstract

To address diagnostic delays in pediatric abdominal emergencies, this study aimed to develop and validate multi-institutional deep learning models for detecting intussusception and splenomegaly on abdominal radiographs, thereby evaluating their potential to enhance clinical triage. This retrospective study included 26,552 radiographs from seven tertiary hospitals (2012–2022). Models were trained with EfficientDet-B2, selected after an ablation study against ResNet-50, DenseNet-121, and other EfficientDet variants. Three strategies were compared: independent binary classifiers, a multiclass model, and transfer learning (fine-tuning from intussusception to splenomegaly). External performance was primarily assessed using a locked external test set, with leave-one-institution-out validation performed as a secondary robustness analysis. Subgroup analyses were performed by institution and body weight (< 10 kg, 10– < 20 kg, 20- < 30 kg), and visualized using area under curve (AUC) with 95% confidence intervals. An ablation study identified EfficientDet-B2 as the best-performing backbone among ResNet-50, DenseNet-121, and other EfficientDet variants. The EfficientDet-B2 models achieved internal AUCs of 0.851 for intussusception and 0.834 for splenomegaly, with locked external AUCs of 0.818 and 0.806, respectively. Secondary leave-one-institution-out (LOIO) AUCs ranged from 0.832 to 0.912, while multiclass training and transfer learning improved intussusception and splenomegaly performance, respectively. Subgroup analyses likewise showed stable AUC across institutions, with reduced sensitivity in the < 10 kg group but maintained performance in those weighing 10– < 30 kg. Multiclass training improved intussusception detection, whereas transfer learning enhanced splenomegaly recognition. With robust performance across institutions and clinically interpretable subgroup analyses, these models show promise for supporting triage of pediatric abdominal radiographs. Prospective validation in real-world workflows is warranted to confirm utility.

**Keywords** Intussusception · Splenomegaly · X-ray · Deep learning · Pediatric imaging

Minsoo Shin and Sungwon Ham contributed equally to this article as first authors.

Yoon Lee and Jung Ok Shim contributed equally to this article as co-corresponding authors.

✉ Yoon Lee
ragabash@korea.ac.kr

✉ Jung Ok Shim
shimjo@snu.ac.kr

[1] Department of Pediatrics, Korea University Medical Center Ansan Hospital, 123 Jeokgeum-Ro, Danwon-Gu, Ansan 15355, Gyeonggi-do, Republic of Korea

[2] Healthcare Readiness Institute for Unified Korea, Korea University Ansan Hospital, Korea University College of Medicine, 123 Jeokgeum-Ro, Danwon-Gu, Ansan 15355, Gyeonggi-do, Republic of Korea

[3] Department of Pediatrics, Korea University Medical Center Anam Hospital, 73, Goryeodae-ro, Seongbuk-gu, Seoul 02841, Republic of Korea

[4] Department of Pediatrics, Seoul National University College of Medicine, 101 Daehak-Ro, Jongno-Gu, Seoul 03080, Republic of Korea

Published online: 12 May 2026

## Introduction

Intussusception is the leading cause of intestinal obstruction in infants and young children and demands urgent recognition to prevent bowel ischemia, necrosis, and perforation [1–3]. Despite its clinical importance, the classic triad of colicky abdominal pain, hematochezia, and a palpable abdominal mass is observed in fewer than half of cases [4]. Instead, many children present with nonspecific symptoms such as irritability and vomiting, which complicate timely diagnosis and often contribute to delayed or missed recognition. Ultrasonography is regarded as the diagnostic standard owing to its excellent sensitivity and specificity [5], but its accuracy is operator dependent and contingent upon pediatric expertise that is not always available in emergency settings, particularly in resource-limited hospitals. In contrast, abdominal radiography is inexpensive, rapid, and universally accessible, but its diagnostic sensitivity is modest [5]. If its potential as a screening modality could be enhanced, abdominal radiography might substantially improve early detection and triage of intussusception in environments where pediatric ultrasonography is not readily available.

Splenomegaly represents another clinically important but diagnostically challenging pediatric finding. Splenic enlargement occurs in a wide spectrum of conditions, including infections, hematologic and hepatic disorders, malignancies, autoimmune disease, and metabolic syndromes [6–8]. Pediatric splenomegaly is defined using age- or body size–specific percentiles, and accurate identification is essential to guide further investigation of underlying systemic disease. However, physical examination by palpation is highly variable, with sensitivity as low as 58% and frequent false positives in children with normally sized spleens [9]. While advanced imaging modalities such as ultrasound, CT, and MRI allow direct measurement, their cost, limited availability, and long examination times restrict their use as frontline screening tools [10]. Abdominal radiography, although limited in delineating splenic margins, remains among the most frequently acquired imaging studies in pediatric practice. An automated approach that could leverage these ubiquitous radiographs for opportunistic splenomegaly detection would provide a valuable adjunct for identifying underlying systemic disease, transforming a routine examination into an opportunistic screening tool.

Artificial intelligence, particularly deep learning with convolutional neural networks, has shown transformative potential in radiology [11, 12]. AI algorithms have achieved high accuracy in detecting pneumonia, tuberculosis, and other conditions on chest radiographs [13, 14]. However, applications in pediatric abdominal imaging remain sparse, and most prior studies have been limited to single institutions or modest datasets, raising concerns regarding reproducibility and generalizability. Furthermore, clinical acceptance requires interpretability, yet only a few works have aligned model outputs with clinically relevant radiographic features, leaving a persistent gap in explainability. These limitations underscore the need for large-scale, multi-institutional validation with interpretable models, which formed the primary motivation for this study.

To address these challenges, we developed and validated deep learning–based classification models for pediatric abdominal radiographs using a large, multi-institutional dataset from seven tertiary hospitals. Focusing on intussusception and splenomegaly, we systematically compared independent binary classifiers, a multiclass model, and transfer-learning strategies. Our objectives were twofold: (1) to determine whether these models could achieve reproducible and interpretable performance across heterogeneous institutions and patient subgroups, and (2) to evaluate their potential role as adjunctive tools for early detection and triage in pediatric emergency imaging workflows.

## Method

### Study Population and Data Acquisition

We performed a retrospective, multicentre study of pediatric abdominal radiographs obtained between 2012 and 2022 from seven tertiary hospitals in South Korea (Hospital A, Hospital B, Hospital C, Hospital D, Hospital E, F and G). Image collection was finalized between July and December 2022. Institutional review board (IRB) approval was obtained at each participating site (Hospital A, No. 2022AS0278; Hospital B, No. SMC2022—02—093—002; Hospital C, No. 2022—0418—0002; Hospital D, No. AJOUIRB-MDB-2022—192; Hospital E, No. CBNUH2022—02—015—002; Hospital F, No. CHAMC2022—04—031—003 and Hospital G, No. 2021—08—063, 2022—05—010), with a waiver of informed consent owing to the retrospective design. The study complied with the Declaration of Helsinki.

For intussusception, radiographs acquired within 24 h of either non-operative or operative reduction were included, identified using the Korean Standard Classification of Diseases (KCD) code K56.1. For splenomegaly, radiographs were selected from children with KCD codes R16.1 or R16.2 in which ultrasound or CT reports explicitly contained the term "splenomegaly" or "hepatosplenomegaly." The abdominal radiograph used for model training was selected as the image closest in time within ± 24 h of the confirming

ultrasound or CT examination. Each patient contributed a single abdominal radiograph to the dataset. When multiple radiographs were available within the defined time window, the image closest in time to the reference examination or procedure was selected. Controls were defined as children without abdominal disease, prior abdominal surgery, or abnormal findings, whose abdominal radiographs had been reported as "normal."

Patients were stratified by body weight into three subgroups: W1 (< 10 kg), W2 (10– < 20 kg), and W3 (20– < 30 kg). To ensure annotation quality, trained research assistants performed initial labeling under the supervision of a pediatric gastroenterologist. One pediatric gastroenterologist and two pediatric radiologists independently reviewed all cases, and disagreements were resolved by consensus.

## Dataset and Splitting

We initially collected 26,552 abdominal radiographs from seven tertiary hospitals (2012—2022), including 4,490 with intussusception, 1,135 with splenomegaly, and 20,927 normal controls. After excluding 790 images with incorrect orientation or excessive artefact (3.4% of the dataset), the final dataset comprised 25,762 radiographs (4,424 intussusception, 1,107 splenomegaly, and 20,203 normal). To prevent data leakage, all dataset partitions were performed at the patient level. For the primary evaluation, one institution was designated as a locked external test set and was not used during model training, validation, hyperparameter tuning, or internal testing. The remaining six institutions formed the development cohort and were split into training, validation, and internal test sets in a 70:15:15 ratio. Model performance was evaluated on both the internal test set and the locked external test set. Separately, leave-one-institution-out (LOIO) validation was performed as a secondary robustness analysis by iteratively holding out each institution once and training the model on the remaining institutions.

## Network Architectures and Model Training

To identify the optimal network architecture, we conducted an ablation study comparing ResNet-50 [15], DenseNet-121 [16], and EfficientDet variants (B0-B4) [17, 18]. EfficientNet-B2 was additionally included as a classification-oriented comparator to assess the potential contribution of the bidirectional feature pyramid network (BiFPN) in our image-level classification task. Based on classification performance and computational efficiency, EfficientDet-B2 was selected as the primary architecture for all subsequent experiments. This network integrates EfficientNet-B2 as the feature-extraction backbone with BiFPN for multi-scale feature fusion, followed by shared prediction heads for classification. Model weights pretrained on the COCO dataset were used for initialization and fine-tuned on pediatric abdominal radiographs [19].

All images were de-identified prior to analysis to remove patient information. Any visible textual annotations or markers embedded in the radiographs were removed when present. Images were resized to a standardized resolution, and intensity normalization and contrast adjustment were applied. Radiologist-defined regions of interest were annotated as pixel-wise masks rather than bounding boxes, and fill-hole correction was applied to generate continuous masks before training.

These masks were incorporated as auxiliary supervision through a binary cross-entropy–based segmentation loss. The overall training objective combined classification and segmentation losses as

$$L_total = L_cls + \lambda L_seg,$$

where L_cls denotes the cross-entropy classification loss and L_seg represents the auxiliary segmentation loss. The weighting coefficient λ was set to 0.3 in our experiments, allowing segmentation supervision to guide spatial attention while maintaining the primary focus on classification. ROI masks were available for abnormal cases (intussusception and splenomegaly), whereas normal images did not contain ROI annotations. For images without ROI masks, the segmentation loss was not applied during training. This auxiliary supervision guided the network's attention toward diagnostically relevant abdominal regions while suppressing irrelevant background signals [20].

All radiographs were resized according to the recommended input resolution of each EfficientDet variant (512 × 512 for B0, 768 × 768 for B2, and 1024 × 1024 for B4), following the compound scaling design of EfficientDet. Images were normalized to the [0–1] range and augmented with random horizontal flipping, small rotations, and intensity variations to improve robustness and generalizability.

Three training strategies were evaluated. In the independent binary setting, separate models were trained for intussusception versus normal and splenomegaly versus normal. In the multiclass setting, a single model was trained to simultaneously classify intussusception, splenomegaly, and normal cases. In the transfer-learning setting, a model pretrained on the intussusception dataset was subsequently fine-tuned for splenomegaly classification, enabling feature representations learned from the larger cohort to be applied to the smaller and more subtle task.

For the subgroup analyses, model performance was evaluated separately across participating institutions to assess external consistency. In addition, patients were stratified into three body-weight groups: W1 (< 10 kg), W2 (10– < 20 kg), and W3 (20– < 30 kg).

All models were trained with the AdamW optimizer (initial learning rate = 0.0001) for up to 1,000 epochs with early stopping based on validation loss [21, 22]. The batch size was set to 32 for the binary and multiclass models, while the transfer-learning model used a reduced batch size of 16 to accommodate GPU memory constraints. Cross-entropy loss was applied, with class weighting in the multiclass setting to mitigate imbalance [23]. The overall workflow, including dataset building, preprocessing, model training, and statistical analysis, is summarized in Fig. 1.

## Evaluation Metrics

Model performance was assessed using accuracy, sensitivity, specificity, F1-score, and the area under curve (AUC). The definitions of these evaluation metrics are summarized in Supplementary Figure S1. Accuracy reflected the overall proportion of correctly classified cases. Sensitivity and specificity quantified the ability to correctly identify positive and negative cases, and the F1-score summarized the balance between precision and recall in the presence of class imbalance. AUC was used as the primary measure of model discrimination and was reported with 95% confidence intervals (CIs) [24, 25]. Pairwise differences in AUCs between models were assessed using the DeLong test, with statistical significance defined as $p < 0.05$ [26]. For subgroup analyses, performance was evaluated using AUC with 95% CIs only, stratified by institution and body-weight groups (W1: < 10 kg, W2: 10– < 20 kg, W3: 20– < 30 kg).

To ensure strict external validation, one institution was designated as a locked external test set, selected to ensure enough cases and a balanced class distribution across target conditions. This institution was not used during model training, validation, or internal testing. The remaining six institutions were used to construct the development cohort, which was further split into training (70%), validation (15%), and internal test (15%) subsets. Model performance was reported on both the internal test set and the external test set. In addition, we performed a LOIO analysis as a secondary analysis to evaluate model robustness across institutions. Detailed results are provided in Supplementary Materials.

## Results

A total of 25,762 pediatric abdominal radiographs were included after exclusion of low-quality images, comprising 4,424 with intussusception, 1,107 with splenomegaly, and 20,203 normal controls. Patient-level stratification was applied to prevent overlap across cohorts. Model performance was evaluated on the internal test set and a locked external test set, with LOIO validation performed as a secondary analysis.

The performance of different backbone architectures in the independent binary classification tasks is summarized in Table 1 and Fig. 2. For intussusception (Fig. 2, left panel), the AUC values were 0.791 for ResNet-50, 0.805 for DenseNet-121, 0.845 for EfficientDet-B0, 0.881 for

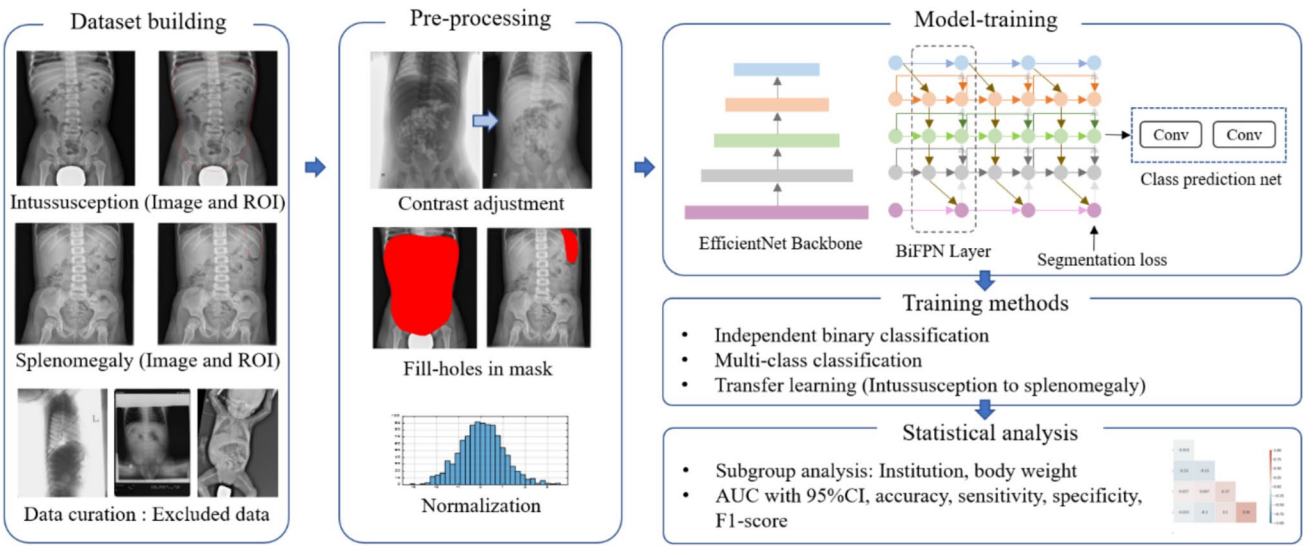



**Fig. 1** The overall workflow of this study. The study workflow consisted of four stages. Dataset building: abdominal radiographs with intussusception and splenomegaly were collected with corresponding regions of interest, and inappropriate or corrupted data were excluded. Pre-processing: images underwent contrast adjustment, mask refinement, and normalization. Model training: EfficientDet-B2 served as the backbone, with training performed using three strategies-independent binary classification, multiclass classification, and transfer learning (fine-tuning from intussusception to splenomegaly). Statistical analysis: performance was evaluated with AUC (95% CI), accuracy, sensitivity, specificity, and F1-score, including subgroup analyses by institution and body weight

**Table 1** Ablation study comparing backbone architectures in the independent binary classification

| | | AUC (95% CI) | Accuracy | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|
| ResNet-50 | Intussusception | 0.791* (0.751 0.838) | 0.800 ± 0.061 | 0.785 ± 0.066 | 0.815 ± 0.039 | 0.784 ± 0.067 |
| | Splenomegaly | 0.798* (0.748 0.841) | 0.804 ± 0.059 | 0.790 ± 0.062 | 0.818 ± 0.037 | 0.792 ± 0.068 |
| DenseNet-121 | Intussusception | 0.805* (0.762 0.840) | 0.816 ± 0.052 | 0.800 ± 0.055 | 0.832 ± 0.042 | 0.794 ± 0.064 |
| | Splenomegaly | 0.811 (0.764 0.851) | 0.822 ± 0.037 | 0.810 ± 0.049 | 0.834 ± 0.044 | 0.812 ± 0.050 |
| EfficientDet-B0 | Intussusception | 0.845* (0.792 0.893) | 0.851 ± 0.033 | 0.840 ± 0.033 | 0.864 ± 0.030 | 0.831 ± 0.044 |
| | Splenomegaly | 0.829 (0.784 0.870) | 0.840 ± 0.036 | 0.830 ± 0.041 | 0.852 ± 0.031 | 0.828 ± 0.035 |
| EfficientDet-B2 | Intussusception | 0.881 (0.840 0.922) | 0.876 ± 0.025 | 0.851 ± 0.035 | 0.870 ± 0.026 | 0.837 ± 0.033 |
| | Splenomegaly | 0.841 (0.805 0.879) | 0.864 ± 0.026 | 0.850 ± 0.038 | 0.878 ± 0.022 | 0.842 ± 0.031 |
| EfficientDet-B4 | Intussusception | 0.856 (0.807 0.901) | 0.847 ± 0.038 | 0.841 ± 0.036 | 0.853 ± 0.036 | 0.830 ± 0.029 |
| | Splenomegaly | 0.838 (0.798 0.879) | 0.859 ± 0.032 | 0.842 ± 0.030 | 0.874 ± 0.024 | 0.837 ± 0.028 |

**The asterisk shows statistically significant differences in the area under curve (AUC) compared to EfficientDet-B2 ($p < 0.05$) based on the DeLong test*

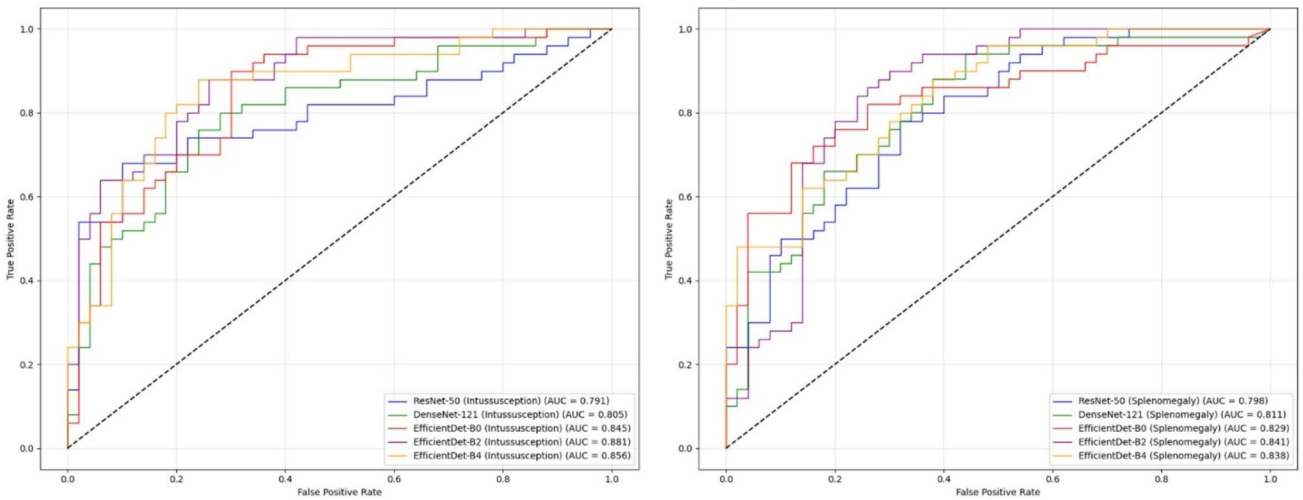



**Fig. 2** Receiver operating characteristic (ROC) curves for the multi-class classification model and independently trained binary classification models for intussusception and splenomegaly. The left panel shows results for intussusception, and the right panel for splenomegaly. Each line represents the performance of a specific architectural setup

EfficientDet-B2, and 0.856 for EfficientDet-B4. Statistical comparisons indicated that the AUCs of ResNet-50, DenseNet-121, and EfficientDet-B0 were significantly lower than that of EfficientDet-B2 ($p < 0.05$). For splenomegaly (Fig. 2, right panel), the AUC values were 0.798 for ResNet-50, 0.811 for DenseNet-121, 0.829 for EfficientDet-B0, 0.841 for EfficientDet-B2, and 0.838 for EfficientDet-B4. The AUCs of ResNet-50 and DenseNet-121 were significantly lower than that of EfficientDet-B2 ($p < 0.05$), whereas EfficientDet-B0 and EfficientDet-B4 showed no significant differences. Based on these results, EfficientDet-B2 was selected as the backbone architecture for all subsequent analyses. Detailed accuracy, sensitivity, specificity, and F1-scores are provided in Table 1.

On the locked external test set, the best-performing model achieved an AUC of 0.818 (95% CI, 0.754–0.858) for intussusception and 0.806 (95% CI, 0.773–0.836) for splenomegaly. On the internal test set, the corresponding AUCs were 0.851 (95% CI, 0.817–0.892) and 0.834 (95% CI, 0.796–0.875), respectively. These results indicate a modest decrease in performance on external data, reflecting real-world generalizability.

To further evaluate the architectural contribution of the BiFPN, we additionally compared EfficientDet-B2 with EfficientNet-B2 under the same experimental setting (Table S1). EfficientDet-B2 showed numerically higher AUCs than EfficientNet-B2 for both intussusception (0.881 vs. 0.857) and splenomegaly (0.841 vs. 0.822), although the differences were not statistically significant according to the DeLong test.

Using EfficientDet-B2, three training strategies were compared (Table 2). For intussusception, the AUC was 0.881 for the independent model and 0.904 for the multi-class model, with the latter showing the highest performance. For splenomegaly, the AUCs were 0.841 for the independent model, 0.839 for the multi-class model, and 0.857 for the transfer learning model, with transfer learning achieving the best performance.

We further evaluated model performance on the internal test set and the locked external institutional test set (Table 3). For intussusception, the model achieved an AUC of 0.851 (95% CI, 0.817–0.892) on the internal test set (sensitivity, 0.828; specificity, 0.847) and 0.818 (95% CI, 0.754–0.858) on the locked external test set (sensitivity, 0.804; specificity, 0.835). For splenomegaly, the model achieved an AUC of 0.834 (95% CI, 0.796–0.875) on the internal test set (sensitivity, 0.813; specificity, 0.825) and 0.806 (95% CI, 0.773–0.836) on the locked external test set (sensitivity, 0.782; specificity, 0.826). These results indicate a modest decrease in performance on external data while supporting generalizability across institutions.

**Table 3** Comparison of internal and locked external test performance

| Task | Dataset | AUC (95% CI) | Sensitivity | Specificity |
|---|---|---|---|---|
| Intussusception | Internal | 0.851 (0.817 0.892) | 0.828 | 0.847 |
| | External | 0.818 (0.754 0.858) | 0.804 | 0.835 |
| Splenomegaly | Internal | 0.834 (0.796 0.875) | 0.813 | 0.825 |
| | External | 0.806 (0.773 0.836) | 0.782 | 0.826 |

Subgroup analyses were conducted to evaluate model performance across institutions (Fig. 3). For intussusception, the AUC values ranged from 0.867 (Hospital D) to 0.912 (Hospital C). For splenomegaly, the AUC values ranged from 0.832 (Hospitals A and G) to 0.850 (Hospital C). Error bars indicate 95% confidence intervals for each institution. The number of cases per institution is summarized in Supplementary Table S2, and accuracy, sensitivity, specificity, and F1-scores are provided in Supplementary Table S3.

Performance was also examined across patient weight groups (Fig. 4). For intussusception, the AUC values were 0.834 in W1 (< 10 kg), 0.901 in W2 (10– < 20 kg), and 0.898 in W3 (20– < 30 kg). For splenomegaly, the corresponding values were 0.782, 0.858, and 0.861 in W1, W2, and W3, respectively. Error bars represent 95% confidence intervals.

**Table 2** Performance comparison between multi-class and independent binary classification models

| | | AUC (95% CI) | Accuracy | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|
| Intussusception | Independent | 0.881 (0.844 0.917) | 0.876 ± 0.019 | 0.841 ± 0.028 | 0.900 ± 0.024 | 0.837 ± 0.033 |
| | Multi-class | 0.904* (0.860 0.948) | 0.874 ± 0.021 | 0.846 ± 0.025 | 0.922 ± 0.015 | 0.847 ± 0.026 |
| Splenomegaly | Independent | 0.841 (0.799 0.872) | 0.864 ± 0.031 | 0.850 ± 0.029 | 0.890 ± 0.022 | 0.842 ± 0.031 |
| | Multi-class | 0.839 (0.785 0.881) | 0.825 ± 0.042 | 0.821 ± 0.035 | 0.875 ± 0.032 | 0.812 ± 0.040 |
| | Transfer | 0.857 (0.806 0.912) | 0.871 ± 0.025 | 0.842 ± 0.024 | 0.881 ± 0.025 | 0.848 ± 0.032 |

**The asterisk represents a statistically significant area under curve (AUC) difference over the multi-class model based on the DeLong test ($p < 0.05$)*

The number of cases in each weight group is shown in Supplementary Table S4, and detailed performance metrics are summarized in Supplementary Table S5. To further assess whether the model focuses on anatomically plausible regions, we compared Grad-CAM activation maps with radiologist-drawn ROI masks (Figure S2).

## Discussion

In this multicenter study of more than 25,000 pediatric abdominal radiographs collected from seven tertiary hospitals, we demonstrated the feasibility of deep learning–based detection of intussusception and splenomegaly. Pediatric abdominal radiographs are among the most frequently performed yet most difficult examinations to interpret, with subtle findings, overlapping gas patterns, and wide inter-reader variability. By leveraging a large-scale, heterogeneous dataset and systematic model evaluation, our study addresses these challenges and provides one of the most comprehensive assessments of AI for pediatric abdominal imaging to date.

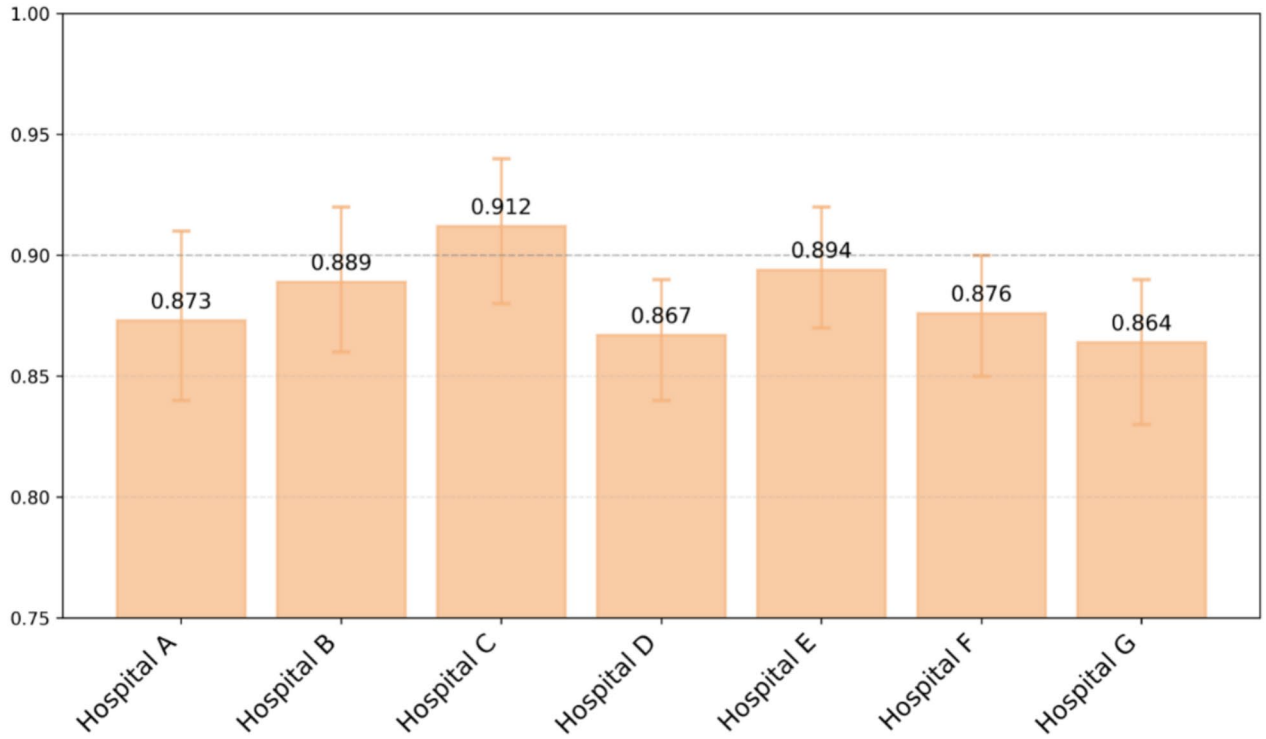
We selected intussusception and splenomegaly as target conditions because they represent clinically important but diagnostically challenging findings. Intussusception is a common pediatric emergency in which radiographic findings are often subtle, and although ultrasonography is the diagnostic gold standard, it remains operator dependent and variably available in acute settings. Splenomegaly, in contrast, is a critical indicator of systemic disease but is especially difficult to recognize in children due to age-dependent normal variations in splenic size and the limited sensitivity of physical examination. To our knowledge, no validated automated approach for splenomegaly detection on radiographs has previously been reported. Thus, our study fills an important gap by systematically evaluating AI models for both conditions on plain radiographs, a modality that is universally available in pediatric practice.

Our ablation experiments comparing ResNet-50, DenseNet-121, and EfficientDet variants demonstrated that EfficientDet-B2 consistently achieved the best overall performance. The use of EfficientDet in this study was motivated by its BiFPN, which enables effective multi-scale feature fusion. In pediatric abdominal radiographs, radiographic findings such as intussusception and splenomegaly may vary considerably in spatial scale, particularly across different body-weight groups. In addition, our framework incorporated mask annotations with an auxiliary segmentation loss to guide the network toward diagnostically relevant abdominal regions. The multi-scale representation provided by BiFPN may facilitate the integration of this spatial supervision, improving the model's ability to capture subtle radiographic patterns. Its compound scaling strategy and BiFPN architecture may therefore facilitate recognition of subtle and variably sized features, supporting its use for subsequent analyses. In a direct comparison with EfficientNet-B2, a classification-oriented backbone without BiFPN, EfficientDet-B2 showed numerically higher performance for both intussusception and splenomegaly, although the differences did not reach statistical significance. These findings suggest that BiFPN-based multi-scale feature fusion may provide a modest benefit in this image-level classification setting.

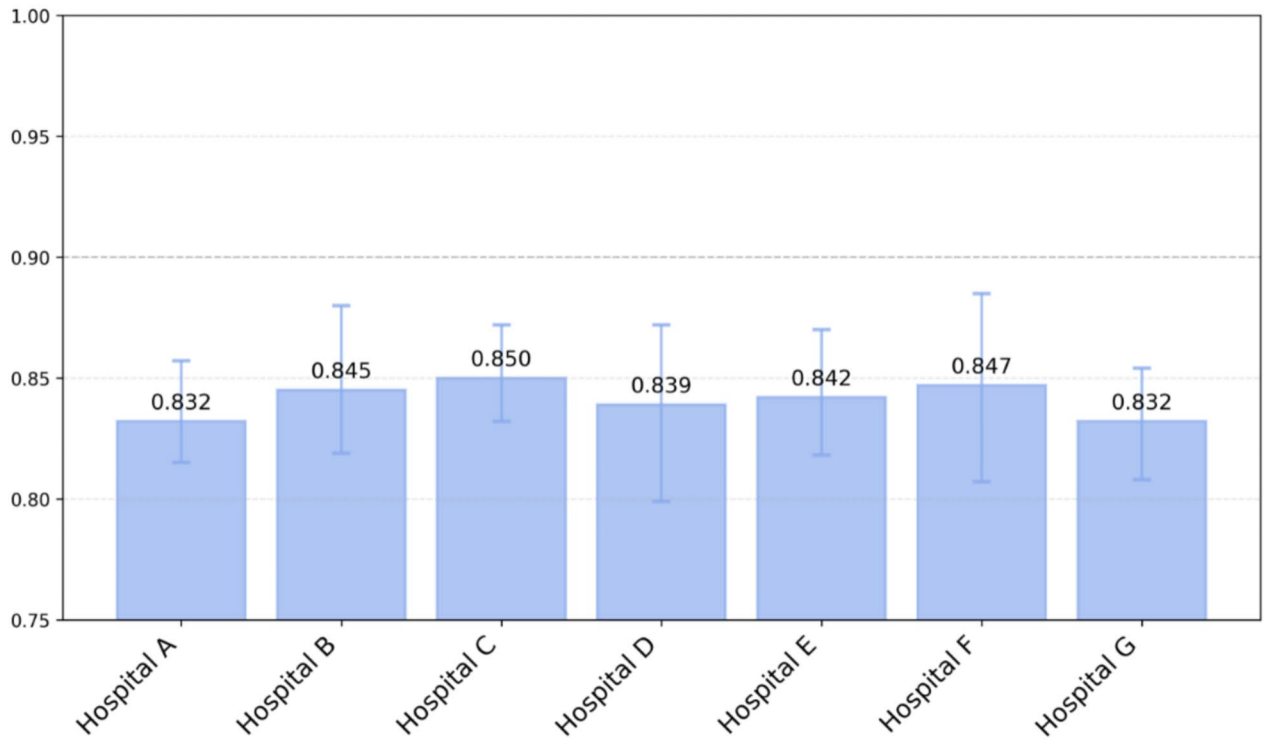
Building on this backbone, we compared three training strategies. For intussusception, multiclass classification yielded the highest AUC, suggesting that simultaneously modeling normal, intussusception, and splenomegaly improved feature discrimination by leveraging shared context. For splenomegaly, however, transfer learning from intussusception provided the greatest benefit, with the highest AUC and F1-score. This finding indicates that representations learned from detecting intussusception, including abdominal contour and organ-level features, could be effectively adapted to the more subtle task of splenic enlargement detection. These strategy-specific differences underscore the importance of tailoring model design to disease characteristics.

Subgroup analyses further supported the robustness and generalizability of the models. Institution-wise performance remained consistently high across seven hospitals with heterogeneous imaging equipment and protocols. The locked external test results provide the primary evidence of generalizability to unseen populations, while LOIO validation further supports robustness across institutions. Weight stratification revealed a size-related effect: relative to W3, W1 showed significantly lower AUCs for both intussusception (0.834 vs 0.898) and splenomegaly (0.782 vs 0.861) ($p < 0.05$; Table S4), consistent with the reduced sensitivity observed in infants.

The relatively lower performance observed in the < 10 kg group may be related, at least in part, to the age and weight distribution of the study cohort, in addition to the greater difficulty of image interpretation in younger infants. Intussusception most commonly occurs in infants aged approximately 1–2 years and is rare in neonates. Consequently, within the < 10 kg subgroup, patients with intussusception are likely concentrated near the upper end of this weight category, whereas the normal group may include a broader age range of younger infants within the same weight category. This difference in age distribution may influence model




(a)




(b)

learning and classification performance. For splenomegaly, even human interpretation on abdominal radiographs is known to have relatively low sensitivity and specificity, particularly in cases of mild splenomegaly, which can be difficult to detect. In addition, image interpretation is inherently more challenging in the W1 subgroup. This is largely

◂**Fig. 3** Institution-wise performance for intussusception and splenomegaly classification. The upper panel presents the AUC with 95% confidence intervals for detecting intussusception, while the lower panel shows corresponding results for splenomegaly. Each bar represents the mean AUC for a given institution, with error bars indicating the confidence interval. Hospital designations (A–G) are as follows: A, Korea University Medicine; B, Seoul Samsung Medical Center; C, Catholic University Hospital; D, Ajou University Hospital; E, Chungbuk National University Hospital; F, CHA Bundang Medical Center; and G, Keimyung University Dongsan Hospital

due to the increased amount and complexity of gastrointestinal gas patterns. Most radiographs in this age group are obtained in the supine position, in which bowel gas is more diffusely distributed, the gastric bubble is less well defined, and displacement of abdominal organs is more difficult to appreciate. Furthermore, splenomegaly is more readily identified during adequate inspiration; however, patients in the W1 subgroup are generally unable to cooperate with breath control during imaging, which may have further contributed to the observed decrease in performance. Future studies with narrower weight ranges may help clarify model performance in younger patients. Several strengths of this study warrant emphasis. First, the dataset size and diversity exceed most prior work in pediatric imaging, reducing center-specific and demographic biases. Second, patient-level stratification prevented overlap across cohorts and ensured robust evaluation across developmental stages. Third, by directly comparing multiple backbones and training strategies, we provide practical insights into model design rather than reporting performance of a single architecture. Finally, subgroup and external validations provide strong evidence for generalizability, a prerequisite for clinical translation.

Nonetheless, several limitations should be acknowledged. In addition, splenomegaly is inherently difficult to assess on abdominal radiographs, even for experienced readers, because radiographic findings are often subtle and less specific than those of other abdominal conditions. Thus, the model performance for splenomegaly may partly reflect the intrinsic limitations of the imaging modality itself. The retrospective design may introduce variability in acquisition parameters and image quality. In addition, only imaging data were used; integration of clinical information, such as symptoms or laboratory values, may further improve patient-specific decision support. The study was also limited to two target diseases, and expansion to a broader range of abdominal conditions will be necessary to develop clinically comprehensive systems.

Although LOIO validation was performed as a secondary analysis, a stricter locked external test set was used to provide a more realistic assessment of generalizability. Prospective validation in real-world clinical workflows is essential to confirm its impact on diagnostic accuracy and turnaround time. As this was a multicentre study, differences in disease prevalence and case-mix across institutions may influence model performance. To address this, we employed a leave-one-institution-out validation strategy to evaluate generalizability across heterogeneous institutional distributions. Nevertheless, residual inter-institutional variability may persist, and future studies with more balanced or stratified datasets are warranted.

In conclusion, our multicenter study demonstrates that deep learning can provide accurate and generalizable detection of intussusception and splenomegaly on pediatric abdominal radiographs. Multiclass training improved performance for intussusception, while transfer learning substantially enhanced splenomegaly detection, reflecting disease-specific optimization of training strategies. With reproducible performance across diverse institutions and

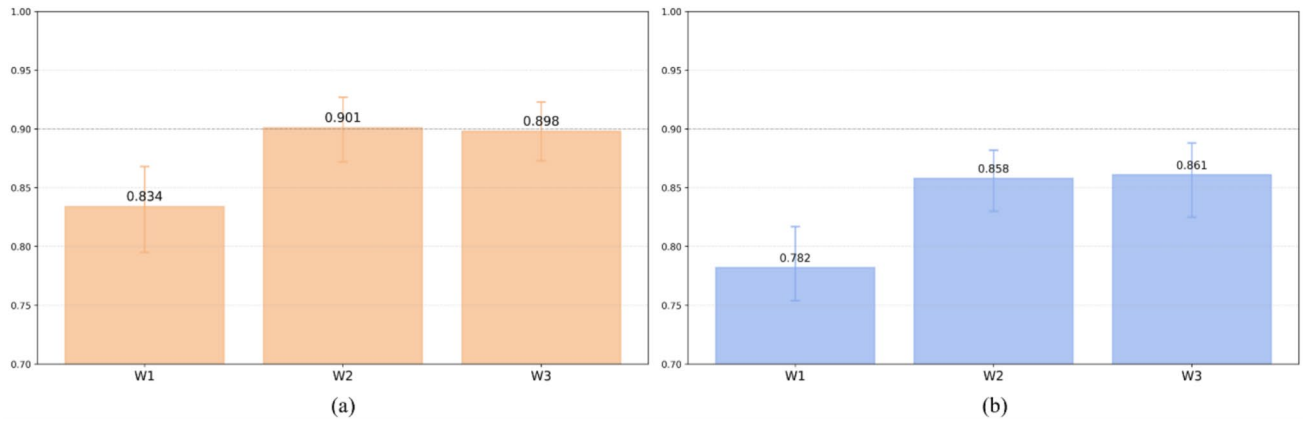



**Fig. 4** Weight group–wise performance comparison for intussusception and splenomegaly classification. The left panel shows results for intussusception, and the right panel for splenomegaly. Weight groups were defined as W1 (< 10 kg), W2 (10– < 20 kg), and W3 (20– < 30 kg). Each bar represents the mean AUC for the corresponding weight group, with error bars indicating 95% confidence intervals

subgroups, these models show promise as decision-support tools to aid radiologists and clinicians in emergency and resource-limited settings. Prospective multicenter studies will be the critical next step toward establishing their role in routine pediatric practice.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10278-026-01994-0.

**Authors' Contributions** Minsoo Shin: data curation, investigation, writing-original draft.

Sungwon Ham: methodology, formal analysis, writing-original draft.

Yoon Lee: data curation, methodology, investigation, project administration, writing-review and editing.

Jung Ok Shim: conceptualization, data curation, methodology, investigation, project administration, writing-review and editing.

All authors reviewed and approved the final manuscript.

**Funding** Open Access funding enabled and organized by Seoul National University.

**Data Availability** The data used in this study are not publicly available due to ethical and institutional restrictions. De-identified data may be available from the corresponding author upon reasonable request and with approval from the participating institutions.

## Declarations

**Ethical Approval** This retrospective multicentre study was approved by the Institutional Review Boards of all participating institutions: Hospital A (No. 2022AS0278), Hospital B (No. SMC2022—02—093—002), Hospital C (No. 2022—0418—0002), Hospital D (No. AJOUIRB-MDB-2022—192), Hospital E (No. CBNUH2022—02—015—002), Hospital F (No. CHAMC2022—04—031—003), and Hospital G (No. 2021—08—063, 2022—05—010). The requirement for informed consent was waived owing to the retrospective design. The study complied with the Declaration of Helsinki.

**Consent to Participate** The requirement for informed consent was waived by the Institutional Review Boards because of the retrospective study design.

**Consent to Publish** Not applicable.

**Conflict of interest** The authors declare no competing interests.

**Competing interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. 1 Stringer, M. D., Pablot, S. M. & Brereton, R. J. Paediatric intussusception. *Br J Surg* **79**, 867-876 (1992). https://doi.org/10.1002/bjs.1800790906
2. 2 Demirkan, A. *et al.* Intussusception in adult and pediatric patients: two different entities. *Surg Today* **39**, 861-865 (2009). https://doi.org/10.1007/s00595-009-3979-8
3. Gluckman, S., Karpelowsky, J., Webster, A. C. & McGee, R. G. Management for intussusception in children. *Cochrane Database Syst Rev* **6**, Cd006476 (2017). https://doi.org/10.1002/14651858.CD006476.pub3
4. 4 Edwards, E. A. *et al.* Intussusception: past, present and future. *Pediatr Radiol* **47**, 1101-1108 (2017). https://doi.org/10.1007/s00247-017-3878-x
5. 5 Carroll, A. G. *et al.* Comparative Effectiveness of Imaging Modalities for the Diagnosis and Treatment of Intussusception: A Critically Appraised Topic. *Acad Radiol* **24**, 521-529 (2017). https://doi.org/10.1016/j.acra.2017.01.002
6. 6 Suttorp, M. & Classen, C. F. Splenomegaly in Children and Adolescents. *Front Pediatr* **9**, 704635 (2021). https://doi.org/10.3389/fped.2021.704635
7. 7 Rosenberg, H. K. *et al.* Normal splenic size in infants and children: sonographic measurements. *AJR Am J Roentgenol* **157**, 119-121 (1991). https://doi.org/10.2214/ajr.157.1.2048509
8. 8 Prassopoulos, P. & Cavouras, D. CT assessment of normal splenic size in children. *Acta Radiol* **35**, 152-154 (1994).
9. Grover, S. A., Barkun, A. N. & Sackett, D. L. The rational clinical examination. Does this patient have splenomegaly? *Jama* **270**, 2218–2221 (1993).
10. 10 Pozo, A. L., Godfrey, E. M. & Bowles, K. M. Splenomegaly: investigation, diagnosis and management. *Blood Rev* **23**, 105-111 (2009). https://doi.org/10.1016/j.blre.2008.10.001
11. 11 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015). https://doi.org/10.1038/nature14539
12. Mathew, Amitha, P. Amudha, and S. Sivakumari. "Deep learning techniques: an overview." International conference on advanced machine learning technologies and applications. Singapore: Springer Singapore, 2020.
13. Alapat, D. J., Menon, M. V. & Ashok, S. A Review on Detection of Pneumonia in Chest X-ray Images Using Neural Networks. *J Biomed Phys Eng* **12**, 551–558 (2022). https://doi.org/10.31661/jbpe.v0i0.2202-1461
14. 14 Chen, C. F. *et al.* A deep learning-based algorithm for pulmonary tuberculosis detection in chest radiography. *Sci Rep* **14**, 14917 (2024). https://doi.org/10.1038/s41598-024-65703-z
15. Koonce, Brett. "ResNet 50." *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*. Berkeley, CA: Apress, 2021. 63–72.
16. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
17. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.
18. Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
19. Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Cham: Springer International Publishing, 2014.
20. Liu, Yifan, et al. "Auxiliary learning for deep multi-task learning." *arXiv preprint* arXiv:1909.02214 *(2019).*

21. Llugsi, Ricardo, et al. "Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito." *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*. IEEE, 2021.
22. Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint* arXiv:1711.05101 *(2017).*
23. Mao, Anqi, Mehryar Mohri, and Yutao Zhong. "Cross-entropy loss functions: Theoretical analysis and applications." *International conference on Machine learning*. pmlr, 2023.
24. Hossin, Mohammad, and Md Nasir Sulaiman. "A review on evaluation metrics for data classification evaluations." *International journal of data mining & knowledge management process* 5.2 (2015): 1.
25. 25 Vujović, Željko. "Classification model evaluation metrics." *International Journal of Advanced Computer Science and Applications* 12.6 (2021): 599-606.
26. DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." *Biometrics* (1988): 837–845.

**Publisher'S Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.