scientific reports

OPEN

Development and validation of automated three-dimensional convolutional neural network model for acute appendicitis diagnosis

Minsung Kim[1,7], Taeyong Park[2,7], Jaewoong Kang[2], Min-Jeong Kim[3], Mi Jung Kwon[4], Bo Young Oh[1], Jong Wan Kim[5], Sangook Ha[6], Won Seok Yang[6], Bum-Joo Cho[2] & Iltae Son[1]

Rapid, accurate preoperative imaging diagnostics of appendicitis are critical in surgical decisions of emergency care. This study developed a fully automated diagnostic framework using a 3D convolutional neural network (CNN) to identify appendicitis and clinical information from patients with abdominal pain, including contrast-enhanced abdominopelvic computed tomography images. A deep learning model—Information of Appendix (IA)—was developed, and the volume of interest (VOI) region corresponding to the anatomical location of the appendix was automatically extracted. It was analysed using a two-stage binary algorithm with transfer learning. The algorithm predicted three categories: non-, simple, and complicated appendicitis. The 3D-CNN architecture incorporated ResNet, DenseNet, and EfficientNet. The IA model utilising DenseNet169 demonstrated 79.5% accuracy (76.4–82.6%), 70.1% sensitivity (64.7–75.0%), 87.6% specificity (83.7–90.7%), and an area under the curve (AUC) of 0.865 (0.862–0.867), with a negative appendectomy rate of 12.4% in stage 1 classification identifying non-appendicitis versus. appendicitis. In stage 2, the IA model exhibited 76.1% accuracy (70.3–81.9%), 82.6% sensitivity (62.9–90.9%), 74.2% specificity (67.0–80.3%), and an AUC of 0.827 (0.820–0.833), differentiating simple and complicated appendicitis. This IA model can provide physicians with reliable diagnostic information on appendicitis with generality and reproducibility within the VOI.

Keywords Artificial intelligence, Acute appendicitis, Convolutional neural network

The inconspicuous structure and variable position of the appendix, along with conditions mimicking appendicitis, create significant challenges in accurately diagnosing appendicitis via radiological imaging[1–5]. Consequently, delays or misdiagnoses impose substantial burdens on medical systems and clinicians, particularly in managing complicated appendicitis—a common, life-threatening abdominal emergency[6–15]. The increasing use of computed tomography (CT) imaging as the preferred diagnostic tool for suspected appendicitis under national health insurance systems[16–18] has led to a disproportionate increase in the demand for image interpretation, resulting in increased workloads and potential burnout among radiologists[19,20].

With recent trends toward a fully automated pipeline from data preprocessing to model training, several studies have employed machine learning (ML) and deep learning (DL) methods for diagnosing appendicitis by proposing standard automated diagnostic systems without human intervention[21]. As alternative methods, these systems offer generalisability and reliability, potentially substituting for the roles of radiologists[3,22–27]. The application of transfer learning (TL) has addressed problems of expert-annotated data scarcity by leveraging

[1]Department of Surgery, Hallym University Medical Center, Hallym Sacred Heart Hospital, Hallym University College of Medicine, 22 Gwanpyeong-ro 170 beon-gil, Pyeongan-dong, Dongan-gu, Anyang, Gyeonggi-do, Republic of Korea. [2]Medical Artificial Intelligence Center, Hallym University Medical Center, Anyang, Republic of Korea. [3]Department of Radiology, Hallym Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Republic of Korea. [4]Department of Pathology, Hallym Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Republic of Korea. [5]Department of Surgery, Dongtan Sacred Heart Hospital, Hallym University College of Medicine, Hwaseong, Republic of Korea. [6]Department of Emergency Medicine, Hallym University Sacred Heart Hospital, Hallym University Medical Center, Anyang, Republic of Korea. [7]Minsung Kim and Taeyong Park contributed equally as the first authors. email: bjcho8@gmail.com; 1tae99@hanmail.net

| https://doi.org/10.1038/s41598-024-84348-6

knowledge from source tasks to achieve high performance in target tasks[28,29]. A critical factor for the success of DL models in these applications is the collection of large datasets as control data based on standard references of normal structures and precise ground-truth labels that correlate with treatment and pathology[30,31]. Moreover, enhanced performance in interpreting cross-sectional medical images requires three-dimensional (3D) architectures reconstructed from multiple slices[16,32]. However, the application of ML using 3D reconstructed CT images for the spectrum of appendicitis, ranging from normal to complicated appendicitis, is yet to be reported.

Therefore, in this study, we hypothesised that a DL model, 'Information of Appendix (IA)', based on the two-stage binary classification algorithm using TL, similar to a clinician's approach for decision treatment, could be useful for patients visiting the emergency room (ER) with acute right or lower quadrant abdominal pain. This study aimed to develop and validate a fully automated diagnostic framework using a 3D convolutional neural network (CNN) that could identify non-, simple, and complicated appendicitis from high-quality ground-truth-labelled data from a large population. Finally, the best-performing model on an external validation dataset serves as a proof of concept of the IA model[33].

Methods

The appendix classification dataset used in this study included four categories, namely normal appendix, appendicitis (with or without complications), and benign and malignant appendiceal tumours, which were categorised in an archive (machinery hierarchy of four appendiceal structures). An automatically generated 3D volume of interest (3D-VOI) dataset, yielding three true classes—non-appendicitis, simple appendicitis, and complicated appendicitis—was utilised for the development and technical assessment of the IA model (Fig. 1). This study was approved by the institutional review board ethics committee of Hallym University Sacred Heart Hospital (IRB number, 2023-04-017) and the data review board (IRB number, 2023-06-002) for external validation. The institutional review board granted a waiver of informed patient consent due to the retrospective nature of our study. All methods were performed in accordance with the relevant guideline and regulations.

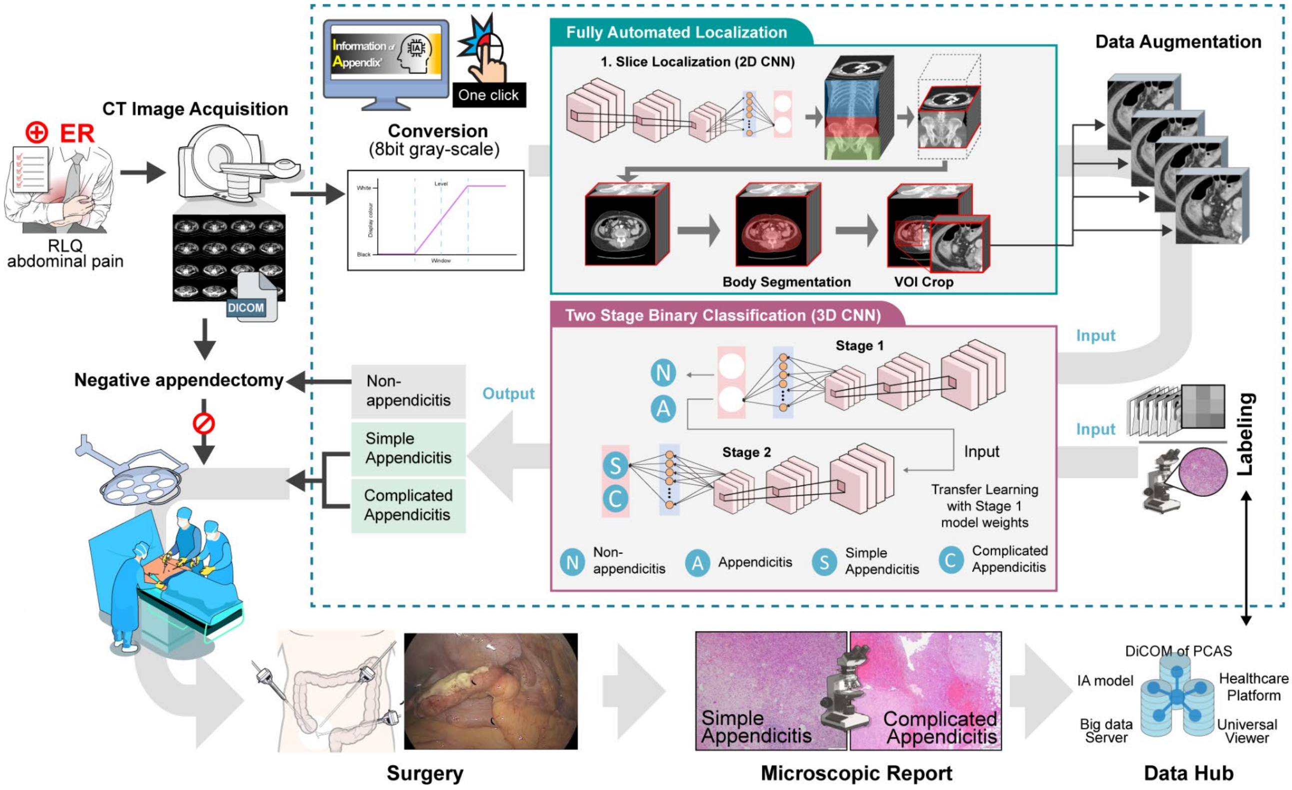


Fig. 1. Integration workflow of IA model. Information of Appendix (IA) represents a fully automated diagnostic framework employing a three-dimensional convolutional neural network (CNN). This model incorporates a pipeline featuring a two-stage binary algorithm connected to transfer learning, enabling the prediction of three classifications: non-appendicitis, simple appendicitis, and complicated appendicitis. DICOM files, sorted from the PACS platform, are pre-processed and anonymised. The Hounsfield unit scale of the extracted CT images is converted into 8-bit greyscale based on a window width of 60 and a level of 400. These grayscale images facilitate the localisation and generation of the volume of interest (VOI), which is then processed using the 3D-CNN algorithm. In the emergency medical system, the IA model simplifies the physician's task to a single button push at the DICOM sorting stage, providing a diagnosis with associated probabilities for the input image within the VOI.

Dataset and labels

The entire dataset, which was labelled in three classes (non-appendicitis, simple appendicitis, and complicated appendicitis) and collected from a single institution, was segmented into training, tuning, and internal validation datasets to develop the IA model. This study focused on the dataset of a large cohort of patients visiting emergency departments who had undergone intravenous contrast-enhanced abdominopelvic CT examinations from January 2014 to May 2022 with right or lower quadrant abdominal pain as their primary complaint. Additionally, a labelled dataset—externally sourced from a single institution via data transfer from January 2020 to June 2020—was employed for the external validation of the IA model.

The standard reference for the control group comprised patients diagnosed with non-appendicitis, with datasets acquired under clinical settings comparable to those of the patients. A negative image in the control group was defined as the absence of abnormal findings in the appendix or mesoappendix, irrespective of disease outside the VOI region. The appendiceal structure, which exhibited pathological findings corresponding to radiological observations, was labelled as a true class following rigorous re-examination of all portal-phase slices in axial images, provided that it satisfied all specified radiological and pathological inclusion criteria (Supplementary Fig. 1)[34,35]. Cases were labelled 'complicated appendicitis' only when both radiologic and pathologic findings indicated perforation. Images from patients with pathologically confirmed negative appendectomies were also labelled negative images.

Parameters of imaging protocol

In this study, the CT protocol parameters were as follows: abdomen or pelvis scans (intravenous contrast, 2 mg/kg, maximum 160 mL), scan timing (portal venous phase), range (from 4 cm above the liver dome to 1 cm below the ischial tuberosity), radiation dose (tube potential, KVP from 100 to 120), pitch 1.75:1, and reconstruction (5 mm cut slice for adults; 3 mm cut slice for children under 12 years).

Exploratory data analysis (EDA)

EDA was performed via cross-validation of DICOM header information, CT protocol parameters, and patient eligibility to mitigate heterogeneity-related bias or errors during DL tuning. Additionally, multiple iterations of EDA were conducted to determine the optimal z-axis-based region of interest (ROI) corresponding to the anatomical boundary of appendicitis, resulting in subsequent revisions and upgrades of the IA model based on EDA outcomes.

Inclusion and exclusion criteria

This study employed broad eligibility criteria, reflecting the rapidly expanding utilisation and interpretation volume of CT in ERs[17,18]. This approach aligns with the practices of several institutions where physicians employ a sensitive approach toward clinical suspicion of appendicitis as a cause of abdominal pain, utilising CT imaging to either confirm or rule out the condition. The exclusion criteria were established based on EDA outcomes (Supplementary Table 1). Patients who did not adhere to the CT imaging protocol were also excluded.

Automated pipeline for VOI extraction

The steps for automatically extracting VOI were as follows: (i) ROI slice localisation on the z-axis; (ii) VOI cropping, including body segmentation on the x- and y-axes; and 3D bounding box generation and optimisation. Initially, during ROI slice localisation, the region was segmented into three areas using 2D DenseNet169[36]: the reference initial ROI, upper region, and lower region. For z-axis range reference, the highest level of the iliac crest and lowest level of the sacroiliac (SI) joint were considered the initial ROIs (Fig. 2a)[37,38].

The automatic detection of the upper and lower regions in axial CT images involved distinguishing between images with and without the corresponding bony structures. This was achieved by identifying the first slice displaying the iliac crest and the final slice of the SI joint, irrespective of whether they were right or left-sided, as shown in Fig. 3. To increase the longitudinal range of the ROI on the z-axis, the upper margin was adjusted by adding thresholds of 25, 50, and 75 mm, and the lower margin was extended by approximately 50 mm down to the upper border of the symphysis pubis (Fig. 2b). The optimal thresholds for the upper and lower margins were determined via EDA (Fig. 4), which resulted in an extension of the initial ROI by an upper margin of 50 mm. Subsequently, the regions that correspond to the x-, y-, and z-axes were integrated to generate the initial 3D bounding box for VOI cropping. The automated extraction process of the VOI proceeded as follows: body segmentation, 3D-bounding box creation, and optimisation along the x- and y-axes for slices that correspond to the extended z-axis of the initial ROI, with an additional upper margin of 50 mm (Fig. 2c). Finally, the reconstructed 3D-VOIs from the entire pipeline were used as input data for the 3D-CNN models.

Two-stage binary classification based on transfer learning

The employed CNN architectures included 3D ResNet[39], DenseNet[36], and EfficientNet[40], which extend two-dimensional (2D) frameworks to 3D, as detailed in the Supplementary Information. The 3D-CNN—applied with a two-stage binary classification based on TL (Fig. 2d)[29]—was designed to classify three distinct classes. In the first step of the pipeline, the stage 1 classification (non-appendicitis vs. appendicitis) utilised Xavier uniform initialisation[41]. The trainable parameters from the first stage were then transferred to facilitate TL for the second stage classification (simple vs. complicated appendicitis). The learning rate, ranging from $1e^{-2}$ to $1e^{-5}$, was adjusted using three methods: reducing by a factor of 10 every 10 epochs, employing cosine-shaped learning rate scheduling, and decreasing by a factor of 10 upon plateauing in the tuning dataset. The Adam optimiser[42] was utilised with batch sizes of 8 and 100 epochs. Binary tests at each stage of the pipeline were configured using the Youden index and a cutoff value of 0.5. Additionally, 3D augmentation was implemented using a random combination of rotation within ±10° and contrast enhancement of up to 10%. Training was conducted on a

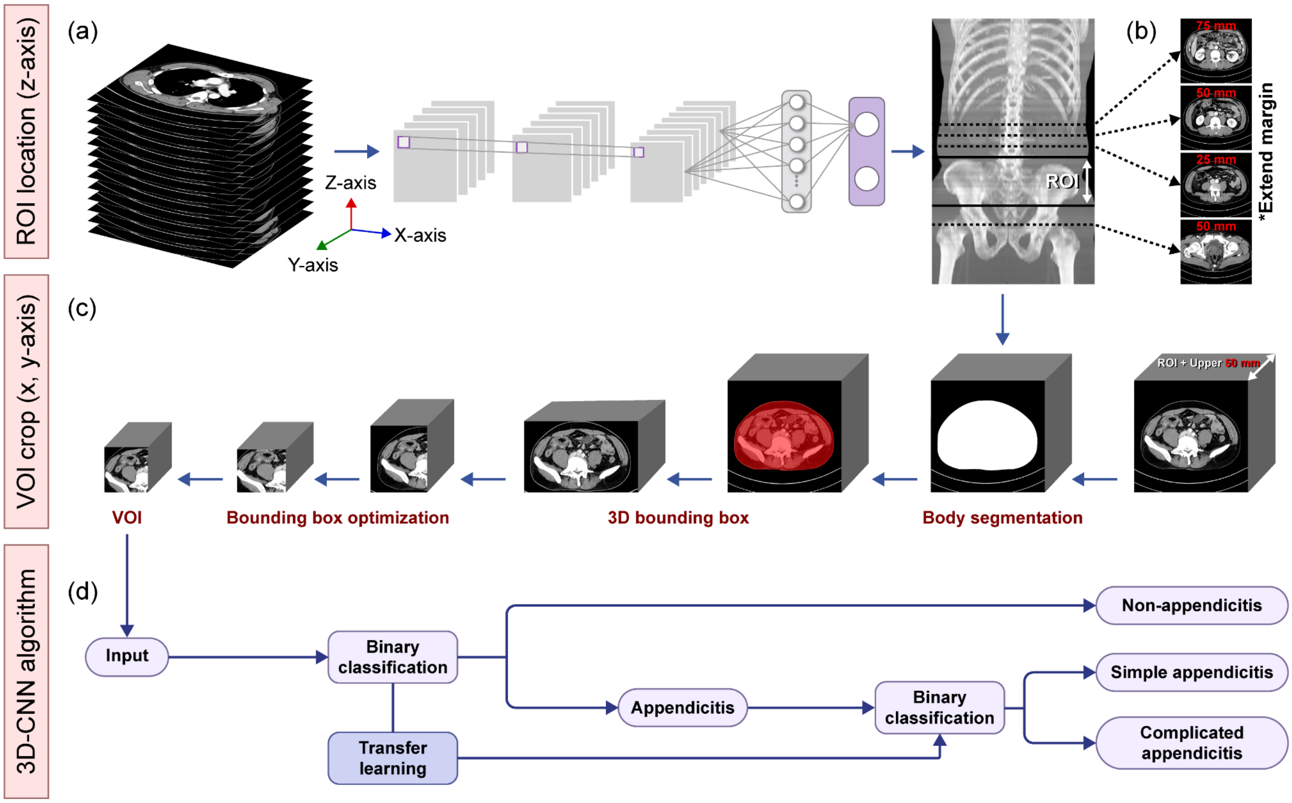


Fig. 2. Automated localisation for region of interest and generation of three-dimensional volume of interest for appendicitis. (**a**) In all cross-sectional slices of abdominopelvic computed tomography (APCT), automated localisation of the region of interest (ROI) on the *z*-axis, corresponding to the anatomical location of appendicitis, is performed using 2D-DenseNet169. The initial ROI range is defined by the highest level of the iliac crest and the lowest margin of the sacroiliac (SI) joint (indicated by the white bidirectional arrow) in the maximal intensity projection view (**b**). Extended upper and lower margins of the ROI (dotted line), with exploratory data analysis (EDA*), lead to the automated extraction of slices based on the extended *z*-axis ROI, with an additional upper margin of 50 mm referred to for the automated extraction of 3D-VOI. (**c**) Entire volume is cropped into the VOI using body segmentation and 3D-bounding box optimisation on the *x*- and *y*-axes. The refinement of the *x*- and *y*-axes of the VOI, achieved through body segmentation, is integrated with the *z*-axis to generate the initial 3D bounding box. Body segmentation that isolates only the patient area involves a sequence of processes including Otsu thresholding, seeded region growing, and morphological filtering. Unnecessary areas such as air and the bed in the *x*- and *y*-axis regions are excluded, especially omitting the left region. Further optimisation of the 3D bounding box entails reducing the range by 10% on three sides: dorsal, ventral, and right-abdominal wall, excluding the medial side. The reconstructed 3D-VOI, averaging 204 in width, 225 in height, and 28 slices in depth, is then input into the 3D-CNN model (**d**). *ROI* region of interest, *EDA**, exploratory data analysis, *VOI*, volume of interest.

server equipped with an Intel(R) Xeon(R) Silver 4216 CPU @ 2.10 GHz, 256 GB RAM, and NVIDIA GeForce RTX 3090 (24 GB) using the PyTorch 1.12.1 DL framework on Ubuntu 20.04.1. The configuration details of the IA model are presented in Supplementary Table 2.

Statistical analysis and outcome assessment

The performance of the model was evaluated at each learning phase by calculating 95% confidence intervals (CIs) for accuracy, sensitivity, specificity, predictive value (PV), F1 score, average precision score, and area under the receiver operating characteristic curve (AUROC). The negative appendectomy rate (NAR), defined as the false positive error, was also computed.

The output images from the model predictions were analysed using gradient-weighted class activation mappings (Grad-CAMs). These mappings visualise the feature maps, shaped as $8 \times 8 \times 2$ outputs from the final convolutional layer, by upscaling and overlaying them onto the input slices to highlight the salient regions used in the classification. The corresponding heatmaps were plotted in alignment with the input slices.

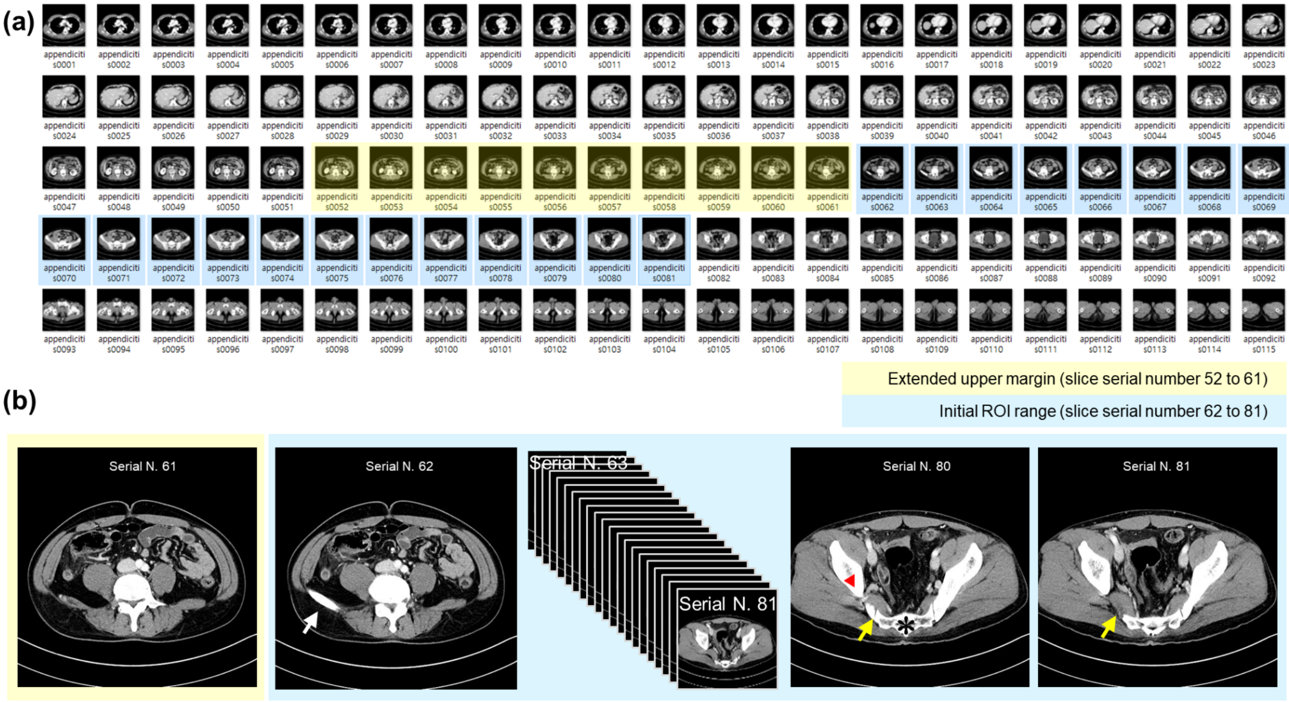


Fig. 3. Automated detection of upper and lower margin for region of interest in abdominal computed tomography slices. (**a**) In a 43-year-old male patient with simple appendicitis, among all axial slices of the abdominal section (serial numbers 1 to 115), thirty slices (from serial numbers 52 to 81) were automatically sorted by DenseNet169 for the initial region of interest (ROI) range (blue box) with an extended upper margin (yellow bow). (**b**) Automated detection of the highest level of the iliac crest (white arrow) at slice serial number 62, and the lowest level of the sacroiliac (SI) joint at slice serial number 81 (yellow arrow), positioned between the iliac bone (red triangle) and sacrum (black asterisk). Bony structure (white arrow) is not visible in slice serial number 61 but is seen from slice serial number 62 onwards. Similarly, the iliac bony structure of the SI joint (yellow arrow) disappears after slice serial number 81.

Results

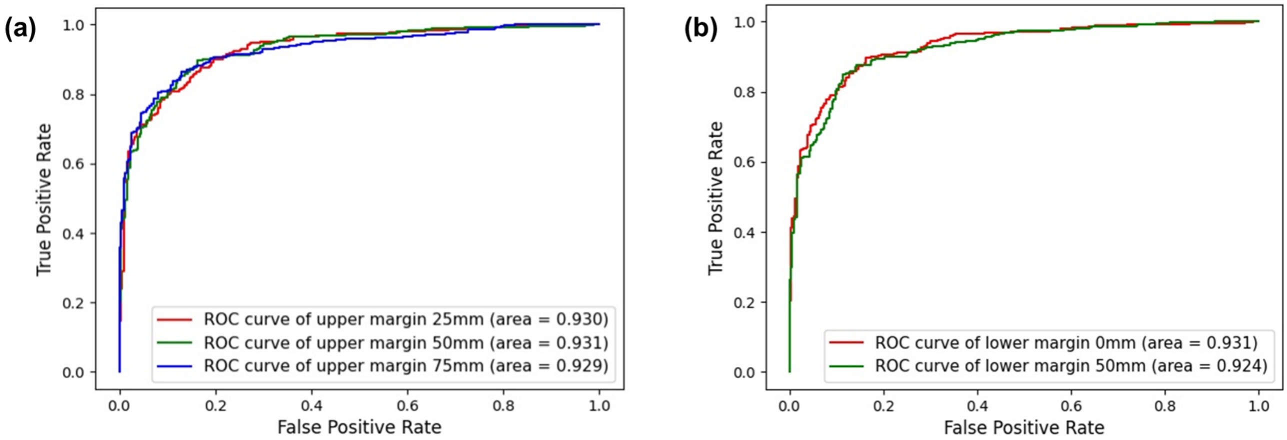
Patient demographics

To develop the dataset, 6,502 patients who visited the ER were included; the patients had a mean age of 38.0 years (SD 17.0 years), with 2,948 men (45.3%) and 3,554 women (54.7%) (Table 1). Among cases of patients who underwent appendectomy, 55 (1.8%) were identified as negative appendectomies. The external dataset exhibited heterogeneity in several aspects: a lower mean age ($p < 0.001$), greater proportion of non-appendicitis cases ($p = 0.05$), and different distribution of diagnoses for non-appendicitis patients ($p < 0.001$), despite the sex distribution being similar to that of the development dataset. In the non-appendicitis group, the most common diagnoses were normal cases with no abnormal findings, followed by terminal ileitis and ascending colitis; these findings were consistent in both the development and external datasets (Supplementary Fig. 2).

Model performance in developing and external dataset

Based on the internal validation on the development dataset, the performance of the seven 3D-CNN models varied according to stage 1 classification (Table 2). However, a pattern of relatively lower performance across all the CNN models was observed in stage 2 classification. The NAR range was 0.114–0.258 (95% CI) in stage 1 and 0.063–0.272 (95% CI) in stage 2, with DenseNet169 achieving the best performance among all the models (Table 2).

In the external validation, the performances of all the 3D-CNN models were relatively lower than those in the internal validation, despite the disease distribution being similar to that in the internal dataset. The DenseNet169 model achieved the highest performance among all the CNN models (Table 2). The NAR ranges for all the models were 0.124–0.360 (95% CI) in stage 1 and 0.231–0.422 (95% CI) in stage 2.



Extended margin	AUROC	Sensitivity	Specificity	PPV	NPV	F1 score
Upper margin (from iliac crest)						
25 mm	0.927 ± 0.002	0.861 ± 0.035	0.844 ± 0.043	0.864 ± 0.027	0.846 ± 0.027	0.861 ± 0.004
50 mm	0.931 ± 0.001	0.840 ± 0.034	0.873 ± 0.026	0.882 ± 0.017	0.830 ± 0.027	0.860 ± 0.010
75 mm	0.925 ± 0.002	0.864 ± 0.007	0.846 ± 0.018	0.864 ± 0.013	0.846 ± 0.005	0.864 ± 0.004
Lower margin (from sacroiliac joint)						
50 mm	0.922 ± 0.003	0.833 ± 0.014	0.856 ± 0.019	0.867 ± 0.016	0.820 ± 0.014	0.850 ± 0.014

Fig. 4. Exploratory data analysis for extended range of region of interest. EDA for 25, 50, and 75 mm-extended upper margins of the ROI on the *z*-axis using DenseNet169 indicated that the performance of each model was not significantly affected, as evidenced by consistent receiver operating characteristic (ROC) values. However, the model with a 50 mm threshold displayed the best performance (**a**). EDA for extended lower margins down to the upper border of the symphysis pubis, in comparison to model performance based on the SI joint, also displayed no significant difference in ROC values (**b**). (**c**) Model performance across three randomly split independent seeds was evaluated for the thresholds of 25, 50, and 75 mm-extended upper margins and the extended lower margin with an additional 50 mm from the lower margin of the sacroiliac joint. The analysis included area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score.

	Total	Non-appendicitis	Simple appendicitis	Complicated appendicitis	*p* value
Developing set	6502	3058 (47.0)	2789 (42.9)	655 (10.1)	
Training	5200	2446 (47.0)	2231 (42.9)	523 (10.1)	
Validation	651	306 (47.0)	279 (42.9)	66 (10.1)	
Test	651	306 (47.0)	279 (42.9)	66 (10.1)	
Mean age ± SD	38.0 ± 17.0	35.1 ± 16.3	38.8 ± 16.7	47.9 ± 17.9	< 0.001
Sex					< 0.001
Male	2948 (45.3)	1096 (35.8)	1480 (53.1)	372 (56.8)	
Female	3554 (54.7)	1962 (64.2)	1309 (46.9)	283 (43.2)	
External data	645	347 (53.8)	242 (37.5)	56 (8.7)	
Mean age ± SD	34.6 ± 15.5	35.4 ± 15.9	35.6 ± 14.6	38.3 ± 18.7	0.477
Sex					0.017
Male	308 (47.8)	148 (42.7)	132 (54.5)	28 (50.0)	
Female	337 (52.2)	199 (57.3)	110 (45.5)	28 (50.0)	

Table 1. Demographics and dataset. *SD* standard deviation.

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Average precision (95% CI)	F1 score (95% CI)	AUC (95% CI)
Internal validation dataset (n = 651)								
Stage 1 classification; non-appendicitis (n = 306) versus appendicitis (n = 345)								
Resnet34	0.836 (0.807–0.864)	0.843 (0.801–0.878)	0.827 (0.780–0.865)	0.846 (0.804–0.880)	0.824 (0.778–0.863)	0.924 (0.922–0.927)	0.831 (0.828–0.834)	0.908 (0.906–0.911)
Resnet101	0.849 (0.822–0.877)	0.878 (0.839–0.909)	0.817 (0.770–0.856)	0.844 (0.803–0.878)	0.856 (0.811–0.892)	0.932 (0.930–0.934)	0.838 (0.835–0.841)	0.921 (0.918–0.923)
Densenet121	0.868 (0.842–0.894)	0.878 (0.839–0.909)	0.856 (0.812–0.891)	0.873 (0.834–0.904)	0.862 (0.818–0.896)	0.936 (0.934–0.938)	0.869 (0.866–0.871)	0.927 (0.925–0.929)
Densenet169	0.868 (0.842–0.894)	0.896 (0.859–0.924)	0.837 (0.791–0.874)	0.861 (0.821–0.893)	0.877 (0.834–0.910)	0.942 (0.940–0.944)	0.872 (0.869–0.874)	0.930 (0.928–0.932)
Efficientnet B1	0.834 (0.806–0.863)	0.916 (0.882–0.941)	0.742 (0.690–0.788)	0.800 (0.758–0.836)	0.887 (0.842–0.920)	0.911 (0.909–0.913)	0.820 (0.817–0.823)	0.899 (0.896–0.901)
Efficientnet B3	0.842 (0.814–0.870)	0.803 (0.758–0.841)	0.886 (0.845–0.917)	0.888 (0.848–0.918)	0.799 (0.753–0.839)	0.925 (0.923–0.928)	0.847 (0.844–0.850)	0.915 (0.912–0.917)
Efficientnet B5	0.820 (0.791–0.850)	0.803 (0.758–0.841)	0.840 (0.795–0.877)	0.850 (0.807–0.884)	0.791 (0.743–0.831)	0.902 (0.899–0.905)	0.824 (0.821–0.827)	0.893 (0.890–0.896)
Stage 2 classification; simple appendicitis (n = 279) versus complicated appendicitis (n = 66)								
Resnet34	0.756 (0.707–0.805)	0.857 (0.750–0.922)	0.728 (0.667–0.782)	0.466 (0.377–0.556)	0.949 (0.905–0.972)	0.676 (0.665–0.687)	0.589 (0.579–0.598)	0.859 (0.853–0.864)
Resnet101	0.832 (0.790–0.874)	0.818 (0.708–0.892)	0.835 (0.783–0.877)	0.581 (0.479–0.676)	0.943 (0.903–0.967)	0.661 (0.648–0.673)	0.638 (0.630–0.645)	0.875 (0.870–0.880)
Densenet121	0.891 (0.856–0.926)	0.719 (0.598–0.814)	0.937 (0.899–0.961)	0.754 (0.633–0.845)	0.926 (0.885–0.952)	0.731 (0.719–0.743)	0.608 (0.597–0.619)	0.870 (0.864–0.877)
Densenet169	0.848 (0.808–0.888)	0.864 (0.760–0.926)	0.844 (0.793–0.884)	0.600 (0.499–0.693)	0.958 (0.922–0.977)	0.704 (0.693–0.715)	0.695 (0.687–0.702)	0.886 (0.881–0.891)
Efficientnet B1	0.832 (0.791–0.873)	0.875 (0.772–0.935)	0.821 (0.769–0.864)	0.554 (0.457–0.648)	0.963 (0.928–0.981)	0.741 (0.730–0.752)	0.671 (0.663–0.679)	0.877 (0.871–0.883)
Efficientnet B3	0.809 (0.762–0.855)	0.862 (0.750–0.928)	0.795 (0.736–0.843)	0.526 (0.427–0.624)	0.956 (0.916–0.977)	0.741 (0.730–0.751)	0.579 (0.570–0.588)	0.890 (0.885–0.895)
Efficientnet B5	0.809 (0.762–0.855)	0.867 (0.758–0.930)	0.793 (0.734–0.841)	0.536 (0.437–0.632)	0.956 (0.915–0.977)	0.734 (0.724–0.744)	0.656 (0.648–0.665)	0.891 (0.887–0.896)
External validation dataset (n = 645)								
Stage 1 classification; non-appendicitis (n = 347) versus appendicitis (n = 298)								
Resnet34	0.781 (0.749–0.813)	0.718 (0.664–0.766)	0.836 (0.793–0.871)	0.790 (0.737–0.834)	0.775 (0.730–0.815)	0.837 (0.833–0.840)	0.750 (0.746–0.755)	0.847 (0.845–0.850)
Resnet101	0.750 (0.717–0.784)	0.879 (0.837–0.911)	0.640 (0.588–0.688)	0.677 (0.629–0.722)	0.860 (0.813–0.897)	0.809 (0.805–0.814)	0.722 (0.718–0.726)	0.834 (0.831–0.837)
Densenet121	0.792 (0.761–0.824)	0.836 (0.789–0.873)	0.755 (0.707–0.797)	0.746 (0.696–0.789)	0.842 (0.798–0.879)	0.852 (0.848–0.856)	0.757 (0.754–0.761)	0.861 (0.858–0.864)
Densenet169	0.795 (0.764–0.826)	0.701 (0.647–0.750)	0.876 (0.837–0.907)	0.829 (0.778–0.871)	0.774 (0.730–0.812)	0.857 (0.854–0.860)	0.772 (0.768–0.775)	0.865 (0.862–0.867)
Efficientnet B1	0.761 (0.728–0.794)	0.768 (0.717–0.813)	0.755 (0.707–0.797)	0.729 (0.678–0.775)	0.792 (0.744–0.832)	0.816 (0.812–0.821)	0.742 (0.737–0.746)	0.835 (0.832–0.838)
Efficientnet B3	0.789 (0.758–0.821)	0.758 (0.707–0.803)	0.816 (0.771–0.853)	0.779 (0.728–0.823)	0.797 (0.752–0.836)	0.856 (0.853–0.859)	0.771 (0.767–0.775)	0.872 (0.870–0.875)
Efficientnet B5	0.761 (0.728–0.794)	0.789 (0.739–0.831)	0.738 (0.689–0.781)	0.721 (0.670–0.767)	0.803 (0.755–0.842)	0.817 (0.813–0.821)	0.745 (0.741–0.749)	0.839 (0.836–0.842)
Stage 2 classification; simple appendicitis (n = 242) versus complicated appendicitis (n = 56)								
Continued								

Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Average precision (95% CI)	F1 score (95% CI)	AUC (95% CI)
Resnet34	0.729 (0.669–0.789)	0.796 (0.663–0.885)	0.709 (0.636–0.773)	0.448 (0.348–0.553)	0.921 (0.861–0.956)	0.563 (0.548–0.578)	0.566 (0.555–0.577)	0.801 (0.794–0.809)
Resnet101	0.756 (0.704–0.808)	0.704 (0.571–0.809)	0.769 (0.707–0.821)	0.442 (0.341–0.547)	0.909 (0.857–0.943)	0.605 (0.592–0.619)	0.529 (0.518–0.540)	0.806 (0.799–0.813)
Densenet121	0.735 (0.680–0.790)	0.824 (0.697–0.904)	0.712 (0.645–0.771)	0.424 (0.331–0.523)	0.940 (0.890–0.968)	0.653 (0.641–0.665)	0.561 (0.551–0.571)	0.820 (0.812–0.827)
Densenet169	0.761 (0.703–0.819)	0.826 (0.692–0.909)	0.742 (0.670–0.803)	0.475 (0.369–0.583)	0.938 (0.882–0.968)	0.629 (0.616–0.641)	0.583 (0.572–0.594)	0.827 (0.820–0.833)
Efficientnet B1	0.642 (0.580–0.704)	0.878 (0.757–0.942)	0.578 (0.505–0.648)	0.361 (0.281–0.451)	0.945 (0.886–0.974)	0.578 (0.563–0.593)	0.543 (0.531–0.556)	0.798 (0.788–0.807)
Efficientnet B3	0.752 (0.696–0.808)	0.765 (0.632–0.860)	0.749 (0.679–0.807)	0.470 (0.366–0.576)	0.916 (0.859–0.951)	0.603 (0.590–0.617)	0.556 (0.546–0.566)	0.808 (0.802–0.814)
Efficientnet B5	0.634 (0.572–0.696)	0.896 (0.778–0.954)	0.567 (0.495–0.636)	0.347 (0.269–0.434)	0.955 (0.899–0.980)	0.557 (0.544–0.571)	0.518 (0.507–0.529)	0.797 (0.790–0.804)

Table 2. Performance of 3D-convolutional neural network models. *95% CI* 95% confidence interval, *PPV* positive predictive value, *NPV* negative predictive value, *AUC* area under the receiver operating characteristic curve.

Grad-CAM image

Figure 5 shows representative Grad-CAM images of subcecal, retrocecal, postileal, and pelvic type appendicitis, overlaid on CT images for comparison with actual locations. All Grad-CAM images were reviewed by clinicians and radiologists, who confirmed that the areas of interest of the DL model aligned with human assessments. In patients with simple appendicitis, the heatmaps predominantly emphasised the appendix and mesoappendix with a redder colour. Conversely, heatmaps for complicated appendicitis highlighted the appendiceal structure, as well as extra-appendiceal findings, including air, fluid collection, cavitary lesions with abscess, and adjacent organ structures, such as the small bowel, cecum, or bowel mesentery.

Discussion

In this study, we developed and validated a fully automated, two-stage CNN model to diagnose acute appendicitis using CT images. The model effectively distinguished between non-appendicitis and appendicitis cases in the first stage, and simple and complicated appendicitis in the second. The DenseNet169 architecture achieved the highest performance, with an AUC of 0.930 for the initial classification (non-appendicitis vs. appendicitis) and 0.886 for the second-stage classification (simple vs. complicated appendicitis). These results underscore the potential of our IA model to support clinical decision-making by providing an informed automated diagnosis of appendicitis using CT images.

This study acknowledges the inherent limitations and potential flaws of the application of AI in the field of medicine, demonstrating the sophisticated architecture and reliability of an IA model via external validation. Among numerous studies employing AI algorithms, only 6.0% of recently published research on the diagnostic analysis of medical images has reported external validation performance[33]. The IA model, developed in this study using a two-stage binary algorithm combined with TL with fine-tuning, embeds parameters learned from a high-quality labelled dataset of a large population across numerous epochs. The model simplifies the role of the clinician to merely activating it with the push of a button at the point of DICOM sorting from patient CT scans, which subsequently provides diagnostic probabilities for the input images. Moreover, from a machine learning perspective, structuring the IA model as a two-stage binary algorithm is advantageous in increasing the classification sensitivity for complicated appendicitis. A single-step multi-class classification model risks the overlooking of cases with complicated appendicitis owing to data imbalance caused by its low incidence rate (Table 1).

However, we remain cautious about overestimating the capabilities of the IA model, which are specifically modified for feature extraction of the appendix within the VOI. The specialisation may limit its generalisability and reproducibility for patients with abdominal pain in ERs; however, it is expected to offer considerable benefits to clinicians and patients in medically underserved regions without radiologists available to read CT images[43].

In the field of medical image analysis, TL facilitates knowledge transfer at the parametric level and is extensively used to optimise ML models. It enhances the learning time and performance by leveraging datasets from the source domain, addressing the lack of data annotations in the target domain and the unavailability of large labelled datasets[29]. However, we acknowledge that clinicians, who integrate patient history, physical examination, review of systems, and laboratory results to formulate a diagnosis before interpreting CT images, do not strictly adhere to the two-stage binary decision-making process inherent in TL. Nonetheless, the rationale for not employing multiclass algorithms involves emulating human cognitive processes as closely as possible. However, it does not imply that clinicians cannot diagnose complicated appendicitis without prior identification of normal or simple cases. Instead, TL facilitates a more efficient learning process for interpreting complicated appendicitis cases by adapting pre-trained tasks, as opposed to starting the ML process from scratch[30].

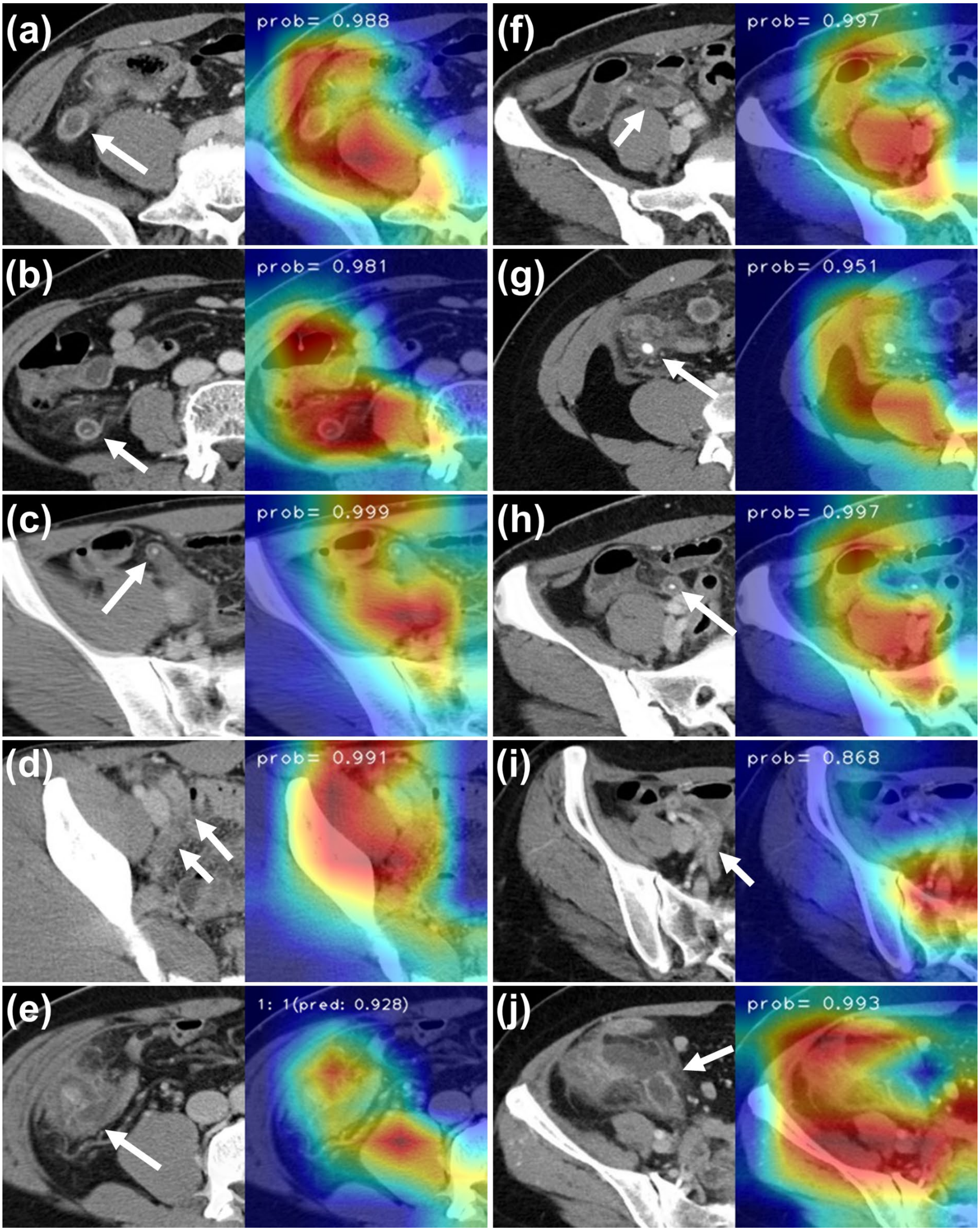
In this study, the VOI provided a feature map that enabled spatial perception of areas beyond the appendix, particularly the extra-appendiceal region. Utilising handcrafted annotations focused solely on the appendix and mesoappendix can result in information loss concerning extra-appendiceal features, such as extraluminal gas, fluid collection, appendicolith, bowel content due to perforation, and adjacent structural changes caused by appendiceal inflammation. Notably, less than 3% of the true class cases were located outside the VOI, underscoring the challenge of expanding the ROI, which may detrimentally affect the performance of the model. The range of the VOI, established to cover all types of appendicitis locations, including pelvic types, was optimised by adjusting the upper and lower margins via EDA[37,38].

However, fatal errors are inevitable when analysing data with potential causes outside of the VOI—a situation applicable to data from patients excluded in real-world clinical settings. Despite being rare, oncologic risks are concerning within the VOI field. Masquerading appendicitis harbouring hidden malignancies, particularly those complicated by perforation, presents a diagnostic challenge and can worsen patient outcomes[44]. An accurate preoperative diagnosis of complicated appendicitis with potential tumour risk is difficult[45].

In this study, the most common cause of false positive errors in both the development and external datasets was ascending colon diverticulitis, followed by terminal ileitis, regardless of the 3D-CNN model employed. Similar to the issues faced by clinicians, the IA model did not effectively address the challenge of interpreting CT images of patients with appendicitis that mimics ascending colon diverticulitis or terminal ileitis[46,47]. An analysis of Grad-CAM images from false class cases revealed that inflammation might alter image texture features, such as edgeness per unit area and local neighbourhood intensity, involving the appendix and surrounding extra-appendiceal structures, such as fat, fluid, or bowel within the VOI[48]. However, the model failed to distinguish discrepancies between salience and activation maps, lacking top–down evidence for the textures of false error images[49]. The technical key to reducing false errors lies in the incorporation of additional data, balanced fitting, and optimising the loss function[30].

This study acknowledges several limitations, including potential iatrogenic perforation due to surgical manipulation and variable responses to antibiotics or delayed surgery between the times of radiologic and pathologic labelling. Despite these limitations, we posit that our robust dataset, which includes data from normal individuals, patients with non-specific conditions, and patients with a broad spectrum of diseases, may yield better results than those of previous studies[3,22] that lacked control data. However, issues such as the automatic uploading of sorted DICOM series from the PACS platform, large amount of excluded data, limitation of DL to the confined VOI rather than the entire abdominal section, and hidden risk of appendiceal tumours within the VOI suggest that practical application in real-world settings may still be distant. To address the limitations of the model, we are currently in the process of conducting a prospective randomised study comparing the diagnosis by non-radiologists and the IA model [ClinicalTrials.gov ID: NCT06175169] to ensure that the model fulfils ethical requirements and is safe. Despite the limitations mentioned, our model has the potential to be applied as a triage tool, particularly in medically underserved regions where radiologists are not readily available. The model could be implemented in an initial screening phase, prioritizing cases flagged as likely appendicitis during stage 1 classification (normal vs. appendicitis) for immediate review by available radiologists or physicians. This triage function could optimize resources by ensuring that cases requiring urgent intervention are assessed without delay, improving the efficiency of clinical workflows.

In conclusion, this study demonstrated that the proposed DL model can identify non-appendicitis, acute appendicitis, and complicated appendicitis from CT scan images of critical patients who presented right-sided or lower abdominal pain in the ER, representing a diagnostic approach for clinicians. The IA model, owing to its generality and reproducibility within the VOI, may alleviate the burden of CT interpretation for physicians, particularly in medically vulnerable systems or regions that lack readily available radiologists or specialist-based interpretations.


(a)
prob= 0.988
(f)
prob= 0.997
(b)
prob= 0.981
(g)
prob= 0.951
(c)
prob= 0.999
(h)
prob= 0.997
(d)
prob= 0.991
(i)
prob= 0.868
(e)
1: 1(pred: 0.928)
(j)
prob= 0.993

◀**Fig. 5**. Gradient-weighted class activation mapping for true classes. APCT (right) and gradient-weighted class activation mapping (Grad-CAM) images (left) of patients visiting the emergency room with right-sided lower quadrant pain are displayed in the internal validation dataset using DenseNet169. In the internal validation dataset, (**a**) dilated appendiceal lumen with the enhancement of a thickened wall (arrow) is shown alongside a Grad-CAM image indicative of subcecal type simple appendicitis, with a probability of 0.988. (**b**) Grad-CAM image corresponding to retrocecal type simple appendicitis (arrow) with a probability of 0.981. (**c**) Appendicolith (arrow) alongside a Grad-CAM image indicative of postileal-type simple appendicitis with a probability of 0.999. (**d**) Enhanced appendiceal lumen situated along the pelvic cavity (arrow) with a corresponding Grad-CAM image for pelvic type simple appendicitis, showing a probability of 0.991. (**e**) Multi-loculated cavity surrounding mural haziness of the appendix (arrow) and a Grad-CAM image indicative of complicated appendicitis with a probability of 0.928. In the external validation dataset, Grad-CAM images (**f–i**) displayed heatmaps correlated to the appendiceal structure (arrow) for subcecal, retrocecal, postileal, and pelvic type simple appendicitis, each with a higher probability. Grad-CAM (**j**) highlighted extra-appendiceal air and fluid collection with mural defect (arrow) correlating to complicated appendicitis, displaying a probability of 0.993.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 3 May 2024; Accepted: 23 December 2024
Published online: 05 March 2025

References

1. Humes, D. J. & Simpson, J. Acute appendicitis. *BMJ* **333**, 530–534. https://doi.org/10.1136/bmj.38940.664363.AE (2006).
2. Ghiatas, A. A. et al. Computed tomography of the normal appendix and acute appendicitis. *Eur. Radiol.* **7**, 1043–1047. https://doi.org/10.1007/s003300050249 (1997).
3. Park, J. J. et al. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. *Sci. Rep.* **10**, 9556. https://doi.org/10.1038/s41598-020-66674-7 (2020).
4. Kim, H. C., Yang, D. M., Jin, W. & Park, S. J. Added diagnostic value of multiplanar reformation of multidetector CT data in patients with suspected appendicitis. *RadioGraphics* **28**, 393–405. https://doi.org/10.1148/rg.282075039 (2008) (**discussion 405–396**).
5. Writing Group for the CODA Collaborative et al. Analysis of outcomes associated with outpatient management of nonoperatively treated patients with appendicitis. *JAMA Netw. Open* **5**, e2220039. https://doi.org/10.1001/jamanetworkopen.2022.20039 (2022).
6. Livingston, E. H., Woodward, W. A., Sarosi, G. A. & Haley, R. W. Disconnect between incidence of nonperforated and perforated appendicitis: Implications for pathophysiology and management. *Ann. Surg.* **245**, 886–892. https://doi.org/10.1097/01.sla.0000256391.05233.aa (2007).
7. Körner, H. et al. Incidence of acute nonperforated and perforated appendicitis: Age-specific and sex-specific analysis. *World J. Surg.* **21**, 313–317. https://doi.org/10.1007/s002689900235 (1997).
8. Addiss, D. G., Shaffer, N., Fowler, B. S. & Tauxe, R. V. The epidemiology of appendicitis and appendectomy in the United States. *Am. J. Epidemiol.* **132**, 910–925. https://doi.org/10.1093/oxfordjournals.aje.a115734 (1990).
9. Drake, F. T. et al. Time to appendectomy and risk of perforation in acute appendicitis. *JAMA Surg.* **149**, 837–844. https://doi.org/10.1001/jamasurg.2014.77 (2014).
10. Giraudo, G., Baracchi, F., Pellegrino, L., Dal Corso, H. M. & Borghi, F. Prompt or delayed appendectomy? Influence of timing of surgery for acute appendicitis. *Surg. Today* **43**, 392–396. https://doi.org/10.1007/s00595-012-0250-5 (2013).
11. Busch, M. et al. In-hospital delay increases the risk of perforation in adults with appendicitis. *World J. Surg.* **35**, 1626–1633. https://doi.org/10.1007/s00268-011-1101-z (2011).
12. Fair, B. A. et al. The impact of operative timing on outcomes of appendicitis: A National Surgical Quality Improvement Project analysis. *Am. J. Surg.* **209**, 498–502. https://doi.org/10.1016/j.amjsurg.2014.10.013 (2015).
13. Abou-Nukta, F. et al. Effects of delaying appendectomy for acute appendicitis for 12 to 24 hours. *Arch. Surg.* **141**, 504–506. https://doi.org/10.1001/archsurg.141.5.504 (2006) (**discussion 506–507**).
14. Reyes, A. M., Royan, R., Feinglass, J., Thomas, A. C. & Stey, A. M. Patient and hospital characteristics associated with delayed diagnosis of appendicitis. *JAMA Surg.* **158**, e227055. https://doi.org/10.1001/jamasurg.2022.7055 (2023).
15. Simmering, J. E., Polgreen, L. A., Talan, D. A., Cavanaugh, J. E. & Polgreen, P. M. Association of appendicitis incidence with warmer weather independent of season. *JAMA Netw. Open* **5**, e2234269. https://doi.org/10.1001/jamanetworkopen.2022.34269 (2022).
16. Berdahl, C. T., Vermeulen, M. J., Larson, D. B. & Schull, M. J. Emergency department computed tomography utilization in the United States and Canada. *Ann. Emerg. Med.* **62**, 486-494.e3. https://doi.org/10.1016/j.annemergmed.2013.02.018 (2013).
17. Hess, E. P. et al. Trends in computed tomography utilization rates: A longitudinal practice-based study. *J. Patient Saf.* **10**, 52–58. https://doi.org/10.1097/PTS.0b013e3182948b1a (2014).
18. Raja, A. S. et al. Radiology utilization in the emergency department: Trends of the past 2 decades. *AJR Am. J. Roentgenol.* **203**, 355–360. https://doi.org/10.2214/AJR.13.11892 (2014).
19. McDonald, R. J. et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198. https://doi.org/10.1016/j.acra.2015.05.007 (2015).
20. Peng, Y. C., Lee, W. J., Chang, Y. C., Chan, W. P. & Chen, S. J. Radiologist burnout: trends in medical imaging utilization under the national health insurance system with the universal code bundling strategy in an academic tertiary medical centre. *Eur. J. Radiol.* **157**, 110596. https://doi.org/10.1016/j.ejrad.2022.110596 (2022).
21. Lam, A. et al. Artificial intelligence for predicting acute appendicitis: A systematic review. *ANZ J. Surg.* **93**, 2070–2078. https://doi.org/10.1111/ans.18610 (2023).
22. Rajpurkar, P. et al. AppendiXNet: Deep learning for diagnosis of appendicitis from A small dataset of CT exams using video pretraining. *Sci. Rep.* **10**, 3958. https://doi.org/10.1038/s41598-020-61055-6 (2020).
23. Hsieh, C. H. et al. Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery* **149**, 87–93. https://doi.org/10.1016/j.surg.2010.03.023 (2011).

24. Reismann, J. et al. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. *PLoS ONE* **14**, e0222030. https://doi.org/10.1371/journal.pone.0222030 (2019).
25. Park, S. Y. & Kim, S. M. Acute appendicitis diagnosis using artificial neural networks. *Technol. Health Care* **23**(Suppl 2), S559–S565. https://doi.org/10.3233/THC-150994 (2015).
26. Ye, Z. et al. Development and validation of an automated image-based deep learning platform for sarcopenia assessment in head and neck cancer. *JAMA Netw. Open* **6**, e2328280. https://doi.org/10.1001/jamanetworkopen.2023.28280 (2023).
27. Hsu, W. et al. External validation of an ensemble model for automated mammography interpretation by artificial intelligence. *JAMA Netw. Open* **5**, e2242343. https://doi.org/10.1001/jamanetworkopen.2022.42343 (2022).
28. Wang, Z., Du, B. & Guo, Y. Domain adaptation with neural embedding matching. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 2387–2397. https://doi.org/10.1109/TNNLS.2019.2935608 (2020).
29. Kim, H. E. et al. Transfer learning for medical image classification: a literature review. *BMC Med. Imaging* **22**, 69. https://doi.org/10.1186/s12880-022-00793-7 (2022).
30. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **9**, 611–629. https://doi.org/10.1007/s13244-018-0639-9 (2018).
31. Teno, J. M. Garbage in, garbage out-words of caution on big data and machine learning in medical practice. *JAMA Health Forum* **4**, e230397. https://doi.org/10.1001/jamahealthforum.2023.0397 (2023).
32. Kocher, K. E. et al. National trends in use of computed tomography in the emergency department. *Ann. Emerg. Med.* **58**, 452–462. https://doi.org/10.1016/j.annemergmed.2011.05.020 (2011).
33. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. *Korean J. Radiol.* **20**, 405–410. https://doi.org/10.3348/kjr.2019.0025 (2019).
34. Ahn, S. LOCAT (low-dose computed tomography for appendicitis trial) comparing clinical outcomes following low- vs standard-dose computed tomography as the first-line imaging test in adolescents and young adults with suspected acute appendicitis: Study protocol for a randomized controlled trial. *Trials* **15**, 28. https://doi.org/10.1186/1745-6215-15-28 (2014).
35. Karande, G. Y. et al. Spectrum of computed tomography manifestations of appendiceal neoplasms: Acute appendicitis and beyond. *Singap. Med. J.* **60**, 173–182. https://doi.org/10.11622/smedj.2019035 (2019).
36. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708 (2017).
37. Davis, J. et al. Computed tomography localization of the appendix in the pediatric population relative to the lumbar spine. *Pediatr. Radiol.* **47**, 301–305. https://doi.org/10.1007/s00247-016-3773-x (2017).
38. Lin, W., Jeffrey, R. B., Trinh, A. & Olcott, E. W. Anatomic reasons for failure to visualize the appendix with graded compression sonography: Insights from contemporaneous CT. *AJR Am. J. Roentgenol.* **209**, W128–W138. https://doi.org/10.2214/AJR.17.18059 (2017).
39. He, K., Zhang, X., Ren, S. & Sun, J. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
40. Tan, M. & Le, Q. In *International Conference on Machine Learning*, 6105–6114 (PMLR, 2019).
41. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceeding of the 13th International Conference on Artificial Intelligence and Statistics*, 249–256 (2010).
42. Adams, R. & Bischof, L. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 641–647. https://doi.org/10.1109/34.295913 (1994).
43. Chen, J. Y., Vedantham, S. & Lexa, F. J. Burnout and work-work imbalance in radiology-wicked problems on a global scale. A baseline pre-COVID-19 survey of US neuroradiologists compared to international radiologists and adjacent staff. *Eur. J. Radiol.* **155**, 110153. https://doi.org/10.1016/j.ejrad.2022.110153 (2022).
44. Son, I. T. et al. Comparison of long-term oncological outcomes of appendiceal cancer and colon cancer: A multicenter retrospective study. *Surg. Oncol.* **25**, 37–43. https://doi.org/10.1016/j.suronc.2015.12.006 (2016).
45. Brunner, M. et al. Risk factors for appendiceal neoplasm and malignancy among patients with acute appendicitis. *Int. J. Colorectal Dis.* **35**, 157–163. https://doi.org/10.1007/s00384-019-03453-5 (2020).
46. Horn, A. E. & Ufberg, J. W. Appendicitis, diverticulitis, and colitis. *Emerg. Med. Clin. N. Am.* **29**, 347–368. https://doi.org/10.1016/j.emc.2011.01.002 (2011).
47. Ziegelmayer, S. et al. Development and validation of a deep learning algorithm to differentiate colon carcinoma from acute diverticulitis in computed tomography images. *JAMA Netw. Open* **6**, e2253370. https://doi.org/10.1001/jamanetworkopen.2022.53370 (2023).
48. Koçak, B., Durmaz, E. Ş, Ateş, E. & Kılıçkesmez, Ö. Radiomics with artificial intelligence: A practical guide for beginners. *Diagn. Interv. Radiol.* **25**, 485–495. https://doi.org/10.5152/dir.2019.19321 (2019).
49. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510. https://doi.org/10.1038/s41568-018-0016-5 (2018).

Acknowledgements

The Medical A.I. Center of Hallym University, Doheon Institute for Digital Innovation in Medicine contributed to this research. The facilitations and tasks were launched under Might Hallym 4.0 vision in April 2019. We acknowledge that this research was made possible by MH 4.0 vision. We worked under Project MHo4 [ClinicalTrials.gov ID: NCT06175169] and constructed the archive (Machinery Hierarchy of Four Appendiceal Structures). This study was awarded the BEST INVESTIGATOR AWARD at the Annual Congress of KSS 2022 and the 74th Congress of the Korean Surgical Society, November 5, 2022, Seoul, South Korea.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: I.S, BJ.C, MJ.K, MJ.K, J.K, BY.O, JW.K, S.H, WS.Y; data collection: M.K; analysis and interpretation of results: T.P, J.K, M.K, I.S, BJ.C; draft manuscript preparation: M.K, I.S. All authors reviewed the results and approved the final version of the manuscript.

Funding

This study was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) and funded by the Korean government (MSIT) (No. NRF-2022R1C1C1010643).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-84348-6.

Correspondence and requests for materials should be addressed to B.-J.C. or I.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.