Annotation-free multi-organ anomaly detection in abdominal CT using free-text radiology reports: a multi-centre retrospective study

Junya Sato,[a,b] Kento Sugimoto,[c] Yuki Suzuki,[a] Tomohiro Wataya,[a,b] Kosuke Kita,[a] Daiki Nishigaki,[a,b] Miyuki Tomiyama,[a,b] Yu Hiraoka,[a,b] Masatoshi Hori,[a] Toshihiro Takeda,[c] Shoji Kido,[a,*] and Noriyuki Tomiyama[b]

[a]Department of Artificial Intelligence in Diagnostic Radiology, Osaka University Graduate School of Medicine, 2-2, Yamadaoka, Suita, Osaka, 565-0871, Japan
[b]Department of Radiology, Osaka University Graduate School of Medicine, 2-2, Yamadaoka, Suita, Osaka, 565-0871, Japan
[c]Department of Medical Informatics, Osaka University Graduate School of Medicine, 2-2, Yamadaoka, Suita, Osaka, 565-0871, Japan

eBioMedicine
2024;110: 105463
Published Online 28 November 2024
https://doi.org/10.1016/j.ebiom.2024.105463

Summary

Background Artificial intelligence (AI) systems designed to detect abnormalities in abdominal computed tomography (CT) could reduce radiologists' workload and improve diagnostic processes. However, development of such models has been hampered by the shortage of large expert-annotated datasets. Here, we used information from free-text radiology reports, rather than manual annotations, to develop a deep-learning-based pipeline for comprehensive detection of abdominal CT abnormalities.

Methods In this multicentre retrospective study, we developed a deep-learning-based pipeline to detect abnormalities in the liver, gallbladder, pancreas, spleen, and kidneys. Abdominal CT exams and related free-text reports obtained during routine clinical practice collected from three institutions were used for training and internal testing, while data collected from six institutions were used for external testing. A multi-organ segmentation model and an information extraction schema were used to extract specific organ images and disease information, CT images and radiology reports, respectively, which were used to train a multiple-instance learning model for anomaly detection. Its performance was evaluated using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1 score against radiologists' ground-truth labels.

Findings We trained the model for each organ on images selected from 66,684 exams (39,255 patients) and tested it on 300 (295 patients) and 600 (596 patients) exams for internal and external validation, respectively. In the external test cohort, the overall AUC for detecting organ abnormalities was 0.886. Whereas models trained on human-annotated labels performed better with the same number of exams, those trained on larger datasets with labels auto-extracted via the information extraction schema significantly outperformed human-annotated label-derived models.

Interpretation Using disease information from routine clinical free-text radiology reports allows development of accurate anomaly detection models without requiring manual annotations. This approach is applicable to various anatomical sites and could streamline diagnostic processes.

Funding Japan Science and Technology Agency.

Keywords: Artificial intelligence; Named-entity recognition; Radiology report; Anomaly detection; Computed tomography

Introduction

Computed tomography (CT) is an essential diagnostic tool in various clinical settings. CT use has increased globally, with the OECD reporting an average increase exceeding 60% in 2021.[1] Thorough search of CT images for abnormalities is a critical routine practice for radiologists; thus, the increase in scans implies increased radiologists' workload.[2,3] Moreover, increased workload correlates with inter-reader interpretive discrepancies.[4] Accordingly, new technological approaches are needed

*Corresponding author.
E-mail address: kido@radiol.med.osaka-u.ac.jp (S. Kido).

Research in context

Evidence before this study

With the recent increase in the number of CT scans, deep-learning-based diagnostic support systems have the potential to improve radiologists' workflow and diagnostic accuracy. However, these systems require training with large datasets that need to be annotated by medical experts. Recent advances in language models have demonstrated effectiveness in extracting information from free-text radiology reports, thereby reducing the annotation burden. A search of PubMed over the 10 years preceding 1 June 2024, using the keywords ["computed tomography" OR "CT"] AND ["radiology report" OR "free-text"] AND ["detection" OR "classification" OR "automat*"] AND ["deep learning" OR "machine learning"] resulted in 53 papers after excluding studies that did not use reports for model training. Of these, 46 studies focused on the classification or summarisation of text alone, five studies used reports for image training, and two studies were review papers. Of the five studies that used images, four studies performed disease detection, and only one study targeted abdominal organs.

Added value of this study

To the best of our knowledge, this study is based on the largest dataset of abdominal CT images and corresponding radiology reports, comprising exams from nine hospitals in various clinical settings. We developed a deep-learning-based pipeline to detect clinical abnormalities in five organs, using a multi-organ segmentation model to process 3D CT images to extract organ-specific regions and an information extraction schema to analyse the accompanying radiology reports to identify disease information. The model demonstrated the high accuracy in abnormality detection across all organs, outperforming models trained with a limited number of radiologist-annotated datasets.

Implications of all the available evidence

Auto-extracted labels from radiology free-text reports can substitute for manual annotations. This approach is broadly applicable across different anatomical sites and can reduce radiologists' workload while heralding significant advances in computer-aided diagnosis.

to streamline diagnostic processes and enhance their accuracy.

Many studies have demonstrated the utility of artificial intelligence (AI) to improve radiologists' workflow[5] and diagnostic accuracy.[6] However, developing accurate AI diagnostic support systems with deep learning requires training with large labelled datasets.[7] Annotating medical images is labour-intensive and time-consuming, and also requires medical expertise to be effective.[8] Furthermore, the scarcity of medical data combined with patient privacy issues limits public dataset availability. Accordingly, innovative approaches are needed to overcome the lack of annotated data to advance medical AI.

The annotation burden can be alleviated by reusing information from medical records obtained during routine clinical practice.[9] After natural language processing to extract pertinent disease information, free-text radiology reports of chest X-ray or head CT images have been used as effective labels, reducing the need for manual annotation.[10–13] However, applying this method to abdominal CT images is difficult, because of their complexity. Abdominal CT, which captures extensive body areas and multiple organs in three-dimensional (3D) images, comprises tens of millions of voxels and varies substantially in individual characteristics and imaging conditions. This complexity necessitates a refined approach to identify each organ accurately, and extract and link disease information from free-text reports. Additionally, the scarcity of large datasets that pair abdominal CT images with detailed reports hampers AI research and development of effective diagnostic support systems.

Here, we propose a fully end-to-end pipeline for detecting abnormalities in five organs (the liver, gallbladder, kidney, spleen, and pancreas). We first extracted organ-specific regions from CT images using a multiorgan segmentation model, which enables focused predictions of specific organs. We then employed an information extraction schema to derive disease information for each organ from radiology reports and used this information as a training label, thereby eliminating the need for additional manual annotations.

Methods

Study design and participants

Our anomaly detection pipeline criteria followed the checklist for Artificial Intelligence in Medical Imaging criteria.[14] In this retrospective study, data were collected from nine institutions—The University of Tokyo, Keio University, Okayama University, Ehime University, Juntendo University, Kyoto University, Kyushu University, Osaka University, and Tokushima University—using the Japan Medical Imaging Database (J-MID). This multicentre study was approved by the ethics committees of Juntendo University (approval ID: E21-0099) and Osaka University (approval ID: K21298), and anonymised data were exchanged among the institutions under a data-sharing agreement. The need for obtaining written informed consent was waived because of the retrospective data acquisition from the J-MID. Participants' gender was self-reported and did not influence the study design. The images were acquired based on protocols used in routine clinical practice at each

institution. The CT vendor information is listed in Supplementary Tables S1 and S2.

Inclusion criteria and data split

Patient enrolment is summarised in Fig. 1. We used abdominal CT images collected from the nine institutions during routine clinical practice, from July 1, 2020, to February 27, 2023. Images with a large number of slices (>300) or a small number of slices (<40) were excluded to eliminate incomplete series, remove images not containing the abdomen, and improve computational efficiency. Axial abdominal exams were selected according to the protocol names and outputs from the information extraction schema. These images were divided into internal training, internal test, and external test cohorts. Among them, images from three institutions were designated as the internal test cohort and those from the remaining six as the external test cohort. For both internal and external test cohorts, the 100 most recent exams from each institution were selected. Data from patients assigned to the internal test cohort were excluded from the internal training cohort. Additionally, images with poor segmentation or those from which disease information could not be extracted using a structured protocol were excluded from the internal training cohort.

Anomaly detection pipeline overview

We used a three-stage approach: labelled dataset extraction, model training, and inference (Fig. 2). We curated a dataset comprising abdominal CT images accompanied by radiology reports and associated patient information from the J-MID database. We then trained a deep-learning-based pipeline to detect abnormal findings across five abdominal organs. During training,

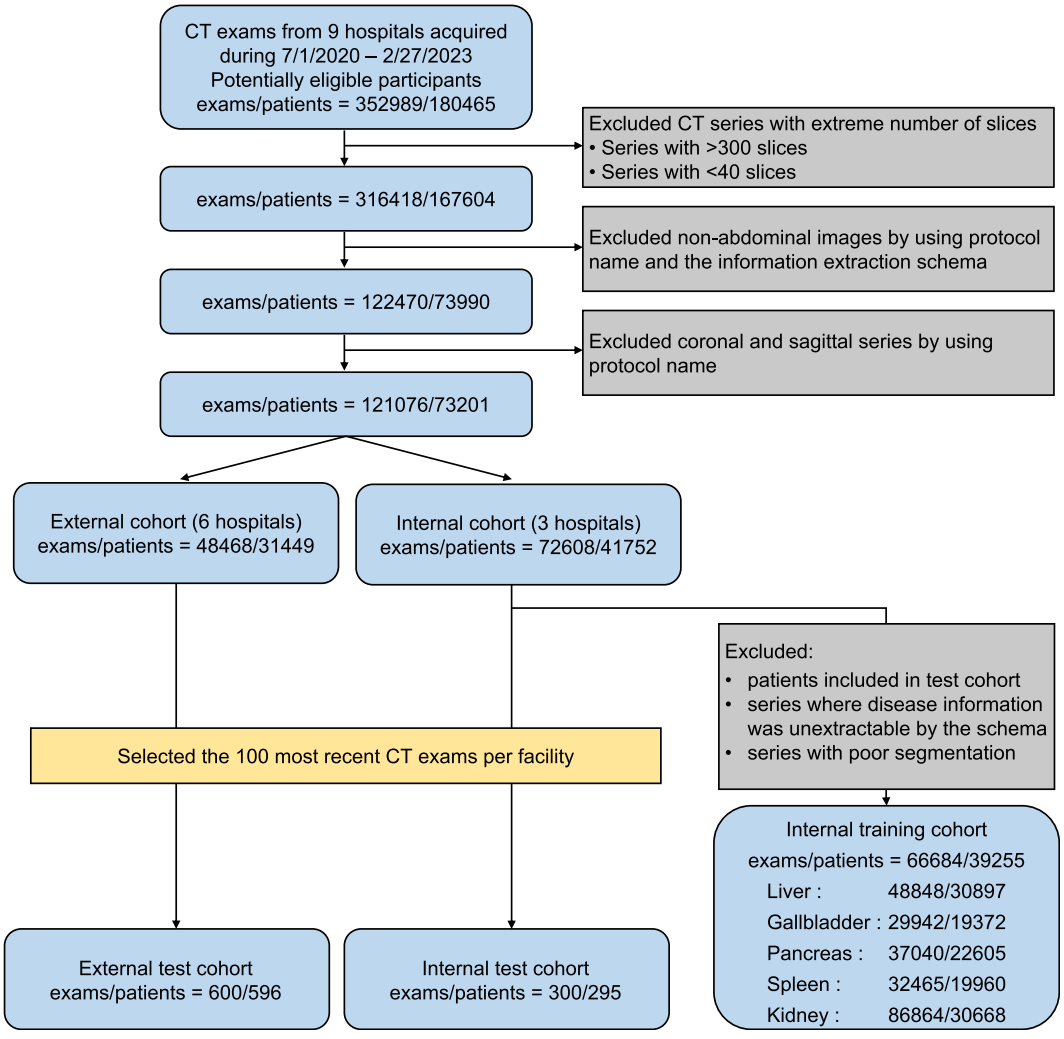


Fig. 1: Flowchart of patient enrolment. If some exams contained multiple series, such as non-contrast and contrast-enhanced images, all series that met the inclusion criteria were used. The kidney images were counted separately for the left and right sides.

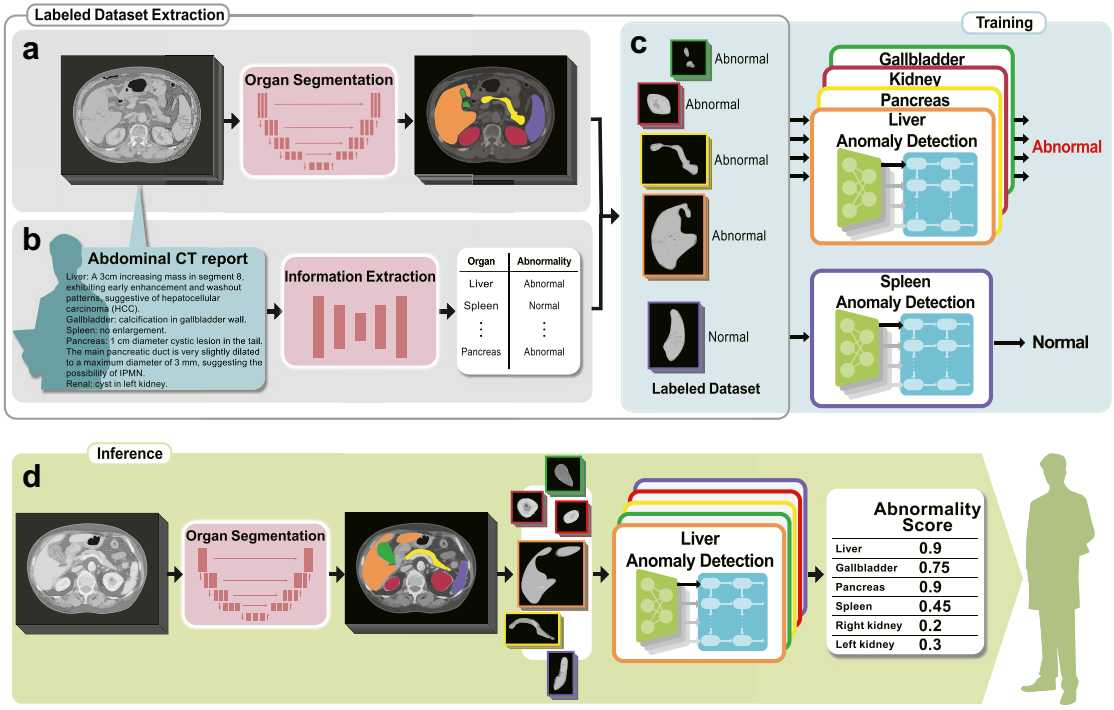


Fig. 2: Our anomaly detection pipeline using free-text radiology reports. (a) In the training process, abdominal computed tomography (CT) images are first input into a multi-organ segmentation model. (b) Radiology reports corresponding to the CT images are checked for the presence of disease in each organ using an information-extraction schema. (c) The extracted disease information is used as supervised labels for training the model with images of the organs. (d) The trained model computes anomaly scores for individual organs from images, offering diagnostic assistance.

3D CT images were input into a multi-organ segmentation model that extracted specific organ information from the entire abdominal image (Fig. 2a). The accompanying radiology reports were processed using an information extraction schema to determine the presence or absence of diseases in these organs (Fig. 2b). We then trained a model to detect diseases in each organ using organ-cropped images as input and the presence of a disease as a training label (Fig. 2c). During inference, combining the segmentation and anomaly detection models enabled processing of whole abdominal images to calculate abnormality scores for each organ, which could assist physicians in their diagnoses (Fig. 2d). These steps are detailed below.

Multi-organ segmentation model

Organ segmentation is effective for efficiently training anomaly detection models by extracting organs from large images, such as abdominal images. This approach has performed well in previous AI tasks for estimating cervical spine fractures.[15] We adopted a custom 3D full-resolution variant of nnUNet,[16] which enables accurate segmentation with big patch and batch sizes.[17] Training was performed using a batch size of 16 and a patch size of 288 × 288 × 64 pixels (width, height, and depth), with all other settings following the original nnUNet study parameters. The model was trained to segment 13 anatomical structures: the liver, gallbladder, pancreas, spleen, left and right kidneys, oesophagus, stomach, duodenum, aorta, left and right adrenal glands, bladder, and prostate/uterus. The model was trained and evaluated with 431 images from two sources: 300 images from AMOS,[18] a large-scale and diverse clinical dataset for abdominal organ segmentation (200 for training and 100 for testing), and 131 images from the Computational Anatomy Project dataset[19] (117 for training and 14 for testing).

Information extraction schema

A large amount of expert-annotated training data is required to develop deep-learning models. To reduce this requirement, we used an information extraction schema to reuse existing medical data. We hypothesised that the information on the presence of abnormal findings obtained from this schema could serve as a surrogate training label. As demonstrated in our previous study,[20] our information extraction schema consisted of two deep learning modules: entity extraction and relation extraction. The first module extracted three types of entities from the radiology reports: observation

entities, which represented observed abnormal features such as "nodule" or "pleural effusion"; clinical finding entities, which included diagnoses based on observations, such as "cancer"; and modifier entities, which described attributes such as anatomical location, certainty, change, characteristics, and size. These extracted entities were then fed into the second module, which predicted the relationships between them. Our model was pre-trained on 911,465 in-house reports and further developed using 1040 annotated reports (728 for training, 104 for validation, and 208 for testing). A certainty score assigned to each observation entity and clinical finding entity indicated the level of confidence in its presence, which ranged from 0 (definite absence) to 4 (definite presence). Absence of lesions (score 0) was categorised as "no finding," whereas any sign of potential lesions (scores 1–4) was classified as an "abnormal finding." The algorithm was described in detail in our previous paper.[21] Subsequently, patients with abnormal findings were categorised. Radiologists defined several disease categories for each organ and determined the most appropriate category for each extracted disease word. The correspondence table between words indicating abnormal findings and their respective categories is presented in the Data Sharing section.

Training protocol and preprocessing of anomaly detection model

Anomaly detection models were trained to identify organ abnormalities, with input of segmented 3D organ images, and output of disease categories. Such training was conducted for each organ through multiclass multi-label learning with labels generated by the information extraction schema (Supplementary Figure S1). 3D images were processed using multiple-instance learning,[22] input into a 2D encoder at set slice intervals, and information across slices was integrated to obtain the final class output. Organ-segmented 3D images, where non-organ voxels were filled with background values (−1000), were resized to 256 × 256 × 64 pixels in the left–right, antero–posterior, and cranio–caudal axes. From these 3D images, sets of five adjacent slices were extracted to form an image with five channels, which were then inputted into a 2D convolutional neural network (CNN) model in two-slice steps (0^{th}, 2nd, 4th, ...). The outputs were inputted to a Long Short-Term Memory network to share information between slices, with the final class output generated through fully connected layers and global average pooling. ConvNeXt v2[23] was used as a 2D CNN encoder. Window level and window width were set to 100 and 300, respectively, and normalised to a range between −1 and 1 before model input.

Training was performed over 12 epochs, using 5-fold cross-validation to select the model with the highest area under the receiver operating characteristic (ROC) curve (AUC) on the validation dataset as the final model for each fold. The final prediction of the test cohort was based on the average of the five models. The threshold for anomaly prediction was determined using the median of the thresholds that yielded the highest F1 scores across the validation dataset for each fold. Cross-entropy loss and the AdamW optimiser were employed with a learning rate of 0.00023 and a cosine-annealing learning-rate scheduler. The anomaly detection model was trained using a computing node with two CPUs and eight NVIDIA A100 graphics cards.

Ground-truth annotation

Each exam in our dataset included a free-text report written by a board-certified radiologist. Since histopathological diagnoses were not available for all images, radiologists' diagnoses were used as ground truth. First, to evaluate the accuracy of our information extraction schema on free-text radiology reports, abnormalities were extracted through review by a radiologist. Abnormalities were defined as findings previously documented as explicitly present. Indirect signs suggestive of abnormalities around the organ were not included. For the test cohort of 900 cases, 2 radiology residents reviewed the reports for each organ (considering the left and right kidneys separately) and listed the abnormalities. A third board-certified radiologist resolved any disagreements.

To evaluate the performance of our anomaly detection model, CT images were reviewed to create ground-truth labels. Although the exams already included reports from board-certified radiologists, another radiology resident double-checked them for diagnostic accuracy. In cases of discrepancies or judgment difficulties between the information extraction schema and the resident's review, a board-certified radiologist reviewed the images for a final decision. The ground-truth established here was based on abnormalities identified by radiologists from CT images, rather than on pathologically confirmed abnormalities. Images were annotated using the SYNAPSE SAI Viewer (FUJIFILM Corporation, Minato, Japan).

Statistical analysis

The segmentation model's performance was evaluated using the Dice similarity coefficient (DSC) and normalised surface Dice (NSD) score, as employed in AMOS. The AI system performance was assessed using the following metrics: AUC, accuracy, sensitivity, specificity, and F1 score. These metrics were calculated using Python (v3.9.12; https://www.python.org/downloads/release/python-3912/), NumPy (v1.23.2; https://numpy.org/), and scikit-learn (v1.0.2; https://scikit-learn.org/stable/whats_new/v1.0.html) packages. Confidence intervals for performance metrics were calculated as the 2.5th and 97.5th percentile of 1000 bootstraps, resampled with replacements from the test cohorts. The agreement

between our information extraction schema and the ground-truth labels for creating the training labels was calculated using Cohen's kappa coefficient. To evaluate the impact of dataset size on accuracy, Delong's test was performed. The code and abnormal class information used for the implementation of our anomaly detection pipeline are available at (https://github.com/jun-sato/sato_j-mid_ad).

Role of funders
The funders of the study had no role in the study design, data collection, data analysis, interpretation, or writing of the report.

Results

Training and evaluation of multi-organ segmentation models
We trained and evaluated a 3D-multi-organ segmentation model using 317 CT images from the AMOS and Computational Anatomy Project datasets for training and 114 images for validation. Boxplots for the DSC and NSD scores across the six organs for segmentation of the test data are presented in Fig. 3a and b, respectively. For all organs, the median DSC exceeded 90%, with the liver showing the highest DSC (0.981 [25th–75th percentile: 0.974–0.986]) and the pancreas showing the lowest DSC (0.911 [0.882–0.934]). The spleen had the highest NSD median value (0.934 [0.902–0.965]), whereas the pancreas had the lowest value (0.770 [0.706–0.831]). Representative examples of test data segmentation are shown in Fig. 3c. Detailed information on the segmentation results is provided in Supplementary Table S3.

Performance of the structured model on free-text radiology reports
We applied our information extraction schema to all free-text radiology reports accompanying included CT exams (Fig. 4a). Within the 900 exams used for our internal and external test cohorts, we examined the extent to which organ-specific abnormal findings, extracted using the information extraction schema,

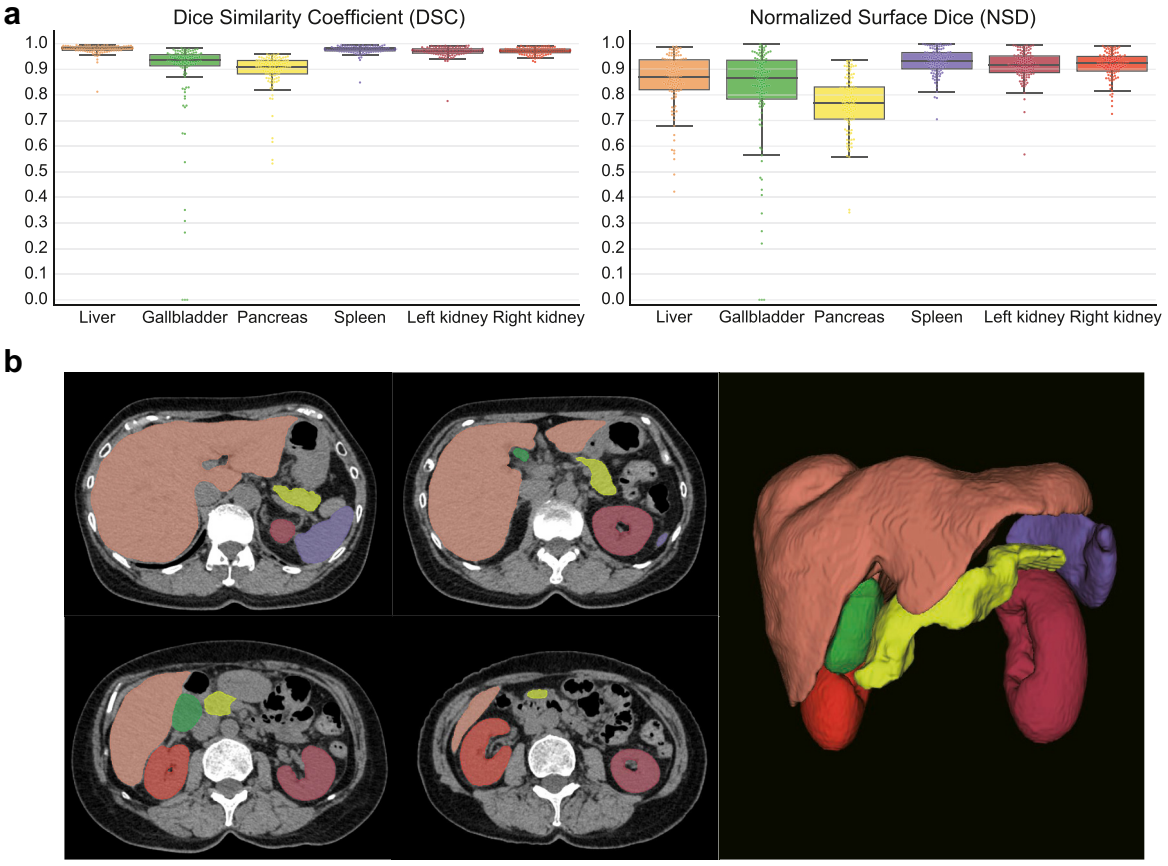
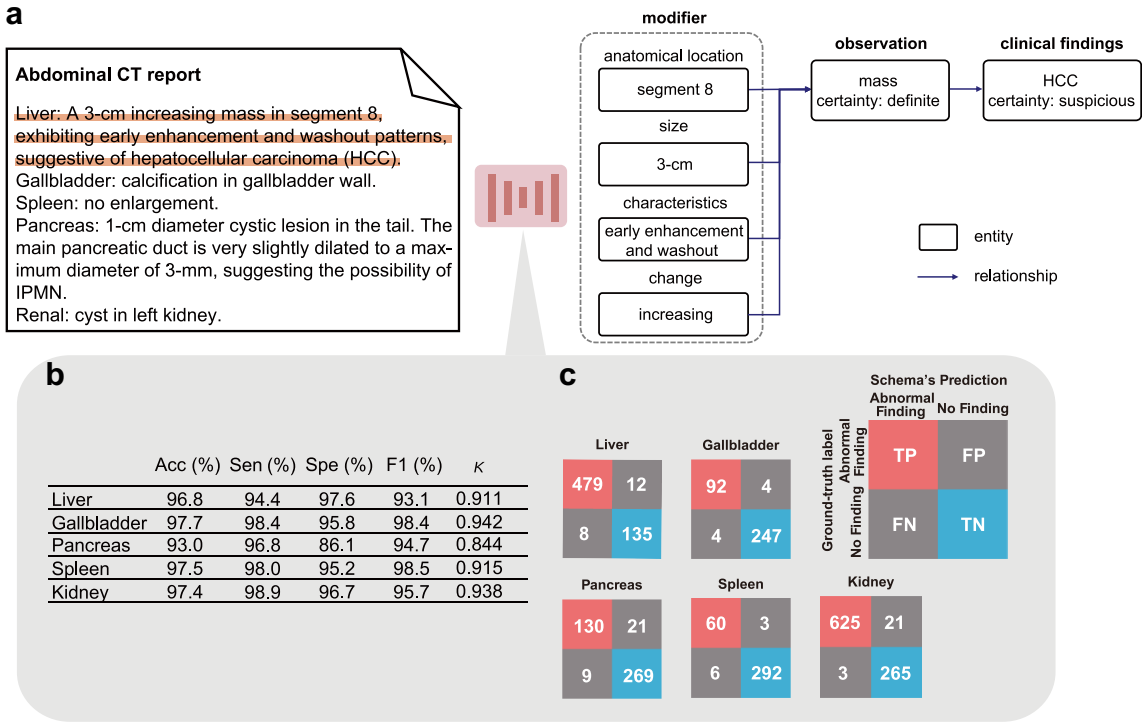


Fig. 3: Multi-organ segmentation performance plots and representative examples. (a) Box plots and swarm plots of Dice similarity coefficient (DSC) and normalised surface Dice (NSD) scores in our organ segmentation model. Box plots are defined as follows: the box's lower and upper bounds are represented by the first and third quartiles of the dataset, respectively, with a median line positioned at the centre. Whiskers stretch from the box up to a maximum of 1.5 times the interquartile range, and down to the minimum and up to the maximum data points lying within this range. (b) Axial computed tomography images and three-dimensional scans with mapped model predictions.



	Acc (%)	Sen (%)	Spe (%)	F1 (%)	κ
Liver	96.8	94.4	97.6	93.1	0.911
Gallbladder	97.7	98.4	95.8	98.4	0.942
Pancreas	93.0	96.8	86.1	94.7	0.844
Spleen	97.5	98.0	95.2	98.5	0.915
Kidney	97.4	98.9	96.7	95.7	0.938

Fig. 4: Representative example of the information-extraction schema applied to a free-text radiology report and illustration of the schema's performance. (a) Our schema extracts imaging findings, associated information, and suspected diseases from radiology reports. (b) Disease extraction performance per organ by the schema is evaluated using accuracy (Acc), sensitivity (Sen), specificity (Spe), F1 score (F1), kappa coefficient, and confusion matrix values. (c) These confusion matrices cross-tabulate the ground-truth labels, manually labelled by radiologists, against predictions from the information extraction schema. True positives (TP) and true negatives (TN) represent accurately predicted cases, while false positives (FP) and false negatives (FN) highlight misclassifications.

matched the ground-truth verified by at least two radiologists. We assessed the performance of the language model in binary classification by identifying the presence or absence of abnormal findings. The accuracy, sensitivity, specificity, F1 score, and confusion matrices of the information extraction schema are presented in Fig. 4b and c. The F1 score ranged between 93.1 and 98.5 across organs. The metrics for each institution are shown in Supplementary Table S4. The agreement with radiologist annotations was high (Cohen's kappa coefficients: 0.844–0.942).

Anomaly detection performance of our model

Our anomaly detection model was trained using images from 66,684 exams (39,255 patients) across three institutions (Fig. 1). This training cohort included 252,762 instances across all organs—48,848 liver instances (13,092 normal, 35,756 abnormal; 30,897 patients), 29,942 gallbladder instances (22,127 normal, 7815 abnormal; 19,372 patients), 37,040 pancreas instances (28,116 normal, 8924 abnormal; 22,605 patients), 32,465 spleen instances (28,069 normal, 4396 abnormal; 19,960 patients), and 86,864 kidney instances (23,546 normal, 63,318 abnormal; 38.822 patients). We validated the model's performance on an internal test cohort of 300 exams (295 individuals) across three institutions as well as an external test cohort of 600 exams (596 patients) from six hospitals. Details of patient characteristics and abnormal label ratios are listed in Tables 1 and 2.

The AUCs for the internal and external test cohorts are shown in Fig. 5. The average AUC for anomaly detection across organs in the external test cohort was 0.886: liver, 0.903 [95% CI: 0.877–0.929], gallbladder 0.898 [95% CI: 0.859–0.934], pancreas 0.838 [95% CI: 0.795–0.882], spleen 0.894 [95% CI: 0.845–0.938], and kidney, 0.898 [95% CI: 0.870–0.923]. The liver, gallbladder, and kidney had higher AUCs in the external cohort than in the internal cohort, with only a slight difference in the average AUC values between the internal (0.881) and external (0.886) cohorts. The accuracy, sensitivity, specificity, and F1 score for the internal and external test cohorts are shown in Table 3, and the precision–recall AUCs are shown in Supplementary Figure S2.

Performance changes in different ratios of training data

A major advantage of our pipeline was its ability to use a large amount of data, obtained using an information extraction schema, without requiring manual

	Liver	Gallbladder	Pancreas	Spleen	Kidney
Tokyo					
Exams	27,881	23,876	29,424	27,504	52,827
Patients	17,172	15,135	17,602	16,645	17,420
Abnormal label	17,177 (61.6)	4499 (18.8)	4634 (15.7)	2429 (8.8)	21,579 (40.8)
Sex					
Male	15,841 (56.8)	13,273 (55.6)	16,725 (56.8)	15,582 (56.7)	29,869 (56.5)
Female	12,040 (43.2)	10,603 (44.4)	12,699 (43.2)	11,922 (43.3)	22,958 (43.5)
Age, years	68 (55–77)	67 (54–76)	68 (55–76)	68 (55–76)	68 (54–76)
Keio					
Exams	9383	1541	2269	1058	14,088
Patients	6347	1157	1498	745	6016
Abnormal label	9178 (97.8)	1359 (88.2)	2014 (88.8)	917 (86.7)	13,782 (97.8)
Sex					
Male	4899 (52.2)	895 (58.1)	1330 (58.6)	617 (58.3)	8759 (62.2)
Female	4484 (47.8)	646 (41.9)	939 (41.4)	441 (41.7)	5329 (37.8)
Age, years	67 (56–76)	70 (59–78)	72 (63–79)	64 (51–73)	71 (61–78)
Okayama					
Exams	11,584	4525	5347	3903	19,949
Patients	7378	3080	3505	2570	7232
Abnormal label	9401 (81.2)	1957 (43.2)	2276 (42.6)	1050 (26.9)	12,717 (63.7)
Sex					
Male	6268 (54.1)	2477 (54.7)	2980 (55.7)	2172 (55.6)	11,488 (57.6)
Female	5315 (45.9)	2048 (45.3)	2367 (44.3)	1731 (44.4)	8459 (42.4)
Unknown	1				2
Age, years	68 (56–75)	67 (54–74)	69 (58–76)	67 (53–74)	69 (58–76)
Overall					
Exams	48,848	29,942	37,040	32,465	86,864
Patients	30,897	19,372	22,605	19,960	30,668
Abnormal label	35,756 (73.2)	7815 (26.1)	8924 (24.1)	4396 (13.5)	63,318 (60.6)
Sex					
Male	27,008 (55.3)	16,645 (55.6)	21,035 (56.8)	18,371 (56.6)	50,116 (57.7)
Female	21,839 (44.7)	13,297 (44.4)	16,005 (43.2)	14,094 (43.4)	36,746 (42.3)
Unknown	1				2
Age, years	68 (56–76)	67 (54–76)	69 (56–77)	68 (54–76)	69 (57–76)

Data are n, n (%), or median (IQR).

Table 1: **Patient characteristics in the internal training cohort.**

annotations, for training. We then evaluated the anomaly detection performance using varying amounts of training data. We randomly selected 300, 1,000, 2,000, 6,000, and 12,000 exams for training across five organs and compared the AUCs obtained. Additionally, we trained models with ground-truth labels provided by radiologists for 300 exams. The results of the ROC curves and AUC for the external test data are shown in Fig. 6. The models trained using 300 expert-labelled data points outperformed those trained using 300 auto-labelled data points. However, while there is a difference in dataset sizes, training models based on a larger dataset (≥2000 cases) with auto-extracted labels significantly outperformed models trained with human labels, despite inaccuracies in the training labels.

Discussion

Here, we developed a deep-learning-based pipeline to detect clinical abnormalities in five organs to aid clinicians in diagnosis. A multi-organ segmentation model processed 3D CT images to extract organ-specific regions, and an information extraction schema was used to analyse the accompanying radiology reports to identify disease information. We then trained the anomaly detection models using organ-cropped images as the input and the presence of a disease as the training label. Our pipeline achieved an AUC of 0.886 against expert-annotated ground-truth labels in an external test cohort derived from six institutions. To the best of our knowledge, no previous study has extensively used free-text radiology reports as training labels for CT image classification, and our study involved the largest data set

	Liver	Gallbladder	Pancreas	Spleen	Kidney
Internal					
Tokyo					
Exams	100	91	100	100	198
Patients	98	90	98	98	98
Abnormal label	52 (52.0)	22 (24.2)	88 (88.0)	8 (8.0)	89 (44.9)
Sex					
Male	63 (63.0)	58 (63.7)	63 (63.0)	63 (63.0)	124 (62.6)
Female	37 (37.0)	33 (36.3)	37 (37.0)	37 (37.0)	74 (37.4)
Age, years	69 (57–76)	68 (54–76)	69 (57–76)	69 (57–76)	69 (56–76)
Keio					
Exams	100	87	100	97	198
Patients	97	85	97	95	97
Abnormal label	65 (65.0)	28 (32.2)	25 (25.0)	9 (9.3)	137 (69.2)
Sex					
Male	55 (55.0)	45 (51.7)	55 (55.0)	52 (53.6)	108 (54.5)
Female	45 (45.0)	42 (48.3)	45 (45.0)	45 (46.4)	90 (45.5)
Age, years	68 (57–74)	64 (56–74)	68 (57–74)	67 (56–74)	69 (57–74)
Okayama					
Exams	100	89	99	99	198
Patients	100	89	99	99	100
Abnormal label	72 (72.0)	22 (24.7)	18 (18.2)	14 (14.1)	104 (52.5)
Sex					
Male	52 (52.0)	48 (53.9)	52 (52.5)	52 (52.5)	102 (51.5)
Female	48 (48.0)	41 (46.1)	47 (47.5)	47 (47.5)	96 (48.5)
Age, years	71 (53–77)	70 (53–77)	71 (53–77)	71 (53–77)	71 (53–77)
External					
Ehime					
Exams	100	88	100	100	193
Patients	100	88	100	100	100
Abnormal label	73 (73.0)	28 (31.8)	23 (23.0)	8 (8.0)	103 (53.4)
Sex					
Male	53 (53.0)	45 (51.1)	53 (53.0)	53 (53.0)	103 (53.4)
Female	47 (47.0)	43 (48.9)	47 (47.0)	47 (47.0)	90 (46.6)
Age, years	NA	NA	NA	NA	NA
Juntendo					
Exams	100	96	100	100	197
Patients	99	95	99	99	99
Abnormal label	64 (64.0)	11 (11.5)	10 (10.0)	5 (5.0)	98 (49.7)
Sex					
Male	62 (62.0)	60 (62.5)	62 (62.0)	62 (62.0)	122 (61.9)
Female	38 (38.0)	36 (37.5)	38 (38.0)	38 (38.0)	75 (38.1)
Age, years	63 (56–74)	63 (56–72)	63 (56–74)	63 (56–74)	63 (56–74)
Kyoto					
Exams	100	90	100	98	198
Patients	98	88	98	96	98
Abnormal label	78 (78.0)	27 (30.0)	19 (19.0)	5 (5.1)	115 (58.1)
Sex					
Male	56 (56.0)	51 (56.7)	56 (56.0)	54 (55.1)	111 (56.1)
Female	44 (44.0)	39 (43.3)	44 (44.0)	44 (44.9)	87 (43.9)
Age, years	72 (56–78)	73 (57–78)	72 (56–78)	73 (56–78)	71 (56–78)
Kyushu					
Exams	100	82	100	99	197
Patients	100	82	100	99	100
Abnormal label	61 (61.0)	13 (15.9)	26 (26.0)	12 (12.1)	99 (50.3)

(Table 2 continues on next page)

	Liver	Gallbladder	Pancreas	Spleen	Kidney
(Continued from previous page)					
Sex					
Male	58 (58.0)	49 (59.8)	58 (58.0)	58 (58.6)	116 (58.9)
Female	42 (42.0)	33 40.2)	42 (42.0)	41 (41.4)	81 (41.1)
Age, years	66 (52–71	65 (49–71)	66 (52–71)	66 (52–71)	65 (50–71)
Osaka					
Exams	100	85	100	95	195
Patients	99	84	99	94	99
Abnormal label	73 (73.0)	23 (27.1)	33 (33.0)	13 (13.7)	108 (55.4)
Sex					
Male	64 (64.0)	55 (64.7)	64 (64.0)	61 (64.2)	124 (63.6)
Female	36 (36.0)	30 (35.3)	36 (36.0)	34 (35.8)	71 (36.4)
Age, years	68 (59–77)	66 (59–76)	68 (59–77)	77 (59–77)	68 (59–77)
Tokushima					
Exams	100	92	100	100	198
Patients	100	92	100	100	100
Abnormal label	63 (63.0)	25 (27.2)	25 (25.0)	12 (12.0)	124 (62.6)
Sex					
Male	65 (65.0)	59 (64.1)	65 (65.0)	65 (65.0)	130 (65.7)
Female	35 (35.0)	33 (35.9)	35 (35.0)	35 (35.0)	68 (34.3)
Age, years	71 (59–78)	71 (57–77)	71 (59–78)	71 (59–78)	71 (59–78)
Overall exams					
Exams	900	800	899	888	1772
Patients	891	793	890	880	891
Abnormal label	603 (67.0)	199 (24.9)	191 (21.2)	86 (9.7)	977 (55.1)
Sex					
Male	524 (58.2)	463 (57.9)	523 (58.2)	515 (58.0)	1034 (58.4)
Female	376 (41.8)	337 (42.1)	376 (41.8)	373 (42.0)	738 (41.6)
Age, years	69 (56–76)	67 (56–76)	68 (56–76)	68 (56–76)	68 (56–76)

Data are n, n (%), or median (IQR).

Table 2: **Patient characteristics in the internal and external test cohort.**

and number of anatomical sites analysed in this way to date.

Previous studies using reports as labels have primarily focused on chest X-rays.[10–12] Although a large amount of data for such 2D images is publicly available, these methods cannot be directly applied to 3D CT images because of differences in size and model structure. Only two studies to date have used reports as labels for CT images: one focused on the head region, with images collected from two institutions,[13] while the other targeted two abdominal regions (liver/gallbladder and kidney), using a total of 9153 images from a single institution, which yielded an anomaly-detection performance of at least 7% lower than that of our pipeline.[24]

Abnormality detection in abdominal CT images is challenging because these scans provide high-resolution images of large areas including multiple organs. These images cannot be directly used in AI due to their size. Reducing their size compromises the details while dividing them into small patches can prevent the model from recognising structural abnormalities across all organs. Radiology reports detailing multiple organs reflect this complexity. To address these challenges, our pipeline employed segmentation models and an information extraction schema to identify and analyse these critical but small regions accurately across the organ spectrum. With the increasing availability of various open-source datasets for organ segmentation,[25] applying our method to other organs might enable more extensive anomaly detection.

We demonstrated that the models trained on a large amount of data labelled by a language model outperformed those trained on smaller datasets labelled by experts, as shown in Fig. 6. Additionally, compared with a previous study that used reports as training labels,[24] our pipeline showed at least a 7% improvement in the AUC for organ-specific abnormality detection. This improvement may be due to the accuracy of the information extraction schema. A previous study on positron-emission tomography/CT applied information extraction algorithms to create labelled datasets from radiology reports to train an anomaly-detection model,[26]

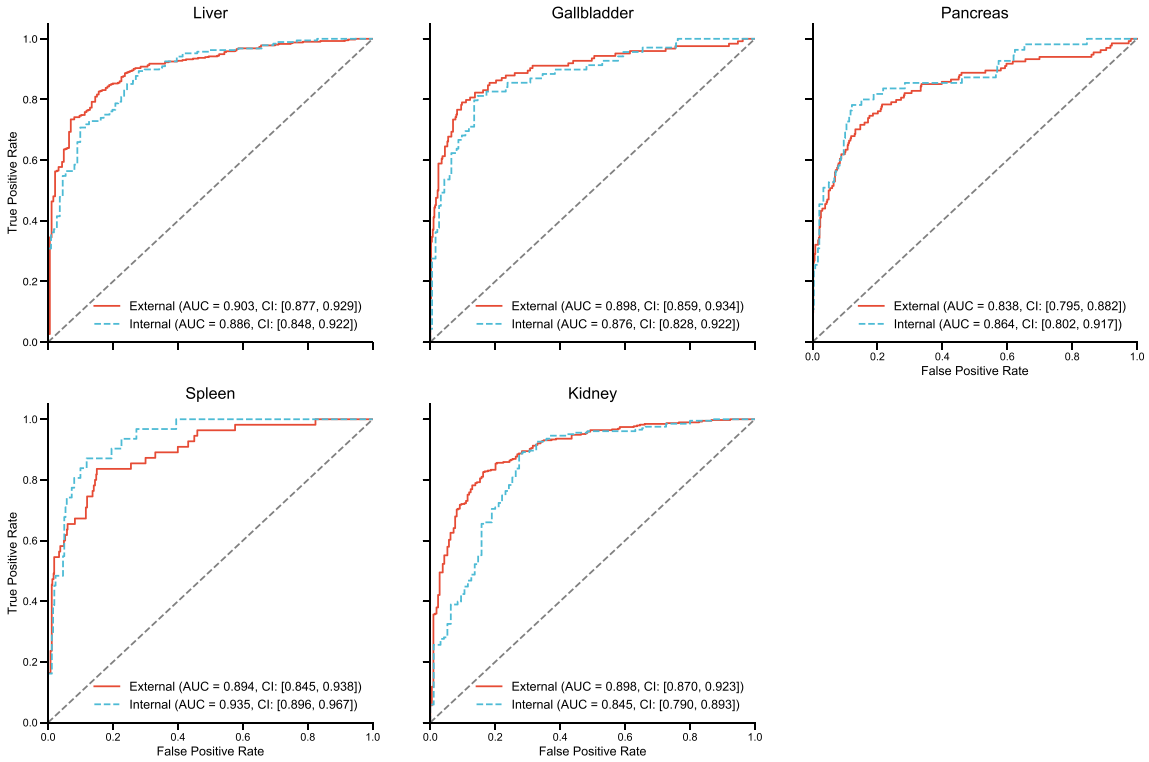


Fig. 5: Receiver operating characteristic curves for the anomaly detection models for each abdominal organ. Receiver operating characteristic curves for each organ in the internal and external test cohorts. The curve presents the true positive rate (sensitivity) and false positive rate (1–specificity) across different cutoffs. The values in each graph represent the areas under the receiver operating characteristic curves (AUCs) and their 95% confidence intervals (CIs) for each cohort.

and achieved a maximum F1 score of 0.888 in label creation. Our schema yielded higher F1 scores across all organs, suggesting that the automatic extraction of information from reports contributed to improved abnormality detection. Moreover, using or combining a wide range of encoder models, such as Transformers,[27] could further improve the accuracy of anomaly detection.

	Accuracy, %	Sensitivity, %	Specificity, %	F1 score, %
Liver				
External	84.8 (82.0–87.5)	88.9 (85.7–91.9)	75.8 (70.0–81.9)	89.0 (86.8–91.1)
Internal	81.6 (77.3–85.6)	92.6 (88.5–95.9)	63.1 (54.1–71.8)	86.4 (82.5–89.5)
Gallbladder				
External	86.6 (83.4–89.5)	79.0 (71.8–85.8)	89.0 (85.7–92.0)	73.7 (67.1–79.4)
Internal	83.5 (79.5–87.8)	81.2 (71.7–89.9)	84.3 (79.4–89.4)	72.7 (64.9–79.8)
Pancreas				
External	81.7 (78.7–84.6)	71.6 (63.7–78.6)	84.6 (81.4–87.8)	63.8 (56.9–69.7)
Internal	79.6 (74.8–84.4)	81.8 (70.8–91.8)	79.1 (73.8–84.0)	60 (49.6–68.9)
Spleen				
External	75.2 (71.6–78.6)	85.5 (75.9–94.2)	74.1 (70.1–77.7)	39.2 (31.3–47.1)
Internal	73.6 (68.5–78.8)	96.8 (89.3–1)	70.9 (65.3–76.3)	43.8 (33.0–53.8)
Kidney				
External	76.8 (73.6–80.1)	68.7 (64.1–73.2)	91.9 (88.4–95.5)	79.4 (76.1–82.7)
Internal	76.5 (71.8–81.2)	76.4 (70.6–82.0)	76.8 (67.8–84.8)	81.6 (77.4–85.3)

Data are % (95% confidence interval).

Table 3: **Anomaly detection performance metrics of the model for internal and external test cohorts.**

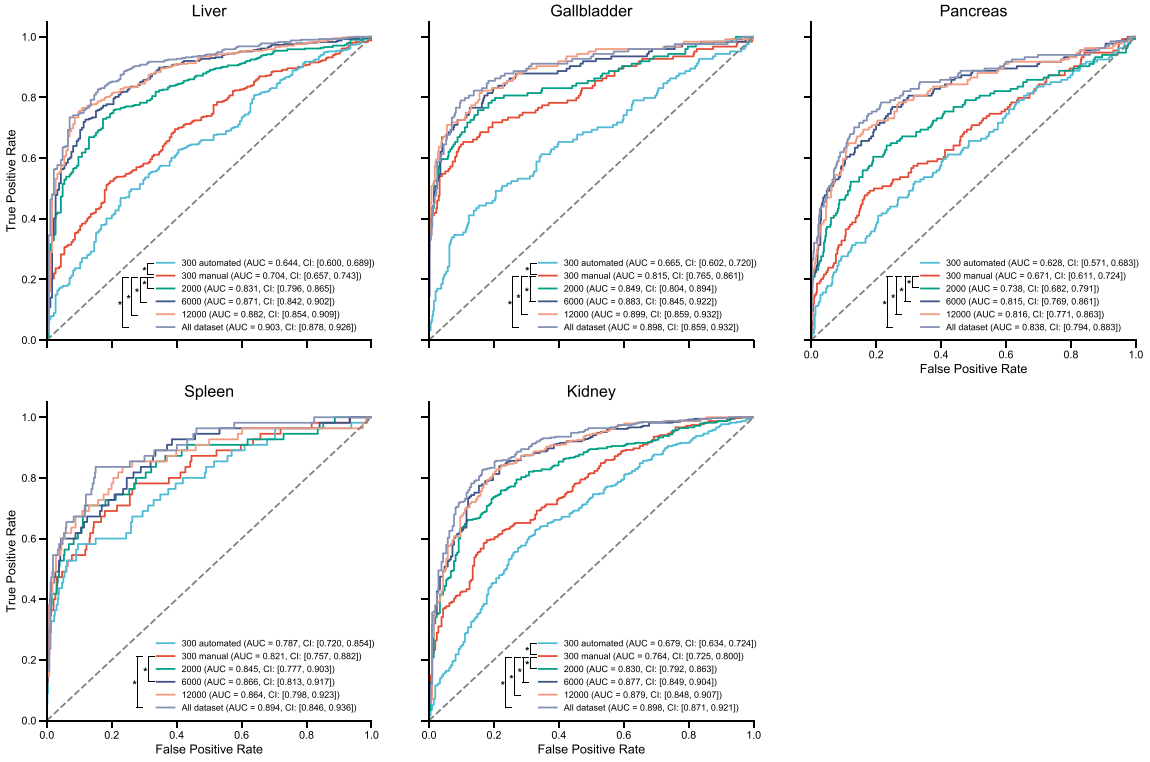


Fig. 6: Impact of training data sizes on receiver operating characteristic curves. The receiver operating characteristic curve for each organ in the external test cohort illustrates the impact of varying training data sizes on the anomaly detection performance. The curve presents the true positive rate (sensitivity) and false positive rate (1–specificity) across different cutoffs. The values in each graph represent the areas under the receiver operating characteristic curves (AUCs) and their 95% confidence intervals (CIs). *p < 0.05.

Extracting information from free-text radiology reports by employing deep-learning models, such as BERT,[10,11] has proven superior to traditional rule-based methods.[7,28] For reports with limited content, such as those for chest X-rays, direct input of the entire text into the model is sufficient for image-level classification.[10–12] However, a different approach is required when identifying abnormalities in multiple organs, as in our case. In our approach, we used a language model to extract organ-specific abnormalities to create organ-specific training datasets and to extract information about the characteristics and longitudinal changes in abnormalities, which could facilitate improved accuracy in future. Additionally, our schema was language-independent, enabling its application to report in various languages. With the emergence of large language models and the growing interest in leveraging existing medical data, our method could be clinically applicable.

Generalisability is a major challenge in deep learning of medical images. Due to patient privacy and copyright concerns, data sharing is restricted, and we must often rely on datasets from a limited number of institutions. This scarcity of diverse datasets negatively affects model training and validation, leading to overfitting and good performance of models only on familiar images.[29] While most available image data with accompanying radiological reports pertain to chest X-rays,[7,30] no publicly available datasets of abdominal CT images with accompanying reports exist. Testing datasets from various institutions helps to avoid overfitting and ensures robustness and generalisability. We collected data from three institutions for our internal cohort and from six institutions for our external cohort, demonstrating the stable anomaly detection capabilities of our model across different imaging environments.

Our pipeline offers several advantages for clinical applications. First, the end-to-end nature of our approach allows for the direct input of clinically acquired images, eliminating the need for pre-processing or manual annotation. This streamlines the workflow and significantly reduces development costs for clinical implementation. Moreover, the high AUC of our model addresses the challenge of increased imaging exams by reducing radiologist reading times. For example, AI for detecting breast cancer in digital breast tomosynthesis[31] and lung nodule detection in chest radiographs,[5] with

AUCs of 0.840 and > 0.9, respectively, significantly shortened reading times. These examples demonstrate the potential of diagnostic support systems like ours to meet the increasing demand for efficient and accurate medical imaging interpretation.

To advance our pipeline towards clinical application, several improvements can be made. First, the performance of our organ segmentation could benefit from training on large-scale datasets like TotalSegmentator.[25] While annotated data is necessary for training segmentation models, using these larger or combined publicly available datasets may allow the model to better adapt to the diverse CT imaging conditions encountered in clinical practice without the need for additional annotation. Additionally, while our information extraction schema shows strong capabilities, variability in reporting styles across institutions and radiologists affects its consistency. Recent advancements in large language models, which can process large volumes of text with accuracy comparable to humans, may help standardise reporting styles and improve extraction performance. The anomaly detection model is also expected to enhance interpretability for clinicians. While our model predicts the presence of abnormalities to prompt careful image review, it cannot provide detailed information about them. Future improvements should focus on recognising details like the type and exact location of abnormalities to enhance clinical understanding. Furthermore, vision-language foundation models provide a more direct way to link images and reports. Currently, processing high-resolution 3D CT images and their disease remains challenging due to variability in both imaging and reporting, as well as computational limitations. However, with more data, improved deep learning models, and enhanced computational power, these models could support the detection of a broader range of diseases.

Our study has several limitations. Retrospective data collection from multiple institutions suggested the need for further validation using prospectively collected test datasets across different disease prevalence rates. The training and evaluation of our model relied on diagnoses made by experienced radiologists, rather than final pathological diagnoses, which could misrepresent the actual abnormalities. A selection bias might also have been present, influenced by the protocol names, results of our information extraction schema, and the exclusion of series with more than 300 or fewer than 40 images. Additionally, exclusion of patients' historical imaging data disallowed leveraging of temporal changes for anomaly detection, potentially limiting the model's ability to reflect real-world clinical scenarios.

In conclusion, we developed a deep-learning-based pipeline encompassing labelled dataset creation, model training, and anomaly detection using CT images and associated free-text radiology reports. The learning process was streamlined by eliminating the need for manual annotations. Our pipeline was trained on a diverse dataset containing 252,762 imaging instances for five organs, which were collected from multiple institutions, and demonstrated high anomaly-detection capabilities for every organ examined. Our approach is broadly applicable across different anatomical sites and diseases, heralding significant advances in computer-aided diagnosis.

Contributors

All authors contributed to the conception and design of this study. JS and YS collected, verified, and analysed the data, and provided access to the raw data. The ground-truth labels of the test data were created by JS, TW, DN, MT, and YH. KS and TT created the information extraction schema. JS wrote the first draft of the manuscript, which was then edited by YS, KS, and SK. All the processes were supervised by YS, TT, MH, SK, and NT. All authors had access to all data, and read and approved the final manuscript. SK was responsible for the decision to submit the manuscript for publication.

Data sharing statement

All codes associated with pipeline development are shared on GitHub (https://github.com/jun-sato/sato_j-mid_ad). The pretrained models for both the multiorgan segmentation module and the anomaly detection module are available from JS. CT images are not available for sharing at this time, to protect the privacy of the participants.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

This work was supported by the Japan Science and Technology Agency SPRING (grant number JPMJSP2138). We utilised the computational resources of SQUID provided by Osaka University through the HPCI System Research Project (Project ID: hp230031). During the preparation of this work the authors used ChatGPT in order to grammatically review the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105463.

References

1 Computed tomography (CT) exams. *Health care use*; 2017. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/computed-tomography-ct-exams/indicator/english_3c994537-en.
2 McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol*. 2015;22:1191–1198.
3 Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ*. 2017;359:j4683.
4 Hanna TN, Lamoureux C, Krupinski EA, Weber S, Johnson J-O. Effect of Shift, Schedule, and volume on interpretive accuracy: a retrospective analysis of 2.9 million radiologic examinations. *Radiology*. 2018;287:205–212.
5 Shin HJ, Han K, Ryu L, Kim E-K. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *NPJ Digit Med*. 2023;6:82.
6 Pyrros A, Borstelmann SM, Mantravadi R, et al. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nat Commun*. 2023;14:4039.
7 Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc Conf AAAI Artif Intell*. 2019;33:590–597.
8 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28:31–38.
9 van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for

research: an international comparative study. *BMC Med Inform Decis Mak*. 2016;16:90.
10 Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng*. 2022;6:1399–1406.
11 Zhou H-Y, Chen X, Zhang Y, Luo R, Wang L, Yu Y. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nat Mach Intell*. 2022;4:32–40.
12 Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. 2020. Contrastive learning of medical visual representations from paired images and text. In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. PMLR; 2022:2–25.
13 Liu A, Guo Y, Lyu J, et al. Automatic intracranial abnormality detection and localization in head CT scans by learning from free-text reports. *Cell Rep Med*. 2023;4:101164.
14 Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029.
15 Lee GR, Flanders AE, Richards T, et al. Performance of the winning algorithms of the RSNA 2022 cervical spine fracture detection challenge. *Radiol Artif Intell*. 2024;6:e230256.
16 Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203–211.
17 Sato J, Kido S. Large batch and patch size training for medical image segmentation. *arXiv [eess.IV]*; 2022. Available from: http://arxiv.org/abs/2210.13364.
18 Ji Y, Bai H, Yang J, et al. AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv [eess.IV*; 2022. Available from: http://arxiv.org/abs/2206.08023.
19 Watanabe H, Shimizu A, Umetsu S, Kobatake H, Nawano S. Semiautomated organ segmentation using 3-dimensional medical imagery through sparse representation. *Trans Jpn Soc Med Biol Eng*. 2013;51:300–312.
20 Sugimoto K, Wada S, Konishi S, et al. Extracting clinical information from Japanese radiology reports using a 2-stage deep learning approach: algorithm development and validation. *JMIR Med Inform*. 2023;11:e49041.
21 Sugimoto K, Wada S, Konishi S, et al. Classification of diagnostic certainty in radiology reports with deep learning. *Stud Health Technol Inf*. 2024;310:569–573.
22 Zhou Z-H. A brief introduction to weakly supervised learning. *Natl Sci Rev*. 2018;5:44–53.
23 Woo S, Debnath S, Hu R, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2023:16133–16142.
24 Tushar FI, D'Anniballe VM, Hou R, et al. Classification of multiple diseases on body ct scans using weakly supervised deep learning. *Radiol Artif Intell*. 2022;4:e210026.
25 Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. 2023;5:e230024.
26 Eyuboglu S, Angus G, Patel BN, et al. Multi-task weak supervision enables anatomically resolved abnormality detection in whole-body FDG-PET/CT. *Nat Commun*. 2021;12:1880.
27 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:10012–10022.
28 Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *arXiv [csCL]*; 2018. https://arxiv.org/abs/1712.05898. AMIA Jt Summits Transl Sci Proc 2018; 2017: 188–196.
29 Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health*. 2019;1:e232–e242.
30 Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6:317.
31 van Winkel SL, Rodríguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. *Eur Radiol*. 2021;31:8682–8691.