Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242950

PEDIATRIC RADIOLOGY

ORIGINAL ARTICLE

Diagnostic accuracy of convolutional neural network algorithms to distinguish gastrointestinal obstruction on conventional radiographs in a pediatric population

Ercan Ayaz[1]
Hasan Güçlü[2]
Ayşe Betül Oktay[3]

[1]Diyarbakır Children's Hospital, Radiology Clinic, Diyarbakır; Current: University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

[2]İstanbul Medeniyet University Faculty of Engineering and Natural Sciences, Department of Biostatistics and Medical Informatics, İstanbul; Current: TOBB University of Economics and Technology, Department of Artificial Intelligence Engineering, Ankara, Türkiye

[3]Yıldız Technical University Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

Corresponding author: Ercan Ayaz

E-mail: ercan.ayaz1@gmail.com

Received 31 October 2024; revision requested 12 December 2024; accepted 13 January 2025.

Epub: 28.02.2025

DOI: 10.4274/dir.2025.242950

PURPOSE
Gastrointestinal (GI) dilatations are frequently observed in radiographs of pediatric patients who visit emergency departments with acute symptoms such as vomiting, pain, constipation, or diarrhea. Timely and accurate differentiation of whether there is an obstruction requiring surgery in these patients is crucial to prevent complications such as necrosis and perforation, which can lead to death. In this study, we aimed to use convolutional neural network (CNN) models to differentiate healthy children with normal intestinal gas distribution in abdominal radiographs from those with GI dilatation or obstruction. We also aimed to distinguish patients with obstruction requiring surgery and those with other GI dilatation or ileus.

METHODS
Abdominal radiographs of patients with a surgical, clinical, and/or laboratory diagnosis of GI diseases with GI dilatation were retrieved from our institution's Picture Archiving and Communication System archive. Additionally, abdominal radiographs performed to detect abnormalities other than GI disorders were collected to form a control group. The images were labeled with three tags according to their groups: surgically-corrected dilatation (SD), inflammatory/infectious dilatation (ID), and normal. To determine the impact of standardizing the imaging area on the model's performance, an additional dataset was created by applying an automated cropping process. Five CNN models with proven success in image analysis (ResNet50, InceptionResNetV2, Xception, EfficientNetV2L, and ConvNeXtXLarge) were trained, validated, and tested using transfer learning.

RESULTS
A total of 540 normal, 298 SD, and 314 ID were used in this study. In the differentiation between normal and abnormal images, the highest accuracy rates were achieved with ResNet50 (93.3%) and InceptionResNetV2 (90.6%) CNN models. Then, after using automated cropping preprocessing, the highest accuracy rates were achieved with ConvNeXtXLarge (96.9%), ResNet50 (95.5%), and InceptionResNetV2 (95.5%). The highest accuracy in the differentiation between SD and ID was achieved with EfficientNetV2L (94.6%).

CONCLUSION
Deep learning models can be integrated into radiographs located in the emergency departments as a decision support system with high accuracy rates in pediatric GI obstructions by immediately alerting the physicians about abnormal radiographs and possible etiologies.

CLINICAL SIGNIFICANCE
This paper describes a novel area of utilization of well-known deep learning algorithm models. Although some studies in the literature have shown the efficiency of CNN models in identifying small bowel obstruction with high accuracy for the adult population or some specific diseases, our study is unique for the pediatric population and for evaluating the requirement of surgical versus medical treatment.

KEYWORDS
Abdominal X-ray, ileus, pediatric radiology, convolutional neural networks, deep learning

You may cite this article as: Ayaz E, Güçlü H, Oktay AB. Diagnostic accuracy of convolutional neural network algorithms to distinguish gastrointestinal obstruction on conventional radiographs in a pediatric population. *Diagn Interv Radiol*. 28 February 2025 DOI: 10.4274/dir.2025.242950 [Epub Ahead of Print].

The imaging of the gastrointestinal (GI) system is challenging in children, and often, the initial modality of choice is either an abdominal radiograph or ultrasound, both in the emergency and outpatient settings. Abdominal radiography is cheap, widely available, exposes less radiation compared with computed tomography (CT), and provides specific appearances for some pediatric conditions such as duodenal atresia and necrotizing enterocolitis (NEC).[1] The common causes of GI obstructions in pediatric patients are more varied and different than in adults and often require dedicated radiological evaluation to recognize peculiar imaging features.[2] The bowel can be obstructed or dilated by a wide range of diseases classified as congenital, developmental, inflammatory, infectious, and neoplastic lesions.[3] Delay in the diagnosis and surgical management of such pediatric acute bowel obstruction increases the risk of bowel necrosis, perforation, and death. Therefore, accurate diagnostic management is crucial to improve patient outcomes.[4] Previous studies in adult populations have revealed that the 3 most sensitive radiographic signs for bowel obstruction are air-fluid levels in loops of the bowel wider than 2.5 cm, 2 or more air-fluid levels, and multiple air-fluid levels within 1 loop of the bowel differing 5 mm.[2]

In recent years, there has been a growing number of studies on integrating artificial intelligence (AI) as a diagnostic support model into image-based medical fields such as radiology and pathology. Artificial neural networks have become the most preferred models for image classification among the subfields of AI due to their high accuracy rates.[5]

Convolutional neural network (CNN), a deep artificial neural network, possesses the ability to distinguish and classify images by extracting and comparing specific features from them. However, the main limitation of CNNs is their need for large datasets for training. The capacity of a CNN trained on a large dataset can be transferred to differentiate similar images.[6] With the proliferation of digital radiography and Picture Archiving and Communication Systems (PACS), significant advancements have been made in acquiring radiographic data in recent years. Although radiography involves single-section and two-dimensional imaging, CT and magnetic resonance imaging provide multi-sectional and three-dimensional imaging. Therefore, radiographs can be processed with simpler deep-learning models.

Main points

- Pre-trained convolutional neural network models can be accurately used in abdominal radiographs with the transfer learning method.
- Fine-tuning should be performed to improve the performance of the model and to decrease the validation and training loss.
- The automated cropping process significantly improves the performance of all models, probably due to factors such as the non-standard nature of the radiographs taken under emergency and outpatient conditions, improper positioning, and inappropriate adjustment of the imaging area.

In daily practice, many abdominal radiographs are performed on children in emergency rooms and outpatient clinics. In Turkey, most of these are not evaluated by radiologists but by emergency or outpatient physicians under time constraints. According to a report prepared by the Turkish Society of Radiology instead of Radiology Association in 2018, the number of radiologists per 100,000 people in Turkey was 5, whereas this number was 2–3 times higher in Organization for Economic Co-operation and Development countries.[7] Due to the lack of sufficient time for evaluating radiographs or the inexperience of the evaluating physician, additional tests may be unnecessarily requested for patients with false-negative evaluations, or patients with a condition may be incorrectly deemed normal and sent home. Conversely, unnecessary treatments or surgical interventions may be performed on patients with false-positive evaluations. Since children often cannot accurately express their complaints and because laboratory findings can change rapidly, radiological examinations hold even greater importance.[4]

Therefore, if the radiographs taken in the emergency room are classified by a CNN model integrated into the PACS system and presented to the relevant physician, it can enable more careful evaluation by the physician.

This study aims to retrain current CNN models on abdominal radiographs and assess which models are more successful in classifying normal and pathological radiographs. It also proposes differentiating between pathological radiographs that resolve with medical treatment (infectious) and those requiring surgical intervention.

Methods

Institutional review board approval was obtained from the Diyarbakır Gazi Yaşargil Training and Research Hospital Non-Interventional Clinical Research Ethics Committee (decision no: 2022/108, decision date: 10.06.2022) for this study's retrospective data collection and analysis. Informed consent was waived because of the retrospective nature of the study.

Image acquisition

After obtaining the approval of the ethical committee, abdominal radiographs taken in the outpatient clinic and emergency department between January 1, 2019, and June 1, 2022, were reviewed using the radiology PACS archive of our institution. They were included if the patients had multiple images within the same disease course and before the surgery. X-ray devices used in the outpatient clinic and emergency department were single-tube Jumong model digital X-ray imaging systems (SG Healthcare Co, Gyeonggi-Do, South Korea). Automatic exposure control (AEC) sensors were used during imaging, and dose parameters for each imaging were adjusted accordingly. Shielding was not used to avoid overexposure due to AEC measurements. Peak tube voltage (kVp), tube current (mA), exposure time (msec), and dose area product (DAP) were recorded for each examination. Due to vast body size variations in the study cohort (0–18 years), peak tube voltage was changed between 80,100 and 120 kVp according to tissue thickness, requiring more photon penetration. For the routine posteroanterior erect abdominal radiograph performed in the outpatient clinic and emergency department, the patient-tube distance was 110 cm.

The images were retrieved for the study using JPEG compression. For comparison, a control group was formed from patients with normal GI findings on abdominal radiographs, who were imaged for other reasons, such as kidney stones, with a balanced age distribution from 0 to 18 years. The dataset consisted of three main groups: (1) patients with GI obstruction requiring surgical intervention [surgically-corrected dilatation (SD)], (2) patients with bowel dilatation/ileus treated without surgery [inflammatory/infectious dilatation (ID)], and (3) a normal control group. The age and sex characteristics of the patients and the diagnoses of the diseases for groups with pathological findings were recorded. The first group requiring surgery was diagnosed surgically. While labeling the second group, if examinations remained indeterminate, the cases were discussed by an experienced pediatric radiologist (with 7 years of experience) and the referring pedi-

atrician. Six cases that remained indeterminate after enhanced clinical-radiological review were excluded, as no meaningful label could be assigned.

Consequently, a total of 612 radiographs with the findings of bowel dilatation or obstruction were included. For the first group (patients who underwent surgery), 298 images from 107 patients were obtained from the archive, and for the second group (patients who did not require surgery), 314 images from 189 patients were obtained. For comparison, a control group of 540 normal abdominal radiographs, 1 for each case, was created, considering a balanced age distribution between 0 and 18 years. The flowchart of the study is presented in Figure 1.

Training, testing dataset, and preprocessing

Images were retrieved from the PACS station with a resolution of 1,040 × 624 pixels and down-sampled by bicubic interpolation automatically in the CNN to match the input layers. Afterward, 32 batches, each including 36 images, were composed of 1,152 images. Each batch was split into training, validation, and test sets using a ratio of 28:3:5, respectively. This ratio was designed to maintain a sufficient training dataset while providing adequate statistical power for the testing dataset. A test set sample size of 160 enabled a statistical power of 0.8 for detecting an area under the curve (AUC) of 0.65 with a type 1 error of 0.025.[8] Data augmentation was performed on the training dataset with horizontal flipping and rotation by Keras library. The images formed with data augmentation would be similar to those not taken in the correct position due to patient rotation during the shooting or sent incorrectly to the PACS system. This approach aims to provide flexibility for the model to evaluate images that are not properly positioned (Supplementary Figure 1).

To determine the impact of standardizing the imaging area on the model's performance, an additional dataset was created by applying an automated cropping process to the data using a cropping code set to remove rows or columns from all edges until a white-toned pixel was found. During the automated cropping process, some images had data labels on the edges of the image, causing the cropping to stop before the model reached the image (Figure 2a-d). This situation represented a limitation of the model compared with manual cropping. Since this study aimed to provide the classification result directly to the physician via automated preprocessing and model analysis of the image obtained from X-ray imaging, manual cropping was not preferred.

Neural networks training and testing

All CNN training, testing, and other processes were performed using the Keras 2.1.5 library with TensorFlow 1.7 as the backend framework in Python (version 3.7.3), and Google Colab was used as a notebook service provider with its integrated graphics processing units.[9-11]

Five CNNs used in this study were publicly available and pre-trained on the ImageNet data set: ResNet50, InceptionResNetV2, Xception, EfficientNet, and ConvNeXt.[12-17] The architecture of the models is briefly described in Supplementary Figure 2. The background of the networks was developed to detect everyday objects such as vehicles, flowers, or animals, but the top layers were completely new and acquired their parameters based on the radiographs used in the training database, called the transfer learning method. Since the datasets on which CNN models were pre-trained contained a large amount of data, during training with our smaller dataset, the process of determining the filter weights for feature extraction was limited to the first few convolutional layers (usually the first three), and the training of the last layers did not occur.[18] To overcome this, a particular process called fine-tuning was applied, and the layers close to the input

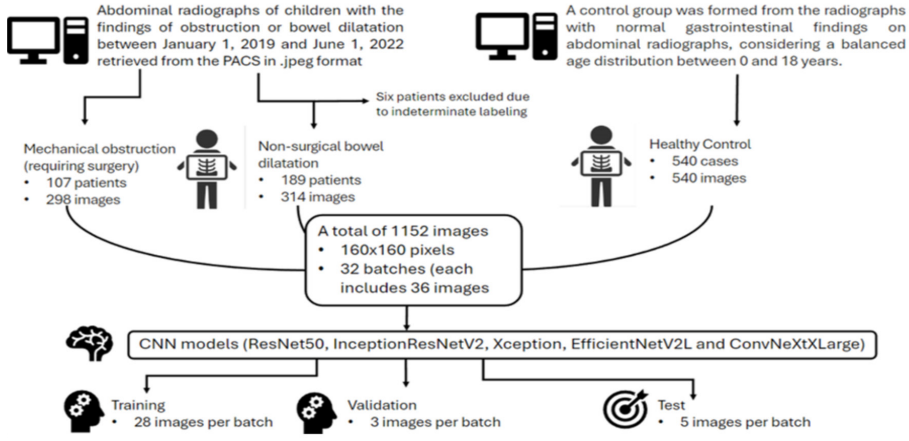


Figure 1. Flow diagram of the study. CNN, convolutional neural network; PACS: Picture Archiving and Communication Systems.

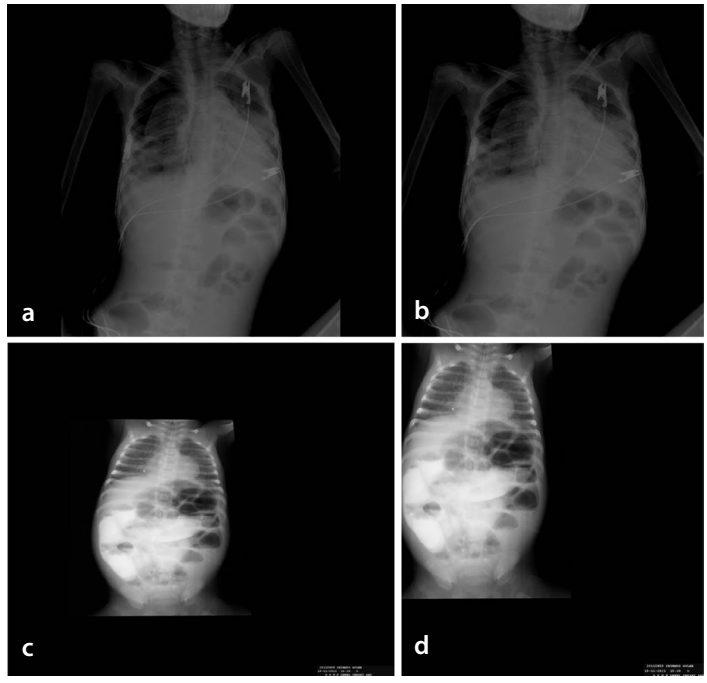


Figure 2. Example of before **(a)** and after **(b)** successful automated cropping to remove unnecessary parts and suboptimal cropping **(c, d)** due to white pixels of a label at the right lower corner of the image.

were frozen to ensure that the weights for feature extraction in the subsequent layers were determined. Fine-tuning was routinely applied, especially in transfer learning methods used in image analysis, and it improved the model's performance. Although epoch durations were longer compared with the standard training process, a significant decrease in training and validation errors was achieved with fewer epochs (Supplementary Figure 3).

The five models used in the experiments were trained for 100 epochs during the training phase. To enhance performance during transfer learning and to allow the models trained on another dataset to adapt to the features of our data, an additional 20 epochs were run for fine-tuning. In this study, the models were first presented with original data and then cropped data from the normal control and abnormal patient groups. All models were tested on 224 images after training, and their success was evaluated using performance metrics. Finally, to determine which diseases, ages, and sexes the misdiagnosed cases (false positives or false negatives) belonged to, the dataset was analyzed using the most successful model.

Statistical analysis

The mean and standard deviation values for age and the median and quartiles were presented. The descriptive statistics of the pathological groups were calculated. Pearson's chi-squared test was used to compare gender data. Kolmogorov–Smirnov test showed that the age data did not follow a normal distribution ($P < 0.001$ for all groups). The median and interquartile ranges were presented for non-normally distributed dose parameters. A non-parametric Mann–Whitney U test was applied since the pathological groups did not show a normal distribution. Statistical analyses were performed using the IBM SPSS version 23.0 software package (IBM Corporation, Armonk, NY, USA). A single receiver operating characteristic (ROC) curve and cut-off analysis were used for the internal test, whereas two ROC curves with independent groups were designed to compare the external and internal validation tests. Two-tailed P values < 0.05 were considered statistically significant. After completing the training phase, the models were tested using the dataset created for testing. The performance of the models was measured by metrics such as accuracy, precision, sensitivity, specificity, F1 score, and the AUC.

Results

The age and sex distribution of the three groups within the dataset are presented in Table 1. No significant difference was found between the two patient groups regarding age ($P = 0.928$). However, there were more boys in the SD group than in the ID group ($P < 0.001$). Regarding dose parameters, 725 examinations were performed with 80 kVp, 338 with 100 kVp, and 89 with 120 kVp. The median tube current was 320 mA (interquartile range 80). The mean (± standard deviation) exposure time was 37.22 ± 7.51 milliseconds, and the median DAP was 165 mGy·cm^2 (interquartile range: 349). In the SD group, a total of 16 different causes of obstruction were identified. The most prevalent cause, ileus due to postoperative adhesions, was observed in 83 radiographs of 27 patients (27.9%). This was followed by complicated appendicitis, seen in 67 radiographs of 30 patients (22.5%), and NEC, found in 35 radiographs of 11 patients (11.7%). It is worth noting that some cases of ileus due to postoperative adhesions were observed during follow-up after surgeries of patients with other etiologies, which is why the total number of cases appears higher than the total number of patients in this group when the cases from both groups are combined. The age and sex distribution according to the types of diseases is presented in Table 2.

NEC, hypertrophic pyloric stenosis, meconium ileus, Hirschsprung's disease, duodenal atresia/stenosis, and inguinal hernia cases are observed in the neonatal and infant periods, whereas abscess/peritonitis secondary to intraperitoneal catheter and intussusception cases occur in early childhood. Complicated appendicitis and Crohn's disease are predominantly seen in the group aged over 10 years. The disease groups with the broadest age distribution are also the two most common diseases: ileus due to postoperative adhesions and complicated appendicitis. Among the common diseases, groups with similar ages were compared statistically using the Student's t-test. The ages of patients

Table 1. Age and sex distribution of the study groups and the control group

	Healthy control group	SD group	ID group
Sex [male (%)/female (%)]	262 (48.5)/278 (51.5)	232 (77.5)/66 (22.1)	180 (57.3)/134 (42.7)
Age (mean ± standard deviation), years	7.29 ± 5.05	5.47 ± 5.82	4.22 ± 4.44
Age [median (interquartile ranges)], years	6.5 (3.1–11.3)	3 (0.3–10.0)	2.1 (1.3–6.0)

SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

Table 2. Number, age, and sex features of patients within the surgically corrected obstruction group

Diagnosis	Number of cases/ number of images	Sex: male (%)/ female (%)	Age: mean ± standard deviation in years
Ileus due to postoperative adhesion	27/83	22 (81.5)/5 (18.5)	6.48 ± 5.32
Complicated acute appendicitis	30/67	22 (73.3)/8 (26.7)	11.33 ± 4.75
Necrotizing enterocolitis	11/35	5(45.5)/6 (54.5)	0.39 ± 0.53
Hypertrophic pyloric stenosis	12/19	12 (100)/0	0.11 ± 0.06
Hirschsprung's disease	7/15	6 (85.7)/1 (14.3)	1.02 ± 1.38
Abscess/peritonitis secondary to intraabdominal catheter	5/15	4 (80)/1 (20)	5.13 ± 6.08
Meconium ileus or meconium plug syndrome	5/12	4 (80)/1 (20)	0.34 ± 0.27
Duodenal atresia or stenosis	2/11	1 (50)/1 (50)	1.15 ± 0.10
Intussusception	9/9	5 (55.6)/4 (44.4)	3.18 ± 4.41
Complicated inguinal hernia	4/8	4 (100)/0	0.95 ± 0.78
Complicated Crohn's disease	2/8	2 (100)/0	11.89 ± 1.48
Midgut volvulus	3/6	3 (100)/0	1.80 ± 2.26
Other	4/9	4 (100)/0	6.87 ± 4.79

with complicated appendicitis were found to be significantly higher than those with ileus due to postoperative adhesions ($P < 0.001$), and the ages of patients with NEC were significantly higher than those with hypertrophic pyloric stenosis ($P = 0.003$). No significant difference was found between cases of postoperative adhesions and catheter infections ($P = 0.379$) or between Hirschsprung's disease and duodenal atresia/stenosis ($P = 0.719$).

In the third group, which included cases of non-ileus, no infectious agent was detected in 142 patients, from whom 231 (73.6%) radiographs were obtained. In 41 patients (68 radiographs, 21.7%), rotavirus was detected in 2 patients (3 radiographs, 1%), adenovirus antigen in 2 patients (6 radiographs, 1.9%), and amoeba in the stool of 2 patients (6 radiographs, 1.9%). In 2 patients with 6 radiographs (1.9%), GI involvement due to multisystem inflammatory syndrome secondary to coronavirus disease-2019 was diagnosed. When comparing the ages of the rotavirus cases and other cases, it was found that rotavirus cases were significantly higher in the younger age groups ($P < 0.001$).

All models were tested separately on 224 images using both the original and cropped datasets after training. The confusion matrices of the models are presented in Table 3, and the performance metrics are presented in Table 4. Although the highest performance metrics of the models were observed in different rankings across different datasets, they were generally achieved with ResNet50 on raw data, EfficientNetV2L between pathological groups, and ConvNeXtXLarge on cropped data and overall. The pixels with the most weight in classification, as indicated by the gradient-weighted class activation map applied to some data in the analysis of the Xception model, are shown in Figure 3.

Finally, to determine which diseases and ages the misclassified cases (false positives or false negatives) belonged to, our dataset was analyzed using the ConvNeXtXLarge model, which had the highest F1 score. The model was run 3 times using 224 images randomly distributed across 7 packages in each analysis. Four images with SD and six with ID were labeled as false negatives in the three-model analyses. In the normal control group, 33 images were classified as false positives across the 3 analyses. Examples of patients who were classified as abnormal but were healthy, according to the model analysis, are presented with their ages and sexes in Supplementary Figure 4. The cases labeled as normal despite being in the SD group are presented in Figure 4. Since the false negative cases occurred in three different disease groups and involved common diseases in the dataset, we could not conclude that a specific disease group was undetectable.

Discussion

Very few studies utilize deep learning applications on abdominal radiographs, and there is even less literature regarding the pediatric population.[4,19-21] Studies on X-rays in the literature primarily focus on chest radiography, mainly due to the large volume of accessible data.[22-25] Our study demonstrated that in classifying normal and abnormal radiographs, an accuracy above 90% was achieved with the ResNet50 (93.3%) and InceptionResNetV2 (90.6%) CNN models. After applying the cropping preprocessing step to the same data groups, an accuracy above 90% was achieved with EfficientNetV2L (94.6%), and an accuracy above 95% was reached with ResNet50 (95.5%), InceptionResNetV2 (95.5%), and ConvNeXtXLarge (96.9%). In the analysis conducted on cropped images to distinguish surgically corrected GI obstruction from other GI dilations, an accuracy above 90% was achieved with InceptionResNetV2 (90.2%), EfficientNetV2L (94.6%), and ConvNeXtXLarge (91.1%). It is evident that the cropping preprocessing step significantly impacts the performance

Table 3. Confusion matrices of the convolutional neural networks' test results used in the study

CNN model	Data type	Labels		Actual	
				Normal (or SD group)	Abnormal (or ID group)
Classification results with **ResNet50** CNN model	Raw images	**Predicted**	Normal	109	15
			Abnormal	0	100
	Cropped images		Normal	109	7
			Abnormal	3	105
	Cropped images		SD group	117	4
			ID group	21	82
Classification results with **InceptionResNetV2** CNN model	Raw images	**Predicted**	Normal	103	3
			Abnormal	18	100
	Cropped images		Normal	119	1
			Abnormal	9	95
	Cropped images		SD group	106	6
			ID group	16	96
Classification results with **Xception** CNN model	Raw images	**Predicted**	Normal	120	10
			Abnormal	28	66
	Cropped images		Normal	100	17
			Abnormal	13	94
	Cropped images		SD group	104	21
			ID group	7	92
Classification results with **EfficientNetV2L** CNN model	Raw images	**Predicted**	Normal	84	28
			Abnormal	0	112
	Cropped images		Normal	118	0
			Abnormal	12	94
	Cropped images		SD group	108	5
			ID group	7	104
Classification results with **ConvNeXtXLarge** CNN model	Raw images	**Predicted**	Normal	102	6
			Abnormal	17	99
	Cropped images		Normal	121	3
			Abnormal	4	96
	Cropped images		SD group	107	7
			ID group	13	97

CNN, convolutional neural network; SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

of all models. This improvement is likely due to factors such as the non-standard nature of radiographs taken under emergency and outpatient conditions, improper positioning, inappropriate adjustment of the imaging area, and the failure to remove contrast-inducing items from patients during imaging.

Abdominal radiographs are generally the first preferred method for GI diseases due to their affordability, widespread availability, rapid application and interpretation (especially with digital radiographs), and ability to comprehensively show intestinal gas distribution. Radiography is superior to ultrasound, particularly for diagnosing GI obstructions.[26] Typical imaging findings are observed in diseases such as NEC and duodenal atresia, which are seen in the neonatal and infant periods. Additionally, in patients with acute severe clinical symptoms where bowel perforation (rupture) is suspected, radiographs can reveal free air in the abdominal cavity. However, the sensitivity of abdominal radiographs in children with abdominal pain is relatively low, with the rate of pathological findings reported between 2% and 20%.[26] Abdominal radiographs in newborns and young children are usually taken while the patient is lying down. In older children, an upright abdominal radiograph may better display air-fluid levels and bowel loop distention, especially in conditions where peristalsis is impaired. In some cases, lateral decubitus radiographs are taken by positioning the patient on their side to show air-fluid levels, free fluid, or free air in the abdomen.

The following studies stood out when reviewing previous deep-learning research in the literature on diagnosing GI diseases using abdominal radiographs. In the study by Kwon et al.[21], 11,384 abdominal radiographs (1,449 with intussusception and 9,935 without) were retrieved from three hospitals to detect intussusception. Diagnosing intussusception from abdominal radiographs is challenging and requires expertise. Therefore, the diagnosis is typically made by ultrasound. The interobserver agreement among radiologists with limited experience in abdominal radiographs is less than 50%.[27] In the study by Kwon et al.[21], for binary classification, the CNN model used was ResNet. The average sensitivity achieved was 81.6%, with a specificity of 92.5%. The highest accuracy reported was 76%, the lowest was 73%, and the average was 74%. In our study, an analysis of the SD cases classified as false negatives revealed that two of the four cases were complicated appendicitis, one was bowel obstruction (ileus due to postoperative adhesions), and one was Hirschsprung's disease. Notably, no misclassification was detected in intussusception cases. Additionally, an accuracy rate of 93.3% was achieved with the ResNet50 model in our study, making it the model with the highest accuracy on raw data.

In another study on small bowel obstruction, a total of 3,663 upright abdominal ra-

Table 4. Performance metrics of the convolutional neural network models according to datasets

CNN model	Dataset	Accuracy	Specificity	Sensitivity	F1 score
ResNet50	Normal vs. abnormal (raw data)	0.933	1.000	0.869	0.930
	Normal vs. abnormal (cropped data)	0.955	0.973	0.938	0.955
	SD vs. ID group	0.889	0.848	0.953	0.868
InceptionResNetV2	Normal vs. abnormal (raw data)	0.906	0.851	0.970	0.905
	Normal vs. abnormal (cropped data)	0.955	0.930	0.990	0.950
	SD vs. ID group	0.902	0.869	0.941	0.897
Xception	Normal vs. abnormal (raw data)	0.839	0.811	0.868	0.776
	Normal vs. abnormal (cropped data)	0.866	0.885	0.847	0.862
	SD vs. ID group	0.875	0.937	0.814	0.868
EfficientNetV2L	Normal vs. abnormal (raw data)	0.875	1.000	0.800	0.889
	Normal vs. abnormal (cropped data)	0.946	0.908	1.000	0.940
	SD vs. ID group	0.946	0.939	0.954	0.945
ConvNeXtXLarge	Normal vs. abnormal (raw data)	0.897	0.857	0.943	0.896
	Normal vs. abnormal (cropped data)	0.969	0.968	0.970	0.965
	SD vs. ID group	0.911	0.892	0.933	0.907

CNN, convolutional neural network; SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

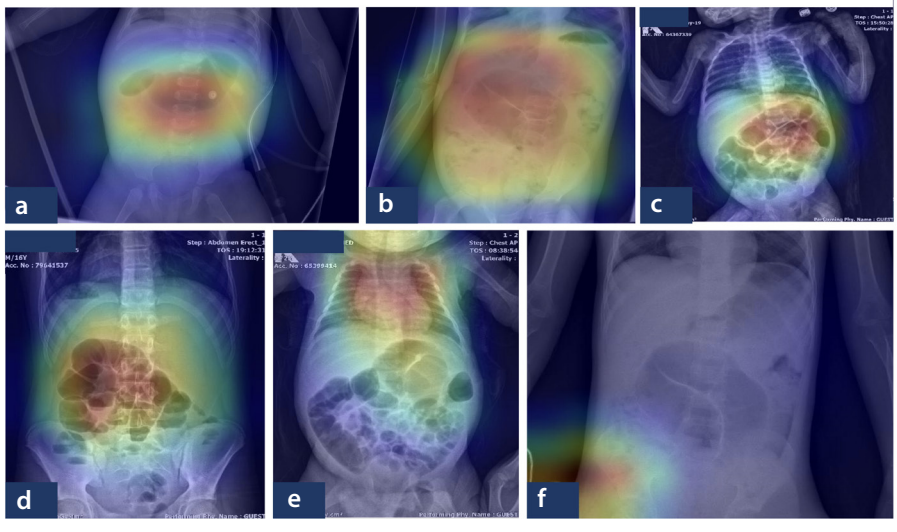


Figure 3. In the gradient-weighted class activation map (Grad-CAM) heatmap, the location of the findings was correctly identified for patients with gastrointestinal dilatation requiring surgery with diagnoses of duodenal atresia **(a)**, midgut volvulus **(b)**, meconium ileus **(c)**, and perforated appendicitis **(d)**. However, in two patients diagnosed with intestinal malrotation/midgut volvulus **(e, f)**, the weight of the Grad-CAM heatmap was incorrectly identified.

diographs (2,210 for training and 1,453 for testing) were used, with 74 showing signs of obstruction.[19] In this study, the pre-trained InceptionV3 CNN model was fine-tuned using the transfer learning method, trained with their dataset, and then tested. The AUC was calculated as 0.84, the sensitivity as 83.8%, and the specificity as 68.1%.

In a subsequent study conducted by the same team, a new dataset consisting of 5,558 radiographs was created using images obtained from their hospital and a second hospital.[20] The average age of the patients in this dataset was 59.1 and 59.9 years, which differed significantly from the causes of obstruction in our patient group. Again, using InceptionV3, the researchers trained and tested the model with the second dataset.

For comparison, 1,453 test images were independently evaluated by three radiologists. The sensitivity of the radiologists ranged from 28.5% to 65.5% (average 44%), whereas the CNN model achieved 82.9%. The specificity of the radiologists ranged from 96.4% to 99.6% (average 98.4%), whereas the CNN model achieved 92.5%. The radiologists' positive predictive value (PPV) ranged from 43% to 78% (average 62%), whereas the CNN model's PPV was 28%. The low PPV in the CNN model was due to a high number of false positives.

Upon examining these false positives, it was found that while the intestinal segments were within physiological limits and considered normal clinically and radiologically, the CNN model identified them as positive. Increasing the number of similar images in the training set could potentially improve the model's performance and address this issue.

In another UK-based study on the same subject, a dataset of abdominal radiographs (445 normal and 445 with GI obstructions) from 990 adult patients was classified using transfer learning and ensemble modeling with five pre-trained CNN models: VGG16, DenseNet121, NasNetLarge, InceptionV3, and Xception.[4] Of the dataset, 800 images were used for training, 80 for validation, and 110 for testing. Among the 110 test images, there were 5 false negatives and 4 false positives. Among the models, DenseNet121 was trained using CheXNet, which consisted of chest radiograph images, whereas the other models were trained with ImageNet. The validation loss rate of the DenseNet121 model was significantly lower than that of the other models, at 43%. In previous studies where CNN models were applied to abdominal radiographs, the highest accuracy rate

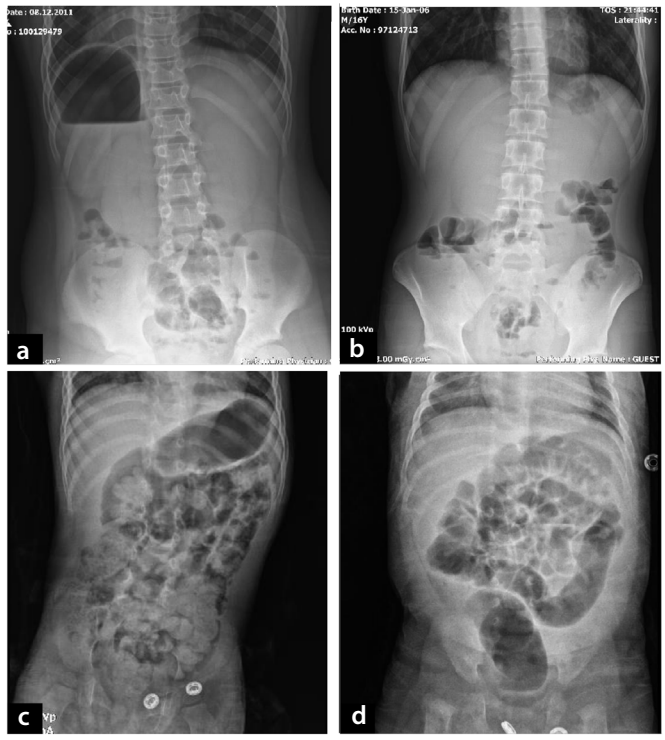


Figure 4. Surgical diagnosis, age, sex, and radiographs of abnormal cases classified as normal (false negatives) when tested with ConvNeXtXLarge are shown. The name labels on images were manually cropped before presenting in the figure. **(a)** An 11-year-old boy with situs inversus and perforated appendicitis; **(b)** a 16-year-old-boy with perforated appendicitis; **(c)** a 2-year-old boy with postoperative adhesions and Ladd band excision; **(d)** a 2-month-old boy with Hirschsprung's disease.

achieved was 92%. Although similar or slightly better performance metrics were achieved in our study, ours is the first to reach these levels in a pediatric patient group. Additionally, upon examining the image samples from the aforementioned study, it is evident that the images were standardized in size and cropped to include only the abdomen. In our study, automatic cropping was applied, but the cropping process only sometimes achieved the desired level in every dataset. This may have caused a decrease in performance metrics. The performance metrics of our study and the aforementioned studies are presented in the Supplementary Table 1.

In all three test runs of the model on our dataset, false-positive results were more frequent than false negatives. At first glance, this could potentially lead to unnecessary surgical or medical treatment. However, since patients with positive results will also be evaluated through laboratory data, clinical examinations, and symptoms, the likelihood of unnecessary surgery due to false positives is very low. It could, however, result in a loss of time and resources due to additional tests and examinations. However, false-negative cases are more dangerous, as they could lead to the oversight of positive cases in the busy working environment of emergency rooms or outpatient clinics. In the model analysis, false negatives were about one-third as frequent as false positives, with 60% of these being patients within the ID group. The false-negative rate was relatively low for more critical SD cases. When examining sensitivity, the performance metric most affected by false-negative data, the sensitivity in the InceptionResNetV2, EfficientNetV2L, and ConvNeXtXLarge models was above 95%.

The main limitation of the study is the small sample size. In CNN models, the amount of data is one of the most important factors for performance improvement. For radiographic studies, there are open-access chest radiograph datasets provided by different institutions, with the number of images approaching 225,000.[28] However, to our knowledge, no such dataset currently exists for abdominal radiographs. In children, radiographs are used far less frequently than in adults due to the potential harm of ionizing radiation. Therefore, multicenter studies are needed to reach sufficient sample sizes. To mitigate this limitation, data augmentation was applied during the training phase. However, data augmentation could result in higher performance metrics than what might be achieved in practical applications.

The SD group in the study included 16 different etiologies, and since the number of cases for each disease was too small when evaluated individually, performance metrics for specific disease groups could not be assessed separately. Another limitation of our study is that some patients had multiple radiographs taken on different days during their illness, and radiographs taken during follow-up after a diagnosis was made were also included in the study to increase the sample size. As the diagnostic process progresses, signs of GI obstruction become more pronounced in radiographs taken later. Therefore, if only radiographs from the initial visit had been used, performance metrics might have been lower.

When creating the control dataset, the aim was to include images representing all age groups between 0 and 18 years to ensure balanced representation during model training. However, patients with abnormal findings were mostly infants and young children. As a result, the average age of the control group (7.29 ± 5.05 years) was higher than that of the patient groups (SD: 5.47 ± 5.82 and ID: 4.22 ± 4.44 years). It is generally expected that there should be no significant difference in the age and sex distribution between the study and control groups, which may have introduced bias in our study. However, we intentionally chose to create a balanced control group for ages 0–18, as we believe our model can be applied across all stages of childhood. In the future, if large open-access datasets are made available, it would be beneficial to use age filters when selecting data for such studies.

In conclusion, this study has verified that training with transfer learning can be used in deep learning to identify GI obstruction in children with high accuracy. The appropriate preprocessing steps and fine-tuning significantly improve the performance of all models. Although there are inconsistent features in the heat map of some correctly labeled cases, these models can also be useful for depicting the location of obstruction requiring surgery and for monitoring dilatation requiring medical treatment.

Acknowledgments

Sample codes used in the study for binary classification can be accessed at https://github.com/RadDrEAyaz/Convolutional-Neural-Netwok-Usage-on-Abdominal-Radiographs-for-the-Diagnosis-of-Gastrointestinal-Obs.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Kandasamy D, Sharma R, Gupta AK. Bowel imaging in children: part 1. *Indian J Pediatr*. 2019;86(9):805-816. [Crossref]
2. Hryhorczuk AL, Lee EY. Imaging evaluation of bowel obstruction in children: updates in imaging techniques and review of imaging findings. *Semin Roentgenol*. 2012;47(2):159-170. [Crossref]
3. Kandasamy D, Sharma R, Gupta AK. Bowel imaging in children: part 2. *Indian J Pediatr*. 2019;86:817-829. [Crossref]
4. Kim DH, Wit H, Thurston M, et al. An artificial intelligence deep learning model for identification of small bowel obstruction on plain abdominal radiographs. *Br J Radiol*. 2021;94(1122):20201407. [Crossref]
5. Atlan F, Pençe İ. An overview of artificial intelligence and medical imaging technologies. *ACIN*. 2021;5(1):207-230. [Crossref]
6. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Arvix*. 2014;27:3320-3328. [Crossref]
7. Turk Radyoloji Derneği. Radyolojik tetkik yoğunlugu. Published in January 2018. Accessed in 08.09.2024. [Crossref]
8. Goksuluk D, Korkmaz S, Zararsiz G, Karaagaoglu AE. easyROC: an interactive web-tool for ROC curve analysis using R language environment. *The R Journal*. 2016;8:213-230. [Crossref]
9. F. Chollet, Keras: the python deep learning library. 2015. [Crossref]
10. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. 2015. [Crossref]
11. Bisong, E. Google colaboratory. In: building machine learning and deep learning models on Google Cloud Platform. Apress, Berkeley, CA. 2019;59-64. [Crossref]
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: 2016:770-778. [Crossref]
13. Szegedy C, Liu W, Jia Y, et al. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: 2015:1-9. [Crossref]
14. Chollet F. Xception: deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017:1251-1258. [Crossref]
15. Tan M, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks. In: *International conference on machine learning*. 2019;6105-6114. [Crossref]
16. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022;11976-11986. [Crossref]
17. Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *CVPR09*. 2009. [Crossref]
18. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging*. 2016;35(5):1299-1312. [Crossref]
19. Cheng PM, Tejura TK, Tran KN, Whang G. Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol (NY)*. 2018;43(5):1120-1127. [Crossref]
20. Cheng PM, Tran KN, Whang G, Tejura TK. Refining convolutional neural network detection of small-bowel obstruction in conventional radiography. *AJR Am J Roentgenol*. 2019;212(2):342-350. [Crossref]
21. Kwon G, Ryu J, Oh J, et al. Deep learning algorithms for detecting and visualising intussusception on plain abdominal radiography in children: a retrospective multicenter study. *Sci Rep*. 2020;10(1):17582. [Crossref]
22. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal Appl*. 2021;24(3):1207-1220. [Crossref]
23. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med*. 2018;15(11):e1002697. [Crossref]
24. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med Hypotheses*. 2020;140:109761. [Crossref]
25. Almutairi TM, Ismail MMB, Bchir O. X-ray based COVID-19 classification using lightweight EfficientNet. *J Artif Intell*. 2022;4(3):167-187. [Crossref]
26. Rothrock SG, Green SM, Hummel CB. Plain abdominal radiography in the detection of major disease in children: a prospective analysis. *Ann Emerg Med*. 1992;21(12):1423-1429. [Crossref]
27. Carroll AG, Kavanagh RG, Ni Leidhin C, Cullinan NM, Lavelle LP, Malone DE. Comparative effectiveness of imaging modalities for the diagnosis and treatment of intussusception: a critically appraised topic. *Acad Radiol*. 2017;24(5):521-529. [Crossref]

28. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33:590-597. [Crossref]

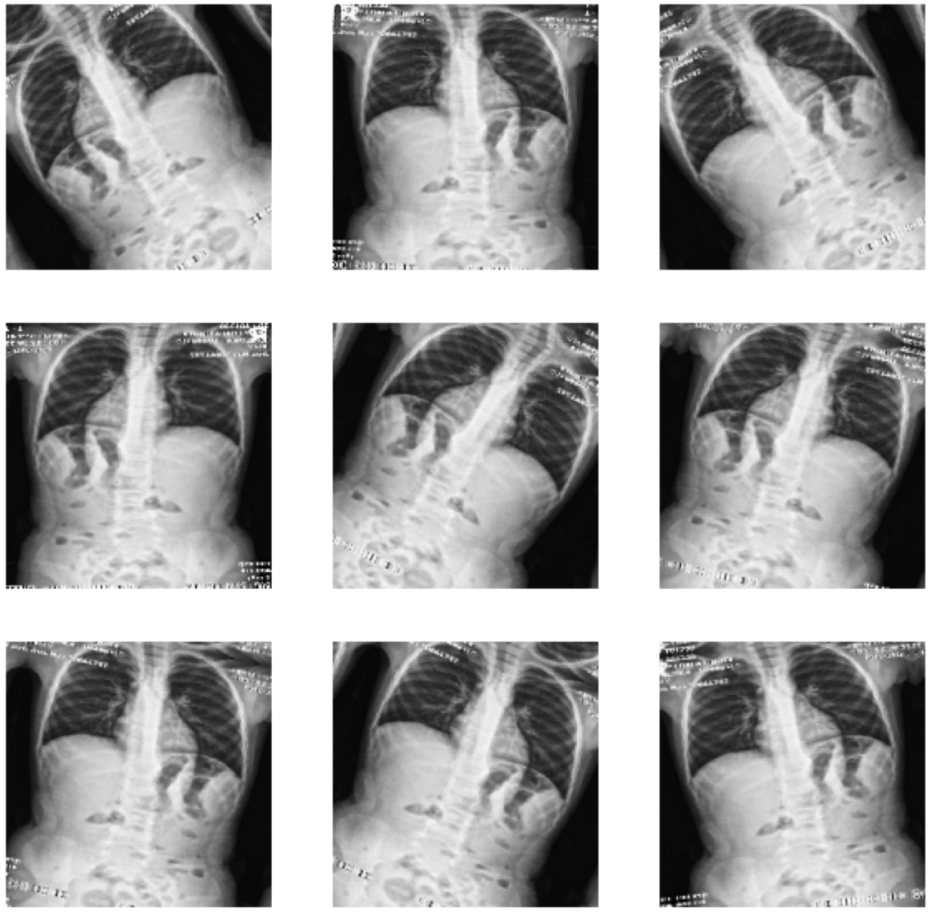
29. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017;31(1):4278-4284. [Crossref]

30. Mehmood A. Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks. *IEEE Access*. 2021;9:138283-138295. [Crossref]

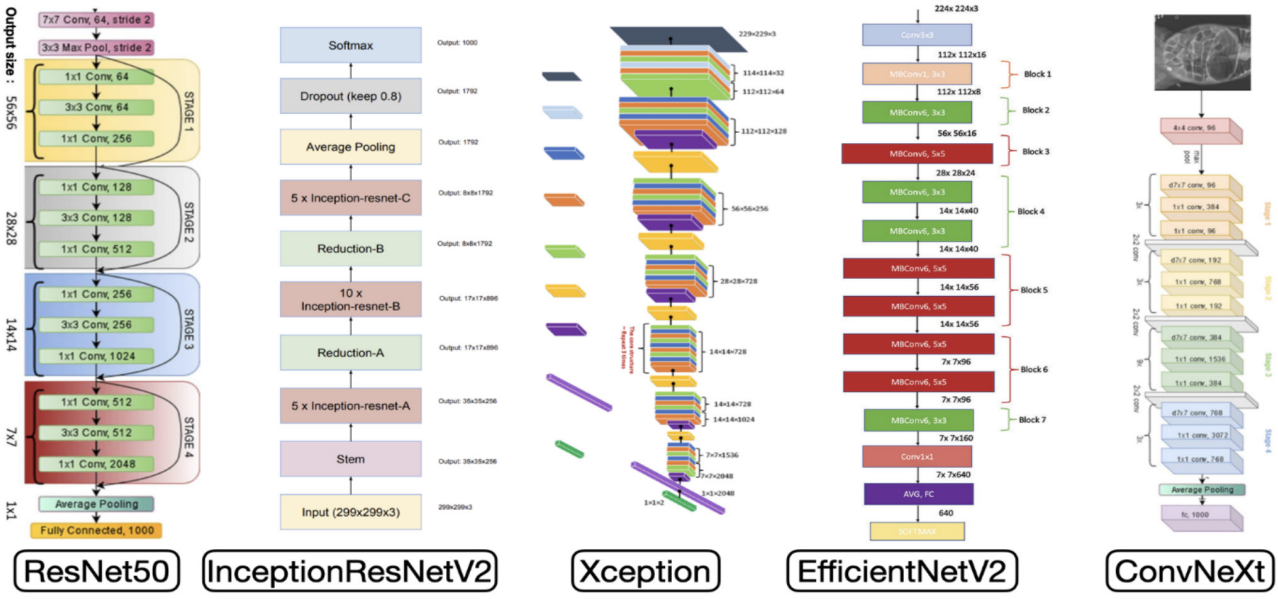
Supplementary Table 1. The performance metrics of the previous studies and the current study on various abdominal X-ray datasets

	Number of images (study/control)	Accuracy	Specificity	Sensitivity
Kwon et al.[21], 2020	11,384 (1,449/9,935)	0.760	0.250	0.816
Cheng et al.[19], 2018	3,663 (74/3,589)	0.685	0.681	0.831
Cheng et al.[20], 2019	5,558 (462/5,096)	N/A	0.925	0.829
Kim et al.[4], 2021	990 (445/445)	0.918	0.927	0.909
Current study (ConvNeXtXLarge), 2024	1,152 (612/540)	0.969	0.968	0.970

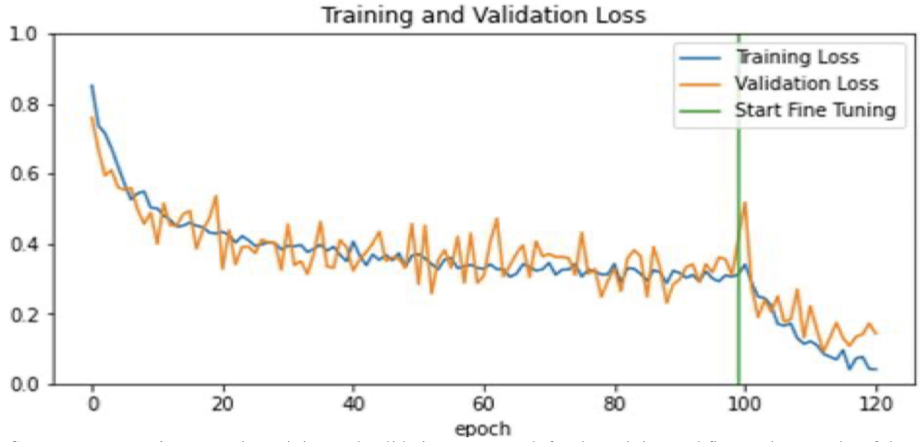
Superscripts indicate the reference number in the main manuscript.



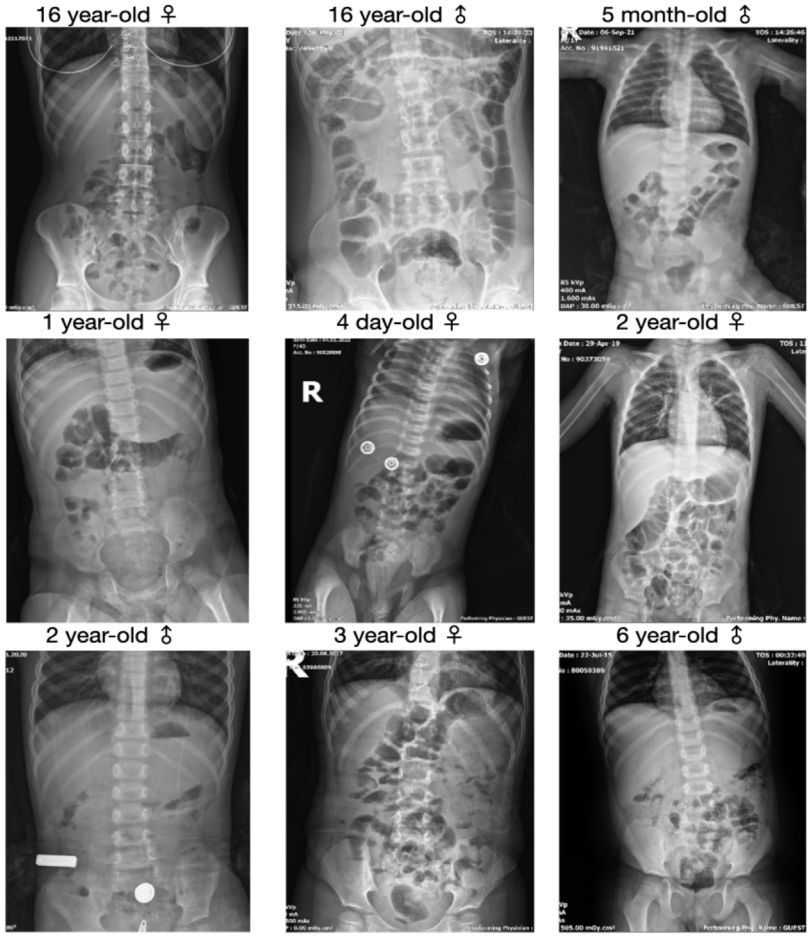
Supplementary Figure 1. Example of the horizontal flipping and rotation performed to an image during data augmentation preprocessing step.



Supplementary Figure 2. Architectural schema of the convolutional neural network models used in this study.[12,16,25,29,30]



Supplementary Figure 3. The training and validation error graph for the training and fine-tuning epochs of the dataset on the ResNet50 model is presented. The epochs after the vertical green line belong to the fine-tuning process, during which the error reduction is much more noticeable.



Supplementary Figure 4. Examples of cases with their age and sex classified as abnormal (false positives) when tested with ConvNeXtXLarge despite being healthy. Name labels on images were manually cropped before presenting in the figure.