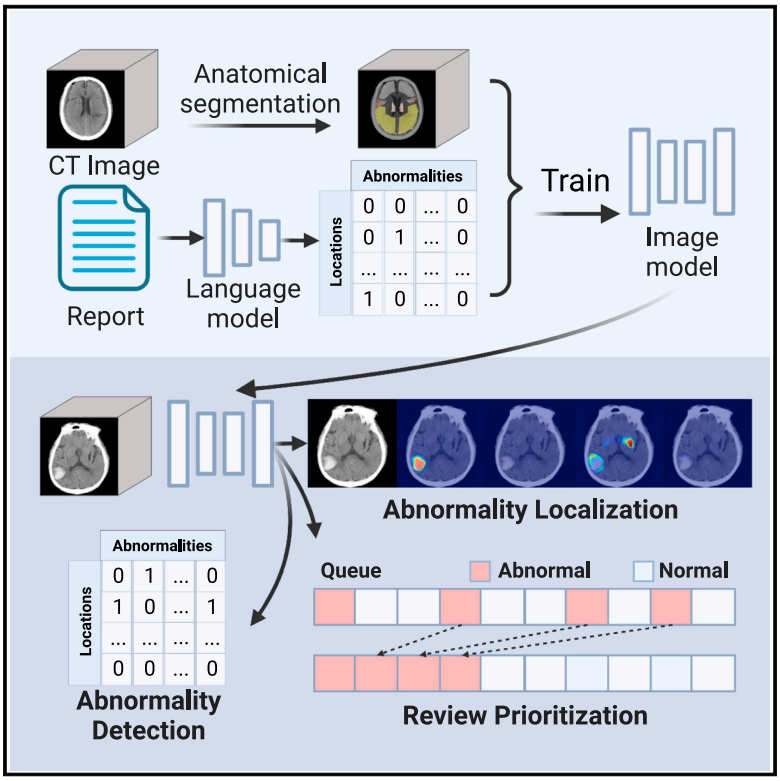
Cell Reports
Medicine

Article

Automatic intracranial abnormality detection and localization in head CT scans by learning from free-text reports

Graphical abstract



Highlights

- Massive data labeled using imaging reports with low annotation cost
- Model detects 4 abnormality types in 17 anatomical regions with high accuracy
- Multi-instance learning enables voxel-level prediction from coarse-grained labels
- Abnormality prediction helps prioritization of scans with various diseases

Authors

Aohan Liu, Yuchen Guo, Jinhao Lyu, ..., Xin Lou, Jun-hai Yong, Qionghai Dai

Correspondence

yuchen.w.guo@gmail.com (Y.G.), xufeng2003@gmail.com (F.X.), louxin@301hospital.com.cn (X.L.), yongjh@tsinghua.edu.cn (J.-h.Y.), qhdai@tsinghua.edu.cn (Q.D.)

In brief

Liu et al. propose a cross-modality deep-learning framework that utilizes free-text imaging reports for intracranial abnormality detection and localization in head CT scans. The learned model can detect, classify, and localize abnormalities at the voxel level and also help downstream tasks including disease classification and review prioritization.

Liu et al., 2023, Cell Reports Medicine *4*, 101164
September 19, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.xcrm.2023.101164

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Automatic intracranial abnormality detection and localization in head CT scans by learning from free-text reports

Aohan Liu,[1,3] Yuchen Guo,[3,*] Jinhao Lyu,[2] Jing Xie,[5,6] Feng Xu,[1,3,7,*] Xin Lou,[2,*] Jun-hai Yong,[1,*] and Qionghai Dai[3,4,*]
[1]School of Software, Tsinghua University, Beijing 100084, China
[2]Department of Radiology, Chinese PLA General Hospital, Beijing 100039, China
[3]Institute for Brain and Cognitive Sciences, BNRist, Tsinghua University, Beijing 100084, China
[4]Department of Automation, Tsinghua University, Beijing 100084, China
[5]Hangzhou Zhuoxi Institute of Brain and Intelligence, Hangzhou, Zhejiang 311100, China
[6]Hanyi Technology (Hangzhou) Co., Ltd., Hangzhou, Zhejiang 311121, China
[7]Lead contact
*Correspondence: yuchen.w.guo@gmail.com (Y.G.), xufeng2003@gmail.com (F.X.), louxin@301hospital.com.cn (X.L.), yongjh@tsinghua.edu.cn (J.-h.Y.), qhdai@tsinghua.edu.cn (Q.D.)
https://doi.org/10.1016/j.xcrm.2023.101164

SUMMARY

Deep learning has yielded promising results for medical image diagnosis but relies heavily on manual image annotations, which are expensive to acquire. We present Cross-DL, a cross-modality learning framework for intracranial abnormality detection and localization in head computed tomography (CT) scans by learning from free-text imaging reports. Cross-DL has a discretizer that automatically extracts discrete labels of abnormality types and locations from reports, which are utilized to train an image analyzer by a dynamic multi-instance learning approach. Benefiting from the low annotation cost and a consequent large-scale training set of 28,472 CT scans, Cross-DL achieves accurate performance, with an average area under the receiver operating characteristic curve (AUROC) of 0.956 (95% confidence interval: 0.952–0.959) in detecting 4 abnormality types in 17 regions while accurately localizing abnormalities at the voxel level. An intracranial hemorrhage classification experiment on the external dataset CQ500 achieves an AUROC of 0.928 (0.905–0.951). The model can also help review prioritization.

INTRODUCTION

Deep learning (DL) has recently yielded promising results for the analysis of various medical imaging modalities,[1–5] demonstrating its potential for improving diagnosis accuracy and optimizing the healthcare workflow.[6] Head computed tomography (CT) is widely used in clinical routine for diagnosing a variety of intracranial diseases and abnormalities, benefiting from its high efficiency and wide accessibility.[7,8] The automatic detection and localization of abnormalities in head CT scans based on DL are crucial for helping radiologists to make faster and more accurate diagnoses.

One key feature of DL models is that they are capable of learning complicated knowledge from massive (e.g., tens of thousands of) labeled data.[9–11] Analogously in medical practice, a large labeled training dataset is a fundamental component for developing accurate, generalizable, and interpretable DL systems. First, abnormalities and diseases have complicated patterns in medical images, and the appearance of images may vary due to different acquisition machines, settings, and operators. Thus, large labeled training datasets are required so that comprehensive knowledge about the patterns and appearance diversity can be learned to ensure the model's accuracy and generalizability. Second, a medical system should provide not only the final prediction for disease types but also details such as the precise locations of lesions to better integrate with human decisions, deal with potential false predictions, and make the final diagnosis more accurate and convincing. However, although spatial clues could be generated using attention-based or gradient-based approaches[12–14] to some extent, accurate localization still requires annotation with spatial information, such as voxel-level segmentation annotations.

The acquirement of annotations for large-scale medical datasets, however, creates substantial obstacles to developing DL systems for medical usage. The major difficulties lie in privacy concerns and the expertise required for manual annotation,[15,16] making it quite expensive, time consuming, and sometimes infeasible to annotate large datasets. This problem becomes more serious when the system is expected to perform pixel-/voxel-level analysis, such as lesion segmentation in 3D CT scans, where it is estimated that it will take more than 5 min for an expert to annotate a single CT scan. Therefore, innovating new algorithms to build accurate, generalizable, and interpretable DL systems for medical practice based on large training

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

datasets with low annotation costs is necessary and can facilitate broader applications of DL in clinical practice.

In this study, we address the above problems by proposing a cross-modality DL (Cross-DL) framework for accurate, generalizable, and interpretable intracranial abnormality detection and localization in head CT scans. Although it is difficult and expensive to obtain manual annotations for CT scans, imaging reports documented by radiologists in their routine work are readily available and informative and could be used as a source of annotation. Benefiting from the low cost and easy accessibility, we constructed a large training dataset with massive CT-report pairs. Cross-DL solved three fundamental problems to learn an image model from free-text reports. (1) What information should be extracted and learned from the free-text reports? Previous works mostly aim to directly learn disease types.[4,17] However, in practice, disease types usually cannot be fully determined without multi-modality information including images, clinical symptoms, and additional examination (e.g., examinations for tumors). Instead, the conclusions that can be solely determined by CT images are the types of abnormalities and their locations, which are usually documented in the "findings" parts of imaging reports and are the prerequisite for radiologists to make the final disease diagnosis. (2) How should useful information from free-text reports be accurately extracted? Different reports may have different linguistic expressions, making it difficult to extract labels using keywords or conventional approaches such as regular expression matching. Cross-DL proposes a discretizer module with natural language processing (NLP) techniques to handle the linguistic variations in different reports. (3) How should a DL model with voxel-level outputs be trained using only discrete annotations extracted from the reports? Cross-DL solves this problem by proposing a dynamic multi-instance learning (DaMIL) approach, enabling the model to not only detect the existence of different types of abnormalities but also to localize them at the voxel level in 3D CT scans.

By addressing the above problems, our framework has the following advantages. Firstly, the utilization of free-text reports enables automatic and fine-grained annotation for large-scale datasets to build DL models in a practical and cost-effective way. Secondly, the detection of abnormalities is generalizable to various diseases, making the model more broadly applicable. Thirdly, this system is capable of providing accurate spatial locations of abnormalities to improve system interpretability and support further expert diagnoses.

With these technical innovations, Cross-DL achieves an accurate, generalizable, and interpretable performance for multi-type intracranial abnormality detection and localization in head CT scans. The system and methodology can facilitate fast and reliable brain abnormality discovery and diagnosis in emergencies to prompt appropriate treatment and mitigate neurological deficit and mortality. We believe that the Cross-DL framework can inspire more studies to build clinically applicable DL systems in the future.

RESULTS

Dataset collection and annotation

To develop and evaluate Cross-DL, we collected a retrospective CT-report dataset from the Chinese PLA General Hospital (PLAGH) containing 31,778 head CT scans and corresponding reports of 18,125 patients between April 2012 and July 2019. The patients in this dataset have an average age of 54.3 years and a gender ratio (male:female) of 1.28 (see Table 1 for more details). All reports were written in Chinese. The retrospective dataset was randomly divided by patient into a training set with 28,472 CT scans of 16,316 patients, a validation set with 1,699 CT scans of 902 patients, and a test set with 1,607 CT scans of 907 patients. We also constructed a prospective test dataset by retrieving 1,500 CT scans and corresponding reports of 1,208 patients between August 2019 and August 2021 for generalizability evaluation.

As our framework uses a discretizer to extract types and regional locations of abnormalities from free-text reports, we constructed a report dataset to train and evaluate the discretizer using a small portion (3.83%) of reports in the retrospective dataset. Specifically, we extracted 4,807 sentences from the "findings" sections of the reports in the retrospective dataset and randomly divided them into a training set of 3,000 sentences, a validation set of 904 sentences, and a test set of 903 sentences.

In this study, we involved three types of abnormality annotations to describe the types and regional locations of abnormalities. The abnormality annotations were represented as binary matrices where each element of a matrix indicated whether a certain type of abnormality existed in a certain anatomical region. "Gold-standard" annotations were labeled by expert radiologists using both the CT scans and the corresponding reports. They were used in the retrospective test dataset (1,607 scans) and the prospective test dataset (1,500 scans). These annotations require expertise but were only used for the final evaluation and not in system development. "Silver-standard" annotations were labeled by non-expert annotators using only imaging reports. They were used in the report dataset (4,807 sentences) for the development and validation of the discretizer module. Experiments showed that our discretizer could work well using small-scale, silver-standard annotations. "Pseudo" annotations were generated by the discretizer from the reports. They were used in the retrospective training and validation datasets for system development and could be obtained fully automatically without involving additional labeling efforts. Besides the abnormality annotations, we used coarse anatomical region segmentation generated by an atlas-based segmentation method[18] for system development, which only required manual segmentation on one CT scan by an expert.

The statistics of the retrospective and prospective datasets are shown in Table 1, along with the distributions of the abnormalities in different anatomical regions in the retrospective test dataset and the prospective dataset based on the gold-standard annotations.

For further evaluation of our system's capability for review prioritization, we constructed a preoperative dataset from the prospective dataset by removing all postoperative scans, resulting in 1,204 CT scans, and then labeling the existence of 12 common diseases in each scan. We also used the publicly available CQ500 dataset[1] consisting of 491 head CT scans with manual annotations (hemorrhage and other diseases) for external evaluation.

CellPress
OPEN ACCESS

Table 1. The statistics of the internal dataset

	Retrospective					Prospective				
No. CT scans	31,778					1,500				
No. patients	18,125					1,208				
Time	April 2012 to July 2019					August 2019 to August 2021				
Patient age (years)[a]	54.3 (41, 69)					55.4 (41, 70)				
Patient gender (male) (%)	56.2					57.3				
CT scanner (%)	Emotion 16 (49.3); Optima CT660 (24.9); uCT 510 (14.3); Emotion 6 (8.6); Sensation Cardiac 64 (1.9); Revolution CT (0.3); SOMATOM Definition (0.2); unknown (0.5)					uCT 510 (57.2); Optima CT660 (34.5); iCT 256 (3.9); Discovery CT750 HD (2.5); LightSpeed VCT (1.8); Revolution CT (0.1)				
Pixel spacing[a] (mm)	0.488 (0.473, 0.488)					0.469 (0.449, 0.488)				
Slice thickness[a] (mm)	4.20 (4.00, 4.80)					4.84 (4.80, 5.00)				
No. slices per scan[a]	43.2 (28, 40)					32.4 (30, 33)				
Abnormality rates (%)	Intra hyper	Extra hyper	Intra hypo	Extra hypo	Any	Intra hyper	Extra hyper	Intra hypo	Extra hypo	Any
L frontal lobe	4.2	5.0	6.5	3.5	13.6	6.8	7.1	9.5	5.7	19.5
R frontal lobe	5.8	6.1	7.7	4.5	16.4	5.9	7.3	9.5	4.8	19.8
L temporal lobe	2.9	3.7	3.2	2.6	8.5	4.4	5.3	6.5	4.3	14.1
R temporal lobe	3.8	4.4	4.0	3.0	10.9	4.2	4.5	4.5	3.7	12.1
L parietal lobe	2.1	5.8	3.2	2.2	10.6	2.6	6.2	5.6	3.2	13.1
R parietal lobe	2.6	5.7	2.8	2.6	10.4	3.4	4.7	5.2	2.6	12.1
L occipital lobe	1.4	1.5	1.9	1.2	4.7	1.4	2.2	2.3	0.5	4.8
R occipital lobe	1.1	1.9	1.4	0.9	4.0	2.0	1.9	2.1	1.1	5.0
L centrum ovale	0.2	0.0	0.6	0.0	0.7	0.1	0.0	0.7	0.0	0.7
R centrum ovale	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.9	0.0	0.9
L basal ganglia	3.8	0.0	7.3	0.0	8.7	3.5	0.0	6.1	0.0	7.9
R basal ganglia	1.7	0.0	7.3	0.0	8.4	2.3	0.0	4.7	0.0	6.1
L lateral ventricle	0.0	3.7	0.0	0.2	4.0	0.0	5.8	0.0	0.1	5.9
R lateral ventricle	0.0	3.7	0.0	0.2	3.7	0.0	5.7	0.0	0.1	5.9
L cerebellum	1.6	0.1	2.1	0.0	2.4	1.9	0.1	2.1	0.1	2.6
R cerebellum	1.5	0.2	2.2	0.0	2.9	1.5	0.0	1.5	0.0	2.0
Brain stem	1.0	0.0	1.9	0.0	2.5	0.5	0.1	0.8	0.1	1.1
Any	19.5	17.4	25.6	7.2	41.1	22.7	22.1	29.5	10.7	50.8

Numbers of patients, age, and gender are the statistics of individuals, while others are statistics of scans. Intra, intraparenchymal; extra, extraparenchymal; hyper, hyperdensity; hypo, hypodensity; L, left; R, right.
[a]For these values, we show the mean, lower quartile, and upper quartile.

Overview of the learning framework

The overall framework of this work is shown in Figure 1. The training of Cross-DL consists of two stages: (1) an annotation generation stage to generate pseudo abnormality annotation and coarse region segmentation in a cost-effective way and (2) an analyzer training stage to train an image analyzer.

In the annotation generation stage, we developed a discretizer with the report dataset and used it to generate the pseudo abnormality annotations for the retrospective training and validation datasets from the corresponding free-text reports. Specifically, the "findings" part of each report was first split into sentences. Then, the discretizer annotated each sentence, and we merged the results of sentences back into an abnormality matrix, indicating the presence of different types of abnormalities in various anatomical regions (Figure 2A). In this study, we focused on 4 commonly observed types of intracranial abnormalities with clinical significance: intraparenchymal hyperdensity, extraparenchymal hyperdensity, intraparenchymal hypodensity, and extraparenchymal hypodensity. Intraparenchymal abnormalities are those inside the brain parenchyma. Extraparenchymal abnormalities are those outside the parenchyma (including intraventricular, subdural, epidural, and subarachnoid abnormalities). As for locations, our system focused on the presence of the above abnormalities in 17 main anatomical regions (see Table 1) and two additional region labels, other region and unclear region, which were used to represent regions outside our defined regions (e.g., "cerebral falx") and regions with vague descriptions (e.g., "surgery area"), respectively. Detailed label definitions and annotation rules are shown in Tables S4 and S5. Meanwhile, the coarse anatomical region segmentation of each CT scan, which

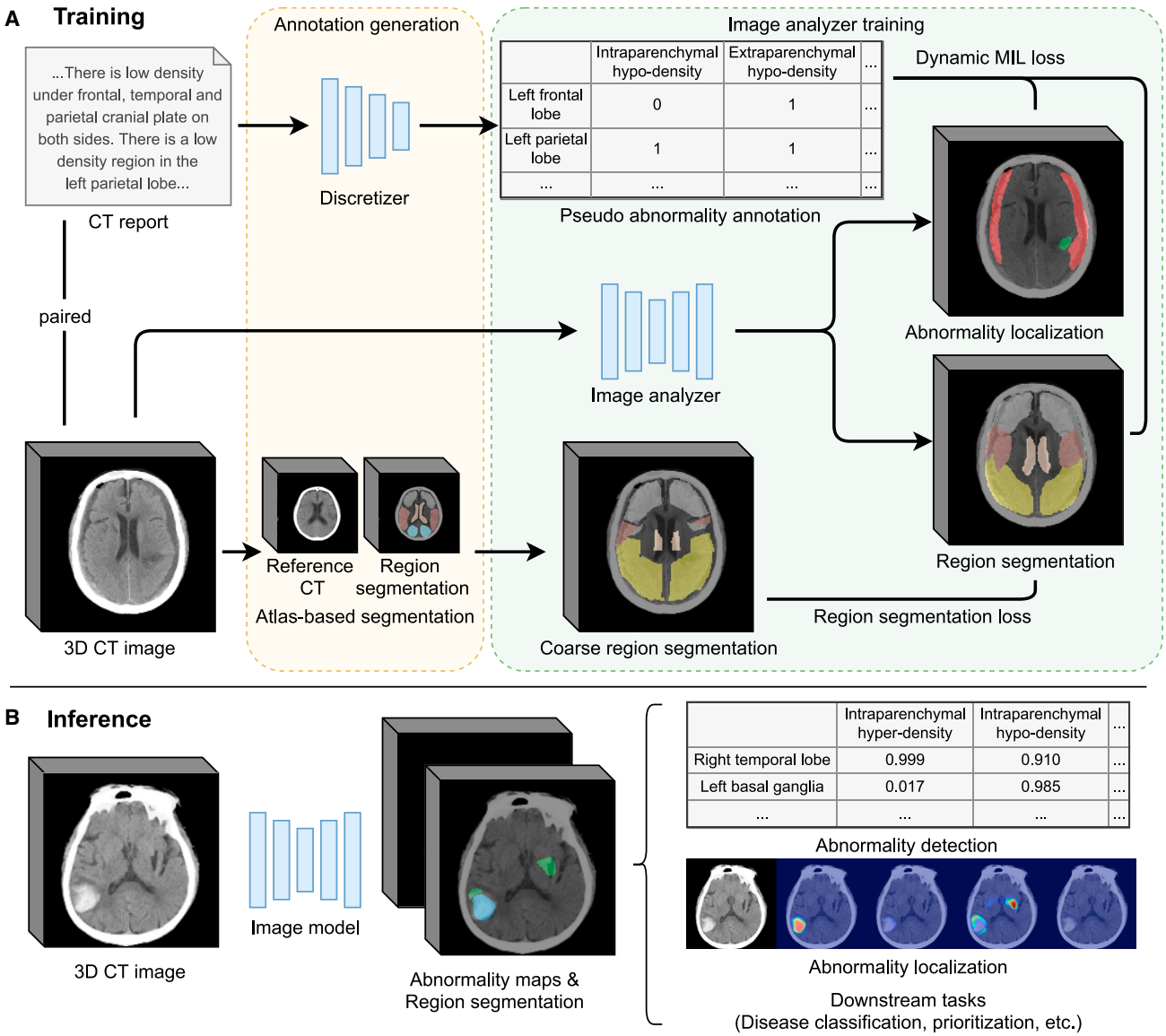


Figure 1. Framework overview

(A) The training of Cross-DL, which contains two stages.

(B) The inference of our obtained image analyzer.

was also required for the training in the second stage, was generated using an atlas-based segmentation method.

In the analyzer training stage, we used the CT scans, together with the pseudo abnormality annotations and the coarse region segmentation, to train an image analyzer to detect and localize intracranial abnormalities in CT scans. The image analyzer is a convolutional neural network, with an input layer that takes a whole CT scan as input, a 3D U-Net-like backbone[19] to extract features, and a multi-task output layer to produce 3D probability maps of the 4 abnormality types and the 17 anatomical regions. The region segmentation prediction was supervised by the coarse region segmentation (see Figure S1 for segmentation results). The predicted 3D probability maps of the anatomical regions and the abnormalities were combined dynamically in a multiple-instance learning (MIL) manner to generate an abnormality matrix, where the matrix obtained in the first stage was used as supervision. Scan-level abnormality prediction could also be acquired using global max pooling, representing the existence of the abnormalities in the whole scan.

After training, the obtained image analyzer could be used to predict the existence of abnormalities in various regions or to localize abnormalities at the voxel level or in downstream tasks such as disease classification and prioritization.

The discretizer generated accurate pseudo abnormality annotation

In this section, we show that the discretizer in Cross-DL, which was trained using 1,500 sentences in the report training set

CellPress
OPEN ACCESS

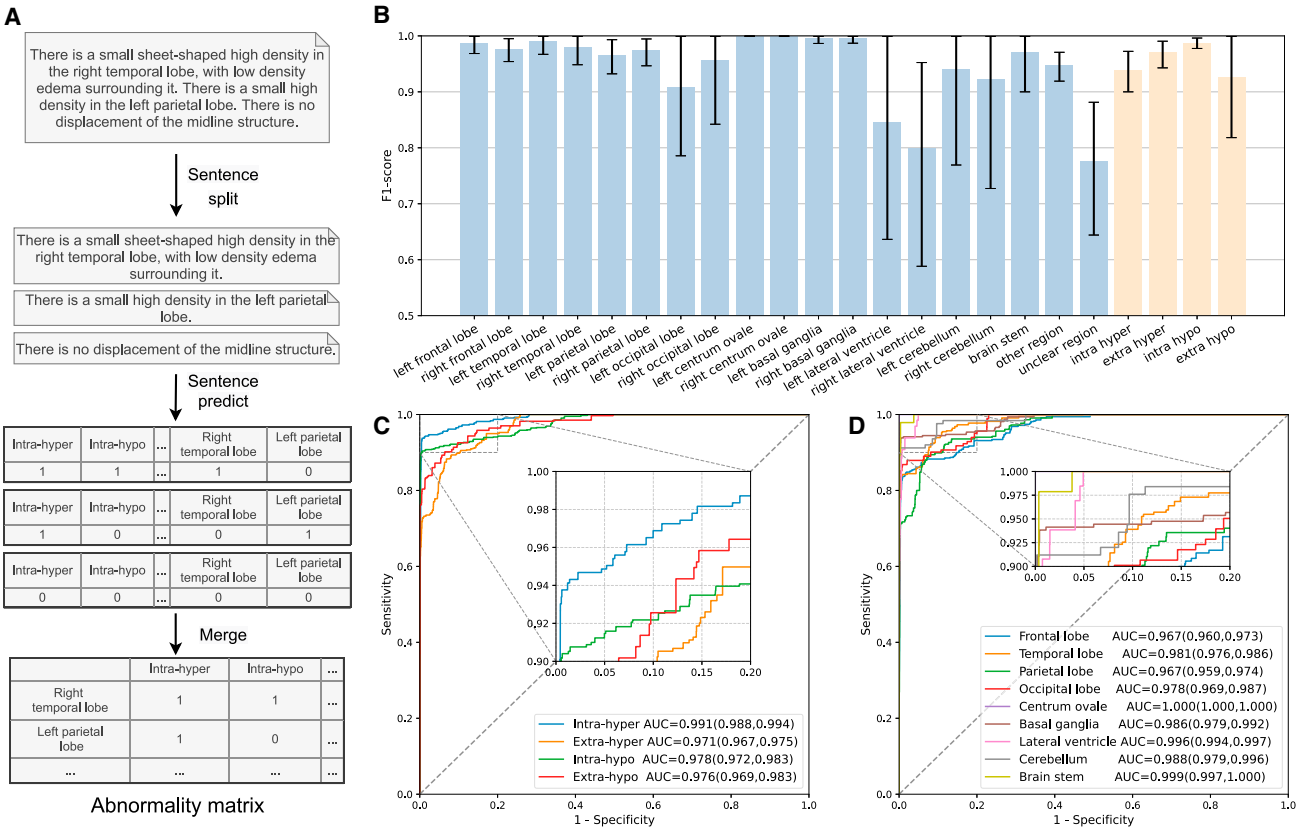


Figure 2. Discretizer evaluation

(A) The sentence-level annotation approach.

(B) The F1 scores of the discretizer to classify if each anatomical region or abnormality type is shown to be positive in a sentence in the report test set under the threshold of 0.5. Regions are plotted in blue, and abnormalities are plotted in yellow. Data are represented as mean and 95% confidence interval.

(C) Average ROC curves of the discretizer to detect four types of abnormalities from a report across all regions in the retrospective test set.

(D) Average ROC curves of the discretizer to detect abnormalities in different regions from a report over all types of abnormalities in the retrospective test set.

(see STAR Methods), could generate accurate pseudo abnormality annotation from the free-text descriptions of imaging reports.

We first evaluated the sentence-level accuracy of the discretizer in extracting the regional locations and abnormality types from sentences on the report test dataset (Figure 2B). Our discretizer achieved an average F1 score of 0.944 in classifying whether each of the 19 regional labels was abnormal, and an average F1 score of 0.956 in classifying whether each of the 4 abnormality types appeared in the sentence. F1 scores on 20 over 23 labels were above 0.90. Even for labels with extremely low positive rates of less than 5% (brain stem and lateral ventricles), the F1 scores were still above 0.80. We then evaluated the scan-level (i.e., report-level) quality of our pseudo abnormality annotation on the retrospective test dataset using the gold-standard annotation. We report the quality of the predicted abnormality matrices in two dimensions: (1) the performance on each abnormality type by averaging the area under the receiver operating characteristic curve (AUROC) over all regions and (2) the performance on each region by averaging the AUROC over all abnormality types (symmetrically corresponding regions were also merged together; see STAR Methods for more details).

Figure 2C shows the average receiver operating characteristic (ROC) curves of the discretizer in detecting the 4 types of abnormalities. The discretizer achieved average AUROCs of 0.991 (95% confidence interval [CI] = 0.988–0.994), 0.971 (95% CI = 0.967–0.975), 0.978 (95% CI = 0.972–0.983), and 0.976 (95% CI = 0.969–0.983) in detecting intraparenchymal hyperdensity, extraparenchymal hyperdensity, intraparenchymal hypodensity, and extraparenchymal hypodensity across all 17 anatomical regions, respectively. Figure 2D shows the ROC curves of the discretizer on different regions, which also show high performance. The results demonstrated that the discretizer was capable of generating reliable pseudo annotations from free-text reports for the training in the next stage.

The image analyzer accurately classified and localized abnormalities

Using the annotations generated in the first stage as supervision, we trained an image analyzer. Our image analyzer can accurately detect and localize intracranial abnormalities in CT scans.

We first evaluated the performance of the image analyzer in detecting abnormalities in anatomical regions. The predicted abnormality matrices of the image analyzer were compared with

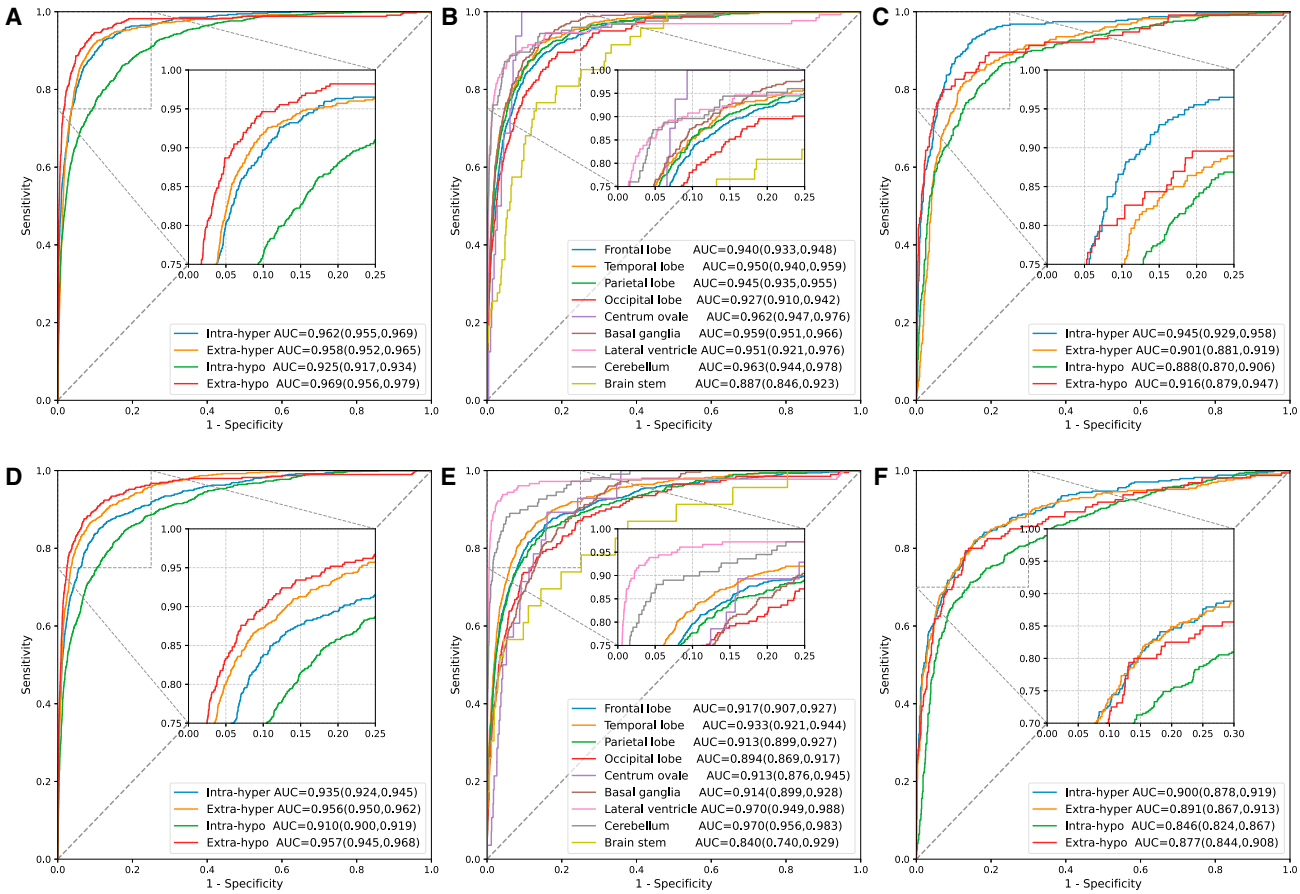


Figure 3. Evaluation of the image analyzer for abnormality detection

(A–C) Retrospective evaluation.

(D–F) Prospective evaluation.

(A and D) Average ROC curves in detecting the four types of abnormalities over all regions.

(B and E) Average ROC curves on different regions.

(C and F) Scan-level ROC curves in detecting the four types of abnormalities.

the gold-standard annotations on the retrospective test dataset. The detailed performance is shown in Table S6. Figure 3A shows the average ROC curves of the image analyzer in detecting the four types of abnormalities over different regions in the test set. The analyzer detected intraparenchymal hyperdensity, extraparenchymal hyperdensity, intraparenchymal hypodensity, and extraparenchymal hypodensity with average AUROCs of 0.962 (95% CI = 0.955–0.969), 0.958 (95% CI = 0.952–0.965), 0.925 (95% CI = 0.917–0.934), and 0.969 (95% CI = 0.956–0.979), respectively. The performance of the intraparenchymal hypodensity was relatively low because of its subtle appearance, especially in cases of early infarction. The analyzer generally performed well on different anatomical regions, with the highest AUROC of 0.963 (95% CI = 0.944–0.978) on the cerebellum and the lowest AUROC of 0.887 (95% CI = 0.846–0.923) on the brain stem (Figure 3B). The low performance on the brain stem was probably because of its small volume, which resulted in a relatively large region segmentation error. Averaging the AUROC scores over all abnormalities and all regions resulted in an average AUROC of 0.956 (95% CI = 0.952–0.959). The analyzer could also generate scan-level predictions using the maximum predicted scores of each abnormality map over all voxels in the scan. The analyzer detected the scan-level existence of intraparenchymal hyperdensity, extraparenchymal hyperdensity, intraparenchymal hypodensity, and extraparenchymal hypodensity with AUROCs of 0.945 (95% CI = 0.929–0.958), 0.901 (95% CI = 0.881–0.919), 0.888 (95% CI = 0.870–0.906), and 0.916 (95% CI = 0.879–0.947), respectively (Figure 3C). As the scan-level abnormality annotation covers abnormalities in all regional labels including other regions and unclear regions, this evaluation shows our model's capability of detecting intracranial abnormalities in the whole scan including those outside the defined anatomical regions.

To demonstrate the generalizability of the analyzer, we conducted experiments on the prospective dataset, which has different distributions of CT scanners and abnormalities from the retrospective dataset (Table S7). Figures 3D–3F show the ROC curves of the analyzer for abnormality detection on the prospective dataset. The analyzer achieved average AUROCs of 0.935 (95% CI = 0.924–0.945), 0.956 (95% CI = 0.950–0.962),

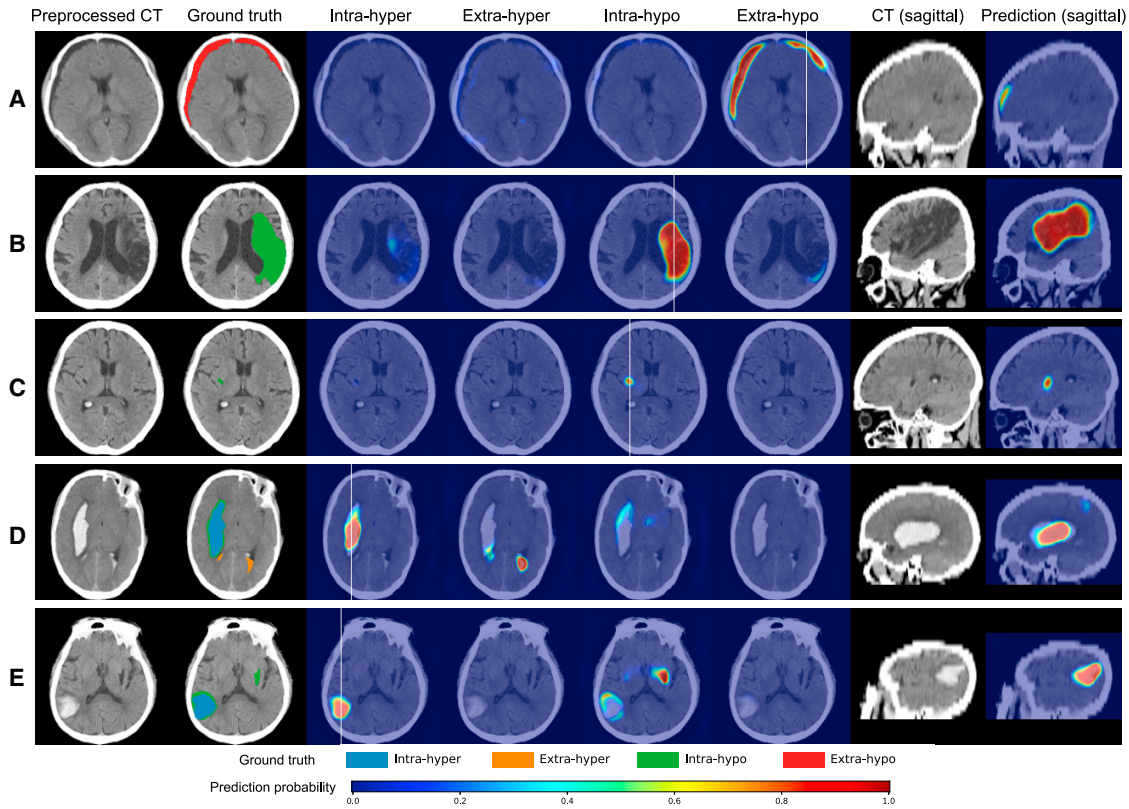


Figure 4. Abnormality localization results

(A–E) Five examples of abnormality prediction results of the image analyzer. Our image analyzer predicts 3D abnormality probability maps for 3D CT images. 2D slices in axial view and sagittal view are shown in the figure. Each example starts with an axial slice of the input preprocessed CT and the ground truth of abnormalities annotated by experts, followed by color maps of predicted probability maps of the four abnormalities and the CT image and probability map of a sagittal slice (the location and abnormality type of which are depicted by the white line in the corresponding axial slice).

(A) Extraparenchymal hypodensity shown under the skull was detected and accurately localized.

(B) Large intraparenchymal hypodensity region localized with probability maps of corresponding shape and size.

(C) Small intraparenchymal hypodensity region localized by the analyzer (calcification in the ventricle was ignored, as we did not regard calcification as a hyperdensity abnormality).

(D) Intraparenchymal hyperdensity and extraparenchymal hyperdensity in ventricles were correctly detected and localized, but the analyzer failed to accurately localize the surrounding slender intraparenchymal hypodensity (edema).

(E) Intraparenchymal hyperdensity, intraparenchymal hypodensity (edema), and intraparenchymal hypodensity (malacia) correctly localized by the analyzer.

0.910 (95% CI = 0.900–0.919), and 0.957 (95% CI = 0.945–0.968) in detecting intraparenchymal hyperdensity, extraparenchymal hyperdensity, intraparenchymal hypodensity, and extraparenchymal hypodensity on the prospective test set, respectively. The results of different regions and scan-level abnormality detection also showed high performance.

Next, we show the accuracy of the image analyzer in localizing abnormalities in 3D CT scans. The image analyzer generated voxel-level prediction maps of different abnormality types, indicating the locations, volumes, and shapes of abnormalities in CT scans (Figure 4). To quantitatively show the segmentation performance, we randomly selected 40 CT scans from the retrospective test set that had at least one of the four abnormality types and manually annotated the abnormality segmentation on each scan. We evaluated the Dice score of the predicted abnormality on each 3D image. The average Dice scores to segment the four abnormalities were 0.287 (95% CI = 0.177–0.408), 0.379 (95% CI = 0.251–0.495), 0.358 (95% CI = 0.280–0.437), and 0.333 (95% CI = 0.105–0.509), respectively. The error mainly came from the boundaries of the abnormalities since many of them were small or narrow. The image analyzer, which did not have any prior knowledge on the abnormality boundaries, may produce inaccurate boundaries, resulting in a relatively large error and in low Dice scores.

The analyzer could detect various diseases and help prioritization

Besides abnormalities, our image analyzer could also be used for review prioritization. As we focused on the detection of abnormalities such as hyperdensity and hypodensity instead of specific diseases, the analyzer could in turn detect various diseases, as they appear as these abnormalities in CT images. To test this, we performed experiments on the preoperative dataset containing 1,204 scans with manual annotations of 12 common diseases. Given a test CT scan, the image analyzer predicted a

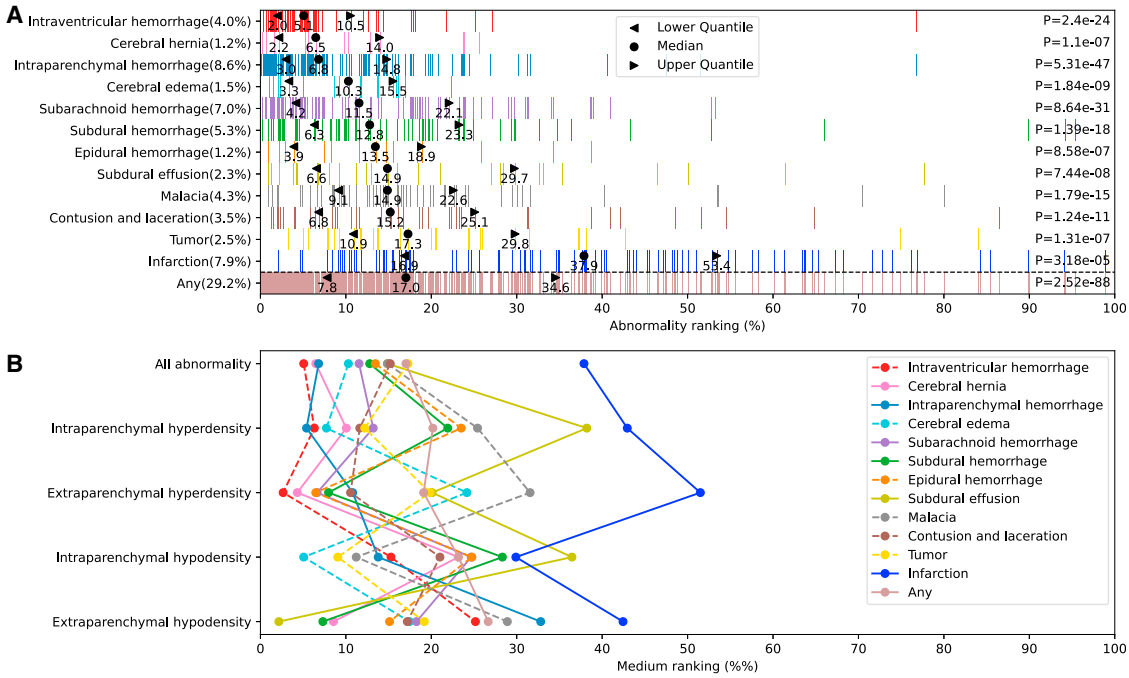


Figure 5. Prioritization evaluation

(A) The performance of the image analyzer in detecting 12 commonly found diseases in head CT scans using the maximum prediction of the four abnormality types as the abnormality score. Each row represents a disease. The last row (any) represents the union of all the 12 diseases. Each vertical strip (colored or not) represents a scan, sorted by the predicted abnormality score in descending order from left to right. A colored strip in a row means a certain scan has the corresponding disease. We show the quartiles of the rankings of CT scans that had the diseases. Diseases are sorted by medium abnormal ranking in ascending order from top to bottom. The positive ratio of each disease is labeled next to the disease name. The p value of the one-sided Wilcoxon rank-sum test for each disease is shown.

(B) Medium abnormal ranking of each disease when using the maximum prediction of four abnormality types and the maximum prediction of each abnormality as the abnormality score separately. Each row corresponds to an abnormality type.

score for each abnormality type at each voxel. The maximum score over all voxels and the four abnormality types was used as the abnormality score of the scan. Based on the abnormality score, we sorted all scans in the preoperative dataset in descending order. The distributions of the 12 diseases on the sorted list are shown in Figure 5A. For all diseases, whether they typically appear as hyperdensity (e.g., different types of hemorrhage) or hypodensity abnormalities (e.g., cerebral edema, malacia, and infarction), CT scans with the disease had significantly higher rankings than those without the disease (two-sample one-sided Wilcoxon rank-sum test, $p < 0.001$). We found that the performance to prioritize infarction cases was the lowest, probably due to their indistinct intraparenchymal hypodensity and the relatively low performance of the analyzer on this abnormality.

As different diseases tend to appear as different types of abnormalities in CT scans, we checked the prioritization performance of the analyzer for each disease using the prediction of different abnormalities. The maximum prediction scores of each abnormality type over all voxels were used as the abnormality scores, respectively. Figure 5B shows the medium abnormality ranking of each disease using four types of abnormalities (see Figure S5 for detailed performance). The performance of each disease on different abnormalities was consistent with medical knowledge. For example, the performance for intraparenchymal hemorrhage, whose typical feature is hyperdensity blood in brain parenchyma,[20] was the best using the score of intraparenchymal hyperdensity. The performance was also good using extraparenchymal hyperdensity and intraparenchymal hypodensity because of complications including other types of hemorrhage and edema and was the worst using extraparenchymal hypodensity. The performance for epidural, subdural, subarachnoid, and intraventricular hemorrhages was the best using extraparenchymal hyperdensity as the blood is accumulated outside the parenchyma in those diseases. The performance for subdural effusion was the best using extraparenchymal hypodensity and the second best using extraparenchymal hyperdensity, mainly because of complications including subdural hemorrhage.[21] The performance for cerebral malacia, edema, and infarction was the best using intraparenchymal hypodensity. The performance for cerebral hernia and contusion and laceration was the best using extraparenchymal hyperdensity, probably because of associated diseases including hemorrhage.[22] The performance for tumors was the best on intraparenchymal hypodensity. The above experiments and the consistency between model performance and medical knowledge indicate our analyzer's accuracy and demonstrate its utility in review prioritization.

CellPress
OPEN ACCESS

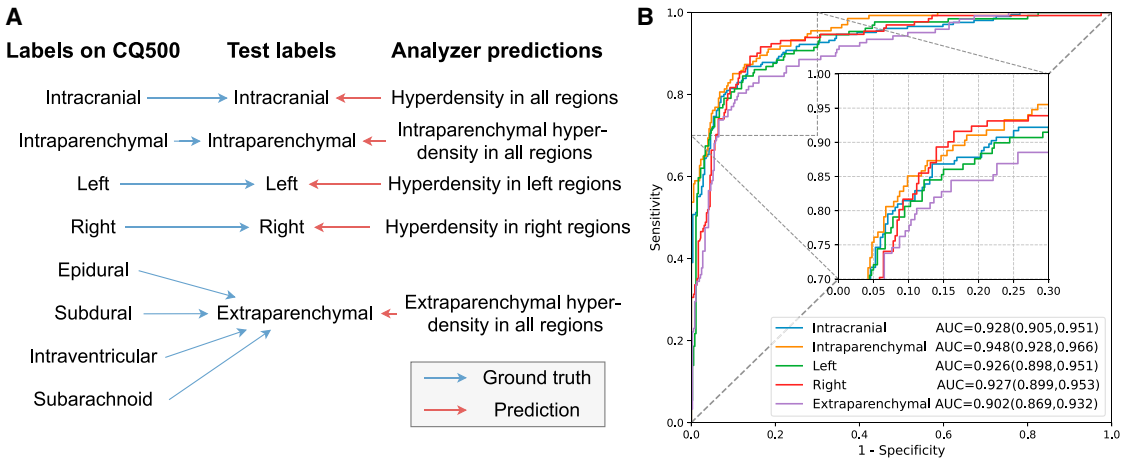


Figure 6. Hemorrhage classification evaluation on the CQ500 dataset

(A) The labels used in our experiment and the correspondence between them and the original labels on CQ500 as well as predictions of our image analyzer.
(B) ROC curves of the analyzer to detect the existence of the five types of hemorrhages on CQ500.

The analyzer generalized well on external data

We further demonstrate the generalizability of our analyzer on external data and its potential use in supporting downstream medical tasks by performing experiments on the CQ500 dataset. CQ500 is a public dataset consisting of 491 head CT scans with manual annotations including the binary labels of intraparenchymal hemorrhage, subdural hemorrhage, epidural hemorrhage, subarachnoid hemorrhage, and intraventricular hemorrhage, as well as bleeding in the left and right sides of the brain. Other annotations (midline shift and mass effect, and calvarial fracture) were not used in this study. As our image analyzer predicted abnormalities rather than diseases as in CQ500, we defined five class labels to relate our prediction to the annotations in CQ500. These labels are intracranial hemorrhage, intraparenchymal hemorrhage, left hemorrhage, and right hemorrhage, all of which were already defined in CQ500, and an additional extraparenchymal hemorrhage label, of which the ground truth was determined by the union of subdural, epidural, subarachnoid and intraventricular hemorrhages. The predictions of these labels could be acquired from the prediction maps of our image analyzer. As hemorrhage typically appears as hyperdensity in CT scans, we used a straightforward approach using the maximum predictions of intraparenchymal and extraparenchymal hyperdensity as the hemorrhage score (Figure 6A).

Figure 6B shows the ROC curves for the five labels on CQ500. Our system achieved AUROCs of 0.928 (95% CI = 0.905–0.951) for intracranial hemorrhage, 0.948 (95% CI = 0.928–0.966) for intraparenchymal hemorrhage, 0.926 (95% CI = 0.898–0.951) for bleeding in the left brain, 0.927 (95% CI = 0.899–0.953) for bleeding in the right brain, and 0.902 (95% CI = 0.869–0.932) for extraparenchymal hemorrhage on the CQ500 dataset. Note that as we used the predictions of hyperdensity directly as probabilities of hemorrhage, errors might occur where other diseases with hyperdensity were classified as hemorrhage and hemorrhage with hypodensity was classified as non-hemorrhage. Despite that, the analyzer still achieved an AUROC above 0.90 on all five labels, showing its generalizability on external data and its utility for disease classification.

DISCUSSION

In this study, we aimed to alleviate the need for massive manual annotations in DL-based medical image analysis. We hold that innovating methods to learn from readily available data sources like imaging reports is important, as it would not only reduce the cost of system development but also enable the utility of more data to help improve system accuracy and generalizability.

Concretely, we show how free-text imaging reports of head CT scans could be utilized to generate pseudo abnormality annotations and how these annotations could be used to train an image analyzer to detect and localize different types of intracranial abnormalities in various anatomical regions. We demonstrated that precise annotations of abnormality location and type could be extracted from free-text imaging reports using a discretizer. The training of the discretizer was easy, as it only requires a small report dataset, which could be easily annotated without requiring professional expertise. We then showed that, using the pseudo abnormality annotation and our training approach, an image analyzer could be trained to not only predict the existence of different types of abnormalities in various anatomical regions but also to accurately localize them in 3D CT scans. This is greatly meaningful for radiologists, as in their routine work, they first examine the images carefully for abnormalities, and then they make diagnoses based on those abnormalities and other factors such as clinical symptoms. The detection of abnormalities in the first step is a prerequisite for disease diagnosis. Our analyzer could help radiologists in the first step by quickly detecting and localizing abnormalities and regions of interest. In this work, we used 1,500 manually labeled sentences as the report training dataset, but it was surprising to find that a good discretizer and a good subsequent image analyzer could be learned even with much fewer data (300 sentences as in

Figure S2). This demonstrated the tolerance of our discretizer to learning from small training sets. Besides the utilization of state-of-the-art NLP techniques in our discretizer model, it was also because imaging reports are more structured than normal languages. Typically, radiologists would select a template based on the CT scan and then make modifications, making the resulting report well-structured and a good source of pseudo annotations. Although our reports were in Chinese, we hold that our method for training the discretizer is generalizable to other languages, based on the model's generalizability that has been shown in hundreds of languages.[23] We conducted prospective and external experiments to show our analyzer's generalizability using different datasets. The analyzer showed slightly lower performance on the prospective set compared with the retrospective test set, which is common in DL-based methods. The primary reason was probably the disparity in the distribution of CT slice thickness and abnormalities between the retrospective and prospective datasets. To make a clinically applicable artificial intelligence (AI) system with high performance (e.g., Aidoc),[24,25] the requirement for model generalization will be higher, and a more diverse training dataset is needed.

We demonstrated our image analyzer's usage in review prioritization. Using our analyzer, most scans with various diseases could be detected and presented to radiologists prior to normal scans. This would save precious time for the diagnosis and treatment of critically ill patients. The order in which different diseases are detected is also important. Although not intended in advance, our analyzer tended to detect CT with severe and urgent diseases first (i.e., gave them higher abnormality scores). When using the maximum prediction over all four types of abnormalities as the score, the analyzer was more sensitive for cerebral hernia, cerebral edema, and four types of hemorrhages (Figure 5A), which are not only serious but also urgent diseases.[26] Scans with these diseases could be presented to radiologists preferentially. Scans with less serious diseases such as subdural effusion and infarction tended to be detected with lower rankings. The performance of the analyzer on each disease when using predictions of different types of abnormalities was also consistent with medical knowledge, indicating the analyzer's accuracy in abnormality prediction.

We also conducted experiments to justify our design of the framework and explore how it performed under different settings (see STAR Methods and Figure S2) in which we showed that (1) the performance of the image analyzer was limited by the discretizer, yet a small number of manual annotations was sufficient to train a discretizer that was good enough. (2) Our framework enabled massive raw medical data to be automatically annotated and used as training data with no additional annotation cost, which improves the performance of the analyzer as the dataset grows and would further improve it with even more data. (3) We extracted fine-grained region-level annotations rather than scan-level annotations, which allowed us to utilize the abnormality location information in imaging reports and boosted the analyzer's performance, especially for abnormality localization. (4) Our dynamic MIL approach to training the image analyzer outperformed the classification and traditional MIL approaches in abnormality detection and localization. (5) As each abnormality type shows common features in different regions, training the analyzer on multiple regions simultaneously improved its accuracy.

In this work, we focused on 17 anatomical regions and 4 types of abnormalities. To make a wide coverage of regional information, we selected 17 anatomical regions with relatively large volumes and relatively high positive rates of abnormalities. Although we mainly used the region-level annotations to train the image analyzer, we also took into consideration abnormalities outside these regions using two additional regional labels (unclear region and other region) and a scan-level loss term (see STAR Methods). This enabled our image analyzer to also detect abnormalities outside the defined regions and resulted in good scan-level accuracy in abnormality detection. As for the abnormality types, we selected these 4 types of abnormalities because hyperdensity and hypodensity are commonly seen in imaging reports and the distinction between intraparenchymal and extraparenchymal abnormalities is important in determining the type of the diseases. More refined and detailed definitions of anatomical regions and abnormality types could be handled by the same framework.

As our framework enables the prediction of abnormality type and location from medical images, which is essential information in imaging reports, this work could be potentially utilized in automatic report generation. Although there are existing works for automatic report generation for chest X-ray images and other medical images,[27–29] the work on the automatic generation of head CT reports is still rare. Many of those works utilized image-captioning techniques and did not introduce such fine-grained abnormality labels, and the quantitative metrics mostly used are those for natural language generation.[30,31] Although the use of these methods and metrics eliminated the need for manual annotation of tags and ensured that the generated reports are formal and fluent, we conjecture that the utilization and evaluation of fine-grained, region-level abnormality labels are still helpful for report generation, as the accuracy of imaging reports is as important as, if not more important than, regularity and fluency.

In conclusion, we developed and validated a weakly supervised learning framework that greatly reduces the cost of manual labeling and produces DL models with high accuracy. We prospectively and externally tested the model for its performance in generating scan-level, region-level, and voxel-level abnormality prediction and for its clinical usability in review prioritization and disease classification. We expect this work and future studies would benefit radiologists in the diagnosing and reporting of head CT exams and other medical imaging modalities.

Limitations of the study

There are some limitations of this study. First, the anatomical regions and abnormality types we used did not cover all the intracranial regions and abnormalities in head CT scans, such as cerebral cistern, midline shift, and mass effect. Particularly, dealing with morphological abnormalities would be more challenging than density abnormalities using our framework. Second, for prioritization, we currently only used the maximum

abnormality score over all voxels of each scan. As abnormalities in different anatomical regions tend to have different clinical manifestations and severity, integrating the anatomical location of abnormalities may result in better prioritization performance. The sizes and volumes of the abnormalities, which could be derived from our abnormality prediction maps, could also be useful for the task.[32] Third, our learning framework can be applied to other imaging modalities or body parts, where the reports have detailed abnormality descriptions corresponding to anatomical regions. However, it may not perform well on regions or organs that are too small or are highly deformable (i.e., vessel and intestine), as they are difficult to segment using the atlas-based method.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Report dataset construction
  - Abnormality annotation definition
  - Manual abnormality labeling
  - Discretizer model and training
  - Anatomical region segmentation
  - CT image preprocessing
  - Image analyzer architecture and training
  - Ablation study
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2023.101164.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (2020AAA0105500 to Y.G. and Q.D. and 2018YFA0704000 to F.X.); the Beijing Natural Science Foundation (M22024 to F.X.); the National Natural Science Foundation of China (81825012, 81730048, and 82151309 to X.L., 81901708 to J.L., 62021002 to F.X., and 61971260 to Y.G.); the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G to F.X.); and the Zhejiang Provincial Natural Science Foundation (LDT23F02024F02 to J.X.). This work is also supported by THUIBCS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission.

AUTHOR CONTRIBUTIONS

Conceptualization, A.L., Y.G., and J.L.; methodology, A.L., Y.G., and F.X.; software, investigation, and formal analysis, A.L.; experimental verification, A.L., Y.G., F.X., and J.L.; writing – original draft, A.L., Y.G., and F.X.; writing – review & editing, A.L., Y.G., F.X., J.L., J.-h.Y., and J.X.; resources, X.L.; project administration and supervision, F.X., X.L., and Q.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 31, 2022
Revised: April 30, 2023
Accepted: July 27, 2023
Published: August 21, 2023

REFERENCES

1. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., and Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet *392*, 2388–2396. https://doi.org/10.1016/s0140-6736(18)31645-3.
2. Gao, X.W., Hui, R., and Tian, Z. (2017). Classification of CT brain images based on deep learning networks. Comput. Methods Progr. Biomed. *138*, 49–56. https://doi.org/10.1016/j.cmpb.2016.10.007.
3. Chen, W., Belle, A., Cockrell, C., Ward, K.R., and Najarian, K. (2013b). Automated Midline Shift and Intracranial Pressure Estimation based on Brain CT Images. JoVE *74*, 3871. https://doi.org/10.3791/3871.
4. Titano, J.J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., et al. (2018). Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat. Med. *24*, 1337–1341. https://doi.org/10.1038/s41591-018-0147-y.
5. Mitani, A., Huang, A., Venugopalan, S., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Liu, Y., and Varadarajan, A.V. (2020). Detection of anaemia from retinal fundus images via deeplearning. Nat. Biomed. Eng. *4*, 18–27. https://doi.org/10.1038/s41551-019-0487-z.
6. Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
7. Wardlaw, J.M., Seymour, J., Cairns, J., Keir, S., Lewis, S., and Sandercock, P. (2004). Immediate Computed Tomography Scanning of Acute Stroke Is Cost-Effective and Improves Quality of Life. Stroke *35*, 2477–2483. https://doi.org/10.1161/01.str.0000143453.78005.44.
8. Papa, L., Stiell, I.G., Clement, C.M., Pawlowicz, A., Wolfram, A., Braga, C., Draviam, S., and Wells, G.A. (2012). Performance of the Canadian CT Head Rule and the New Orleans Criteria for Predicting Any Traumatic Intracranial Injury on Computed Tomography in a United States Level I Trauma Center. Acad. Emerg. Med. *19*, 2–10. https://doi.org/10.1111/j.1553-2712.2011.01247.x.
9. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118. https://doi.org/10.1038/nature21056.
10. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA *316*, 2402–2410. https://doi.org/10.1001/jama.2016.17216.
11. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05225.
12. Springenberg, J.T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6806.
13. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int. J. Comput. Vis. *128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7.

14. Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V.N. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. Preprint at arXiv. https://doi.org/10.1109/wacv.2018.00097.
15. Cohen, I.G., and Mello, M.M. (2019). Big Data, Big Tech, and Protecting Patient Privacy. JAMA *322*, 1141–1142. https://doi.org/10.1001/jama.2019.11365.
16. Price, W.N., and Cohen, I.G. (2019). Privacy in the age of medical big data. Nat. Med. *25*, 37–43. https://doi.org/10.1038/s41591-018-0272-7.
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R.M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Preprint at arXiv. https://doi.org/10.1109/cvpr.2017.369.
18. Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., and Cuadra, M.B. (2011). A review of atlas-based segmentation for magnetic resonance brain images. Comput. Methods Progr. Biomed. *104*, e158–e177. https://doi.org/10.1016/j.cmpb.2011.07.015.
19. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Preprint at arXiv. https://doi.org/10.1007/978-3-319-24574-4_28.
20. Huisman, T.A.G.M. (2005). Intracranial hemorrhage: ultrasound, CT and MRI findings. Eur. Radiol. *15*, 434–440. https://doi.org/10.1007/s00330-004-2615-7.
21. MURATA, K. (1993). Chronic Subdural Hematoma May be Preceded by Persistent Traumatic Subdural Effusion. Neurol. Med.-Chir. *33*, 691–696. https://doi.org/10.2176/nmc.33.691.
22. Riveros Gilardi, B., Muñoz López, J.I., Hernández Villegas, A.C., Garay Mora, J.A., Rico Rodríguez, O.C., Chávez Appendini, R., De la Mora Malváez, M., and Higuera Calleja, J.A. (2019). Types of Cerebral Herniation and Their Imaging Features. Radiographics *39*, 1598–1610. https://doi.org/10.1148/rg.2019190018.
23. Pires, T., Schlinger, E., and Garrette, D. (2019). How Multilingual is Multilingual BERT?. Preprint at arXiv. https://doi.org/10.18653/v1/p19-1493.
24. Ginat, D.T. (2020). Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. Neuroradiology *62*, 335–340. https://doi.org/10.1007/s00234-019-02330-w.
25. Voter, A.F., Meram, E., Garrett, J.W., and Yu, J.-P.J. (2021). Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage. J. Am. Coll. Radiol. *18*, 1143–1152. https://doi.org/10.1016/j.jacr.2021.03.005.
26. Caceres, J.A., and Goldstein, J.N. (2012). Intracranial Hemorrhage. Emerg. Med. Clin. *30*, 771–794. https://doi.org/10.1016/j.emc.2012.06.003.
27. Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating Radiology Reports via Memory-driven Transformer. Preprint at arXiv. https://doi.org/10.18653/v1/2020.emnlp-main.112.
28. Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R.M. (2018). TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. Preprint at arXiv. https://doi.org/10.1109/cvpr.2018.00943.
29. Jing, B., Xie, P., and Xing, E. (2018). On the Automatic Generation of Medical Imaging Reports. https://doi.org/10.18653/v1/p18-1240.
30. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU. https://doi.org/10.3115/1073083.1073135.
31. Lin, C.-Y., and Hovy, E. (2002). Manual and Automatic Evaluation of Summaries. https://doi.org/10.3115/1118162.1118168.
32. Tuhrim, S., Horowitz, D.R., Sacher, M., and Godbold, J.H. (1999). Volume of ventricular blood is an important determinant of outcome in supratentorial intracerebral hemorrhage. Crit. Care Med. *27*, 617–621. https://doi.org/10.1097/00003246-199903000-00045.
33. Klein, S., Staring, M., Murphy, K., Viergever, M.A., and Pluim, J.P.W. (2010). elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Trans. Med. Imag. *29*, 196–205. https://doi.org/10.1109/tmi.2009.2035616.
34. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
36. Guimond, A., Meunier, J., and Thirion, J.-P. (1998). Automatic Computation of Average Brain Models (Springer), pp. 631–640. https://doi.org/10.1007/bfb0056249.
37. Dietterich, T.G., Lathrop, R.H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. *89*, 31–71. https://doi.org/10.1016/s0004-3702(96)00034-3.
38. Quellec, G., Cazuguel, G., Cochener, B., and Lamard, M. (2017). Multiple-Instance Learning for Medical Image and Video Analysis. IEEE Rev. Biomed. Eng. *10*, 213–234. https://doi.org/10.1109/rbme.2017.2651164.
39. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., and Takeuchi, I. (2020). Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. Preprint at arXiv. https://doi.org/10.1109/cvpr42600.2020.00391.
40. Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S., Metaxas, D.N., and Zhou, X.S. (2016). Multi-Instance Deep Learning: Discover Discriminative Local Anatomies for Bodypart Recognition. IEEE Trans. Med. Imag. *35*, 1332–1343. https://doi.org/10.1109/tmi.2016.2524985.
41. Wang, Y., Li, J., and Metze, F. (2019). A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. Preprint at arXiv. https://doi.org/10.1109/icassp.2019.8682847.
42. Jadon, S. (2020). A survey of loss functions for semantic segmentation. Preprint at arXiv. https://doi.org/10.1109/cibcb48159.2020.9277638.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CQ500 dataset	Qure.ai	http://headctstudy.qure.ai/dataset
Codes	This paper	Github: https://github.com/liuaohanjsj/Cross-DL
Software and algorithms		
Python 3.6.9	Python Software Foundation	https://www.python.org
Keras 2.1.2	Google	https://keras.io/
TensorFlow 1.15.0	Google Brain	https://www.tensorflow.org/
SciPy 1.7.1	SciPy	https://scipy.org/
Elastix	Stefan Klein & Marius Staring	https://elastix.lumc.nl/
Pretrained BERT	Google AI Language	https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Feng Xu (xufeng2003@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The CT images and reports data collected from the Chinese PLA General Hospital would not be publicly available due to privacy concerns. The external dataset CQ500 is available at http://headctstudy.qure.ai/dataset. All original code has been deposited on https://github.com/liuaohanjsj/Cross-DL and is publicly available as of the date of publication. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study is approved by the Research Ethics Committee of the Chinese PLA General Hospital (S2018-154-01). Patient consent was waived. 31,778 non-contrast head CT scans of 18,125 patients between April 2012 and July 2019 in the retrospective dataset and 1,500 non-contrast head CT scans of 1,208 patients between August 2019 and August 2021 in the prospective dataset were collected by retrieving the hospital's Picture Archiving and Communication Systems.

Specifically, we first collectd 56, 535 head and neck CT scans of 27, 162 patients (Table S3A). These CT scans were documented between September 2008 and March 2020. All CT scans were collected and stored in Digital Imaging and Communications in Medicine (DICOM) format. We then matched them with 251,463 CT imaging reports of 201,383 patients between April 2012 and July 2019 collected using rules (Table S3B), so that each CT scan was paired with the report with the smallest time difference within 24 h, resulting in 37,453 scan-report pairs. All the reports were written in Chinese.

To improve the quality of the dataset, we removed low-quality scans, including 280 scans with less than 10 slices as they are too short to cover the full brain and would be inadequate for intracranial abnormality detection, 424 scans reconstructed for the bone window as such scans have low signal to noise ratio (SNR) under soft tissue window, and 755 manually filtered scans with extreme artifacts (e.g., have motion artifacts, contain lung area). The manual examination of scan quality was only performed on 4,875 scans with poor registration results (i.e., Elastix[33] final metric value above −0.8), and took a non-expert about two and a half hours. Examples of these removed scans are shown in Figure S3. This resulted in 35,994 scan-report pairs left. As we do not regard ischemia lesions as intraparenchymal hypo-density abnormality in this work, while in some reports it is difficult to distinguish between ischemia

lesions and other hypo-density abnormalities in each region, we removed reports where ischemia lesions show together with other hypo-density abnormalities. This resulted in the final 31,778 scan-report pairs of 18,125 patients in the retrospective dataset.

The retrospective dataset was randomly divided by patient into a training set containing 28,472 CT scans of 16,316 patients, a validation set containing 1,699 CT scans of 902 patients, and a test set containing 1,607 CT scans of 907 patients. The validation set was used for model selection and early stopping during the training process, and the test set was used for the final evaluation. All the prospective dataset was used for evaluation.

METHOD DETAILS

Report dataset construction

We constructed a labeled report dataset using a small portion of the reports in the retrospective dataset to develop and validate our discretizer.

We used the informative 'Findings' section of each report in our report dataset, as it has detailed descriptions of abnormality types and their anatomical locations. The 'Findings' section of each report was composed of several semantically independent sentences. After splitting the 'Findings' sections of the 31,778 reports in the retrospective dataset into sentences, we got 125,430 sentences. We randomly selected 4,807 (3.83%) sentences and split them randomly into a training set of 3,000 sentences (only 1,500 used in the training of the final image analyzer, others used in the ablation study), a validation set of 904 sentences and a test set a 903 sentences. A sentence might appear more than once in the dataset as radiologists use templates when writing reports, and some sentences in the training, validation, and test set might be identical.

Abnormality annotation definition

We focused on N intracranial anatomical regions and M types of abnormalities commonly found in head CT reports ($N = 17$ and $M = 4$ in this work), and two additional region labels: *other region* and *unclear region*, representing intracranial regions outside our defined regions (e.g., "cerebral falx") and intracranial regions with unclear descriptions (e.g., "left hemisphere", "surgery area"), respectively. The location of an abnormality was represented by the anatomical region where it appears. We used 19 regional labels in this work, whose definitions and typical descriptions in imaging reports are shown in Table S4. We focused on 4 types of abnormalities in this work: intraparenchymal hyper-density, extraparenchymal hyper-density, intraparenchymal hypo-density, and extraparenchymal hypo-density, the definitions of which are shown in Table S5.

The abnormality annotation of the 'Findings' section in each report could be represented as a binary abnormality matrix $A \in \{0,1\}^{(N+2)\times M}$, where A_{ij} equals 1 indicating the i-th region has the j-th abnormality, and vise versa.

Instead of directly labeling the abnormality matrix for each report, we found it was easier to first label each sentence independently and then merge the results back into the abnormality matrix. The typical format of a sentence describing abnormalities is "There are [one or multiple types of abnormalities] in [one or multiple anatomical regions]", where all the anatomical regions mentioned in the sentence share all those abnormalities. For example, the sentence "There is a hyperdensity area with clear boundary in the left frontal lobe and temporal lobe, with an irregular hypodensity area in the periphery." indicates that there are intraparenchymal hyperdensity and intraparenchymal hypodensity in both the left frontal lobe and the left temporal lobe. Hence the abnormality information of a single sentence could be simply represented by a vector $r \in \{0,1\}^{N+2}$ indicating if each region is abnormal, and a vector $t \in \{0,1\}^{M}$ indicating if each type of abnormality exists. The abnormality matrix of a report could be obtained using the abnormality vectors of the sentences in its 'Findings' section as:

$$A = \max_{s}\left(r_s t_s^{T}\right)$$

where s is the sentence in the 'Findings' section.

Manual abnormality labeling

Our gold-standard abnormality annotations were labeled by senior radiologists based on both the CT images and corresponding reports. In practice, each report was first written by a radiologist and then reviewed and modified by a peer radiologist during work. During the labeling process, a third senior radiologist was given both the CT image and the report and asked to annotate the abnormality region and type for each sentence using our self-designed program. The *unclear region* label was discarded for the gold-standard annotation as the radiologist would determine the exact abnormal regions based on the CT images. The retrospective test set and the prospective set were labeled by two senior radiologists (with more than 11 years of working experience) respectively.

The silver-standard abnormality annotations were given to sentences in the retrospective training and validation set by non-experts, given specific definition and samples of anatomical regions and abnormality types (Tables S4 and S5).

Discretizer model and training

We built a discretizer based on NLP techniques and used it to predict the abnormality matrix for each report. The discretizer was first trained and evaluated on the report dataset and then used to predict abnormality labels for unlabeled imaging reports in the

retrospective training and validation dataset. Sentences in a report were first processed individually before being merged into the final prediction.

We used the BERT pre-trained model[34] in our discretizer based on the Transformer architecture,[35] to predict the abnormality vector of each sentence. Each sentence was first tokenized and transformed into the input form of BERT by adding a [CLS] token in the front and a [SEP] token at the end before being fed into the model. The output feature of the model at the [CLS] position was connected to an output layer (a fully connected layer followed by a sigmoid operation) to generate a vector $\hat{r} \in [0,1]^{N+2}$ representing the probabilities of the N+2 regions appearing in the sentence as abnormal, and a vector $\hat{t} \in [0,1]^{M}$ representing the probabilities of the M types of abnormalities shown in these regions. The sentence-level prediction model could be formulated as

$$G(x_1, x_2, ..., x_n|\theta_G) = (\hat{r}, \hat{t})$$

where $x_1, x_2, ..., x_n$ are the input tokens of the sentence, and θ_G denotes the model parameters. We trained the model using binary cross-entropy (CE) loss:

$$\theta_G^* = \underset{\theta_G}{\operatorname{argmin}} \frac{1}{L} \sum_{i=1}^{L} [L_{bce}(r_i, \hat{r}_i) + L_{bce}(t_i, \hat{t}_i)]$$

where L is the number of training sentences, and r_i and t_i are manual abnormality vectors of the i-th sentence.

We initialized the parameters of the model with $\text{BERT}_{\text{base}}$ pre-trained on Chinese and fine-tuned it on 1,500 sentences in the report training set. The maximum length of the input sentence was set to 128. During training, we used a data augmentation of left-right flipping. There was a 50% chance that characters "left" and "right" in the sentence were replaced by each other, with the region labels replaced by the corresponding regions on the other side of the brain (e.g., label *left frontal lobe* replaced by *right frontal lobe*). We used an Adam optimizer with a learning rate of 0.00005. We used a batch size of 16. An early stopping with the patience of 1 epoch was used based on the loss on the validation set and the maximum training epoch was set to 5. The model with the minimum loss on the report validation set was selected as the final model.

Once trained, the sentence-level prediction model was used to predict abnormality vectors for sentences in unlabeled imaging reports. We binarized the model prediction using the threshold of 0.5. The binary predictions of sentences in each report were then merged into the pseudo abnormality annotation matrix.

Anatomical region segmentation

We used atlas-based segmentation to generate coarse anatomical region segmentation for each CT scan (3D image) in the retrospective training set. 10 CT scans were first randomly selected, registered to a reference CT scan, and then averaged to obtain an average CT image[36] with an image size of 512 × 512 × 32. Anatomical region segmentation of the average image was then annotated by a senior radiologist. To get the segmentation of a target CT image, the average image was registered to it using an affine transformation followed by a B-spline transformation. The same transformations were then applied to the region segmentation mask of the average image, resulting in the segmentation of the target CT. For better registration performance, all CT images were first preprocessed so that the head was located in the middle of the image and was as bilaterally symmetric as possible using a simulated annealing algorithm. We used the Elastix software[33] for the registration and the transformation operation. The coarse segmentations were used to train our image analyzer to dynamically segment the anatomical regions.

Our image analyzer was robust in predicting anatomical region segmentation. Examples of coarse region segmentation and analyzer-predicted region segmentation are shown in Figure S1. As we used an atlas-based segmentation approach that requires little manual annotation, the generated coarse segmentation masks tended to have rough boundaries due to 3D interpolation operations. However, the segmentation predicted by the image analyzer had much smoother and more accurate boundaries. Moreover, the analyzer achieved good segmentation performance even when the atlas-based method generated extremely bad results due to the failure of the atlas-based method (Figure S1D).

As our atlas-based segmentation required a reference scan, the choice of the scan might affect the accuracy of the coarse region segmentation and the resulting image analyzer. To check this, we randomly selected and labeled 3 more CT scans as the reference scan, and trained the image analyzer using coarse region segmentation generated using these reference scans. The results are shown in Table S1. As can be seen, the choice of the reference CT affects the image analyzer's performance in a very limited way, with the AUROC changing by less than 1%.

CT image preprocessing

The CT scans were automatically preprocessed before being fed to the image analyzer. First, for an input CT in DICOM format, we converted it into an 8-bit image by cropping the HU values using a window level of 35 and a window width of 90, i.e., $y = min(max(x, -10), 80)/90 \times 255$, where x denotes the HU value and y denotes the resulting image pixel value. By stacking the slices of a CT scan we got a 3D image of size 512 × 512 × D, where D denotes the number of slices in the scan. Then we automatically cropped the empty bars on the image boundaries and only reserved the head region. We then resized the cropped image to the size of 160 × 160 × 80 using bilinear interpolation. Note that the resulting image might have different spatial resolutions because

of the difference in the spatial resolutions of the original images and the cropping operation. Finally, we normalized the voxel values in the images based on the mean and standard variation of all voxels in the training dataset.

Image analyzer architecture and training

The 3D CT images, along with the pseudo abnormality annotation generated by the discretizer, and the coarse region segmentation generated by the atlas-based method, were used to train an image analyzer, which simultaneously predicted a segmentation map for each anatomical region and a probability map for each type of abnormality.

The image analyzer had a 3D CNN architecture and was trained using our dynamic multi-instance learning approach. The architecture of the image analyzer is shown in Figure S4. The input 3D image first went through a backbone with a 3D U-Net architecture. The output feature map was then passed to an output layer that generated N segmentation maps of anatomical regions and M prediction maps of abnormalities. In order to improve training speed, the number of the upsampling layers is 1 less than the number of the maxpooling layers in the U-Net backbone, resulting in prediction maps with a half spatial resolution of the input image. The image analyzer could be formulated as

$$F(I|\theta_F) = (\widehat{S}, \widehat{D})$$

where $I \in R^{H\times W\times D}$ is the input CT image, θ_F denotes the model parameters, $\widehat{S} \in [0,1]^{\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times N}$ denotes the segmentation predictions of the N anatomical regions, and $\widehat{D} \in [0,1]^{\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times M}$ denotes the prediction maps of the M abnormality types. The anatomical region segmentation was optimized by a segmentation loss. The abnormality prediction was optimized using our dynamic multi-instance approach.

We used Dice loss for the region segmentation task. The segmentation loss of an input CT scan was the average dice loss on all anatomical regions, defined as

$$L_s = \frac{1}{N}\sum_{i=1}^{N} L_{dice}(S_i, \widehat{S}_i)$$

where S_i is the coarse segmentation of the i-th region generated using the atlas-based method, and $\widehat{S}_i$ is the predicted segmentation of the i-th region.

The abnormality prediction task of our image analyzer was based on multi-instance learning (MIL).[37] MIL is a weak supervision learning paradigm, where data is composed of bags with labels, and each bag is composed of several unlabeled instances. In the most typical MIL situation, a bag is positive if and only if at least one of its instances is positive. MIL is usually used in medical studies by viewing each scan or each slice of image as a bag,[38–40] which is hard to utilize the fine-grained localization information provided by the imaging reports.

In this work, we used a fine-grained and dynamic MIL approach (DaMIL), where bags were anatomical regions whose instances were dynamically predicted by the analyzer. Specifically, we regarded each anatomical region as a bag, and voxels in that region as its instances. An anatomical region was considered to have a type of abnormality, if and only if some of its voxels had such abnormality. In each CT image, there would be multiple labeled bags. Apart from predicting the abnormality labels of each bag from the CT image, the labels of instances (i.e., voxel-level prediction) could also be generated.

There are multiple MIL pooling methods to get the final prediction of a bag using the predictions of its instances.[41] In this work, we used the simple max MIL pooling method as a baseline, which means the prediction of a bag was set as the maximum prediction of all its instances. For simplicity, we would refer to the i-th anatomical region as R_i and the j-th abnormality type as A_j. For region R_i, the probability prediction of abnormality type A_j showing in it was defined as

$$\hat{y}_{ij} = \max_{v\in R_i}(\widehat{D}_{j,v})$$

where v denotes voxels in region R_i, and $\widehat{D}_{j,v}$ is the value of the prediction map of A_j at voxel v.

As the coarse region segmentation masks tended to have artifacts on the boundaries because of the registration and transformation process in the atlas-based segmentation, we used the segmentation predicted dynamically by the image analyzer to determine the instances (voxels) in each bag. In practice, the probability of each voxel being an instance of region R_i and at the same time showing abnormality A_j was calculated as the product of these two probabilities. Formally, the probability of abnormality A_j shown in region R_i in an image was calculated as

$$\hat{y}_{ij} = \max_{v\in I}(\widehat{S}_{i,v}\widehat{D}_{j,v})$$

where v denotes a voxel in the 3D CT image I, and $\widehat{S}_{i,v}$ denotes the value of prediction map of region R_i at voxel v.

In this way, we got the region-level abnormality prediction matrix of shape $N \times M$, which was optimized by a weighted cross-entropy (CE) loss. As the frequencies of regions showing abnormalities were not only very low but also varied on different regions and abnormalities, we used different weights for positive and negative samples during the calculation of the cross entropy loss based on the positive rates.[42] The region-level MIL loss of a CT image was defined as

$$L_{ra} = \frac{-1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\sqrt{\frac{1-f_{ij}+\varepsilon}{f_{ij}+\varepsilon}}y_{ij}\ln(\hat{y}_{ij}) + (1-y_{ij})\ln(1-\hat{y}_{ij})\right]$$

where f_{ij} is the frequency of abnormality A_j showing in region R_i, ε is a small number, y_{ij} is the pseudo binary label, and $\hat{y}_{ij}$ is the prediction of the image analyzer.

Using the simple max MIL pooling would already result in accurate region-level abnormality detection. However, when the abnormality region is large, it is sometimes hard for the prediction map to cover the full abnormality region. To tackle this problem, we used a multi-scale output approach as we found that this problem was alleviated when the output had lower spatial resolution. An output layer was added to each of the last few deconvolution layers in the U-Net, generating region segmentation and abnormality prediction maps of different scales. The segmentation loss of each scale was the same as the baseline, despite having different spatial resolutions. The MIL loss of the first scale (the scale with the lowest output resolution) was also the same as the baseline. The MIL loss of the l-th scale (l > 1) was calculated with the assistance of the abnormality prediction of the former scale l-1.

In the l-th scale (l > 1), the max MIL pooling was not used anymore to get the prediction of each bag. Instead, we used the average prediction of several instances with the largest prediction values as the prediction of the bag. This helped the analyzer output smoother maps by paying attention to multiple voxels instead of focusing on a single voxel. The number of instances used to calculate the bag-level prediction was determined by the number of instances whose prediction was above a certain threshold in scale l-1. The prediction of abnormality A_j showing in region R_i in the l-th scale was calculated as

$$\hat{y}_{l,i,j} = mean\left(TOP_{k_{l,i,j}}\left(\hat{S}_i^l\hat{D}_j^l\right)\right)$$

where $TOP_k()$ denotes the operation of selecting the top k values in a 3D image, $k_{l,i,j}$ is the number of instances in region R_i whose prediction of abnormality A_j is above a threshold t in the former scale $l-1$:

$$k_{l,i,j} = \sum_{v\in l}1\left(\hat{S}_{i,v}^{l-1}\hat{D}_{j,v}^{l-1} > t\right)$$

For a K-scale analyzer, the abnormality MIL loss of an image was defined as the average MIL losses on all scales:

$$L_{ra} = \frac{-1}{KNM}\sum_{l=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\sqrt{\frac{1-f_{ij}+\varepsilon}{f_{ij}+\varepsilon}}y_{ij}\ln(\hat{y}_{l,i,j}) + (1-y_{ij})\ln(1-\hat{y}_{l,i,j})\right]$$

In this work, we set the value of K as 2. The 2 outputs each have a resolution of 40 × 40 × 20 and 80 × 80 × 40 respectively. The prediction of the highest scale is used as the final prediction of the model.

Besides the region-level MIL loss, we also used a scan-level MIL loss for each scan defined as

$$L_{sa} = \frac{-1}{M}\sum_{i=1}^{M}\left[y_i^s\ln(\hat{y}_i^s) + (1-y_i^s)\ln(1-\hat{y}_i^s)\right]$$

where y_i^s is the pseudo scan-level label of the i-th abnormality type calculated using the union (max) of abnormality labels on all regions (including *other region* and *unclear region*), and $\hat{y}_i^s$ is the scan-level prediction of the i-th abnormality type. We trained the analyzer using a linear combination of the segmentation loss, the region-level MIL loss, and the scan-level MIL loss:

$$\theta_F^* = \underset{\theta_F}{\operatorname{argmin}}\frac{1}{L}\sum_{i=1}^{L}\left(\lambda L_s^{(i)} + L_{ra}^{(i)} + \alpha L_{sa}^{(i)}\right)$$

where L is the size of the training set, λ and α are two hyper-parameters to balance these terms. We used a dynamic weight λ in this work. The segmentation task was simpler and easier to converge than the abnormality prediction task, and the abnormality prediction task relied on the segmentation task to determine instances in each bag. Therefore, we used a larger weight λ in the beginning. During the training process, λ was exponentially decayed, in order to increase the influence of the abnormality prediction task.

The analyzer was trained in an end-to-end manner on the retrospective training set. During training, the MIL loss of scans with a positive *unclear region* label was set to 0 to avoid misleading information. We used augmentations of left-right flipping and axial rotation. There was a 50% chance that the input 3D image was left-right flipped, along with the region segmentation maps and the abnormality labels. The input image was then randomly rotated in the axial view clockwise or counter-clockwise by up to 15°. We used an Adam optimizer whose initial learning rate was set to 0.0005 and decayed by half every epoch since the 3rd epoch. The weight λ was initialized as 1, and reduced by half every epoch until reaching the minimum value of 0.01. The weight α was set to 0.1. We used a batch size of 1 because of the memory limitation. The model with the lowest region-level MIL loss on the validation set was saved for final evaluation. We used an early stopping with the patience of 3 epochs. The model was trained on an NVIDIA Tesla V100 GPU with 32GB memory. The full training process took around 48 h. The model was implemented using TensorFlow and Keras.

Ablation study

We justify our design of the framework using a series of experiments.

First, We show that a small number of manual annotations were sufficient for training a discretizer that generated pseudo annotations well enough. We trained the discretizer by varying the size of the training report dataset. Then we trained the image analyzer using pseudo annotations generated by the corresponding discretizer. Figures S2A and S2B show the results of the discretizer and image analyzer in detecting different abnormalities using different report training sizes. Our framework could achieve high and stable performance using only 1,500 sentences to train the discretizer, demonstrating its efficiency in learning from scarce manual annotation.

To show the generalizability of our discretizer, we manually labeled the silver-standard annotations for reports in the prospective dataset and tested the discretizer. The result is shown in Figure S2H. The discretizer's performance to extract regional labels from report sentences was similar in the retrospective and prospective datasets. However, the performance to predict abnormality labels decreased. This suggests that the distribution of the reports might shift over time. However, as the discretizer could be trained using a very small amount of data, re-training or finetuning would be convenient.

To demonstrate the need for the discretizer and the superiority of training on large-scale pseudo annotations over a small amount of more precise annotations, we manually labeled the silver-standard abnormality annotations of 1,169 randomly selected CT reports (which contain 4,831 sentences, similar to the scale of our report dataset), and trained the image analyzer on these data. The average AUROCs of the resulting image analyzer to detect the four abnormalities over the 17 regions were 0.624 (95% CI = 0.605–0.643), 0.735 (95% CI = 0.721–0.750), 0.702 (95% CI = 0.684,0.719), and 0.768 (95% CI = 0.750,0.786) respectively, which were significantly lower than the performance of the image analyzer trained on the 28,472 CT scans with pseudo annotations ($p < 0.001$).

Then we show that the image analyzer's performance continuously improved as the size of the training size grew. Figures S2C and S2D shows the performance change of our image analyzer using 2,000, 5,000, 10,000 and 28,472 training samples. As the size of the training dataset increased, the image analyzer performed better in detecting all 4 types of abnormalities (Figure S2C). The AUROC of extraparenchymal hypodensity, the abnormality with the lowest positive rate, showed the largest relative increase of 9.86% among all 4 abnormalities, increasing from 0.882 (95% CI = 0.863–0.900) when trained with 2,000 scans to 0.969 (95% CI = 0.956–0.979) when trained with 28,472 scans. The performances of the image analyzer in detecting abnormalities on different anatomical regions also increased as the size of the training dataset increased (Figure S2D), except for regions with extremely low abnormal rates such as brain stem. It is foreseeable that with more raw data automatically annotated and added to the training dataset, the performance of the image analyzer would continue to improve.

Next, we show that the fine-grained, region-level abnormality annotations allowed us to utilize the abnormality location information in imaging reports and improved the performance of the image analyzer, especially its ability to localize abnormalities. We show the performance of the analyzer trained using region-level annotations (the region-level analyzer) and the analyzer trained using the scan-level annotations (the scan-level analyzer) in Figures S2E and S2F. The scan-level analyzer was slightly inferior to the scan-level analyzer in detecting scan-level abnormalities. However, the abnormality localization performance was seriously degraded in the scan-level analyzer (Figure S2F). For detecting region-level abnormalities, the scan-level analyzer yielded much lower AUROC than the region-level analyzer, with a relative decrease in AUROC of 28.7%, 24.8%, 34.4% and 26.3% in detecting the four abnormalities in region-level, respectively. The reason was that although the scan-level analyzer accurately predicted the scan-level existence of abnormalities, the localization map tended to be activated at a certain location rather than the correct locations of abnormalities.

To show the superior performance of our DaMIL method, we compared the performance of our image analyzer trained using DaMIL with the following models. A, a classification model. The classification model had the same convolutional layers as the DaMIL model. Unlike the DaMIL model, it did not predict voxel-level predictions but had $(N + 1) \times M$ binary classification heads to classify if each of the N anatomical regions and the entire scan has each of the M abnormality types. B, a classification model with spatial attention. It was the same as the classification model except for having spatial attention for each region before the classification heads. C, a region-level MIL model. This model predicted voxel-level abnormality maps and was trained using region-level abnormality annotations. Unlike the DaMIL model, it did not predict the anatomical region segmentation dynamically but used the coarse region segmentation to determine the instances in each bag. D, the DaMIL model, except that during inference the voxels in each anatomical region were determined by the coarse anatomical segmentation instead of the anatomical segmentation predicted by the model. The average AUROCs of these models (image analyzers) to detect the four abnormality types over the 17 regions are shown in Table S2. As can be seen, the MIL-based models (C, D, and DaMIL) performed better than the classification-based models (A and B) as they utilized the information of the anatomical region segmentation. Our dynamic MIL method was superior to the traditional region-level MIL method (C) as it dynamically determined the instances in each bag during training and inference, which were supervised not only by the coarse region segmentation but also by the abnormality annotations, thus fixing the potential errors in the coarse region segmentation. For the DaMIL-trained model, using model-predicted region segmentation during inference was better than using the coarse segmentation generated by the atlas-based method (D vs. DaMIL).

We also compared the abnormality localization performance of our DaMIL analyzer with the GradCAM method.[13] Specifically, we used GradCAM to generate the 3D activation maps for the scan-level predictions of the classification model for each abnormality type. The GradCAM activation maps achieved Dice scores of 0.014 (95% CI = 0.007–0.021), 0.051 (95% CI = 0.017–0.075), 0.054 (95% CI = 0.020–0.103), and 0.110 (95% CI = 0.020–0.203), which were on average 83.1% lower than our method.

Finally, we show that training the analyzer simultaneously on multiple anatomical regions improved its performance. As a certain type of abnormality shows common features as it appears in different regions in the brain (e.g., intraparenchymal hyperdensity in the frontal lobe would look similar to intraparenchymal hyperdensity in the occipital lobe), the feature learned by the analyzer in one region should help abnormality detection in other regions. To check this, we trained the image analyzer on each single region (single-region analyzers) as opposed to training on multiple regions (the multi-region analyzer). The performances of the single-region analyzers are shown in Figure S2G, compared with the performance of the multi-region analyzer. The multi-region analyzer achieved a higher average AUROC in 15 out of 17 regions, except for lateral ventricles, probably due to the appearance difference of ventricles from the brain parenchyma.

QUANTIFICATION AND STATISTICAL ANALYSIS

To quantify the abnormality detection performance we report the area under the receiver operating characteristic curve (AUROC). The discretizer predicted pseudo abnormality annotations from reports, and the image analyzer predicted region-level abnormality existences from CT scans. Both of these predictions were matrices of shape $N \times M$ suggesting the existence of M types of abnormalities in N anatomical regions, which were compared to our gold-standard annotations for evaluation. Besides reporting the AUROC on each element of the matrices (i.e., the AUROC of an abnormality in a region, Table S6), we summarize the AUROC of these matrices in two dimensions. First, we report the performance to detect each abnormality by averaging the AUROC over different regions. Second, we report the performance on each region by averaging the AUROC over different abnormalities. For simplicity of presentation, when summarizing the performance on each region, the AUROC on symmetrically corresponding regions were also averaged, as the performances on these regions were similar (i.e., performance on the left frontal lobe and the right frontal lobe were averaged into 'frontal lobe'). The averaging operations were done in a micro manner, meaning that the ground truth and predictions on several groups were combined together to draw a ROC curve the area under which was then computed. Despite the above AUROCs, we also report the scan-level AUROC for each abnormality type. For each scan, the scan-level ground truth of an abnormality type denotes if any anatomical region has that abnormality, and the scan-level prediction is acquired using the maximum of the prediction map of that abnormality type.

To evaluate the abnormality segmentation performance of the image analyzer, we computed the Dice score on 40 randomly selected CT scans from the retrospective test dataset whose abnormalities were manually labeled by a senior radiologist. For each abnormality, the Dice score was computed only on the scans that had that abnormality. We did not use the whole test set due to the expensive voxel-level annotation cost.

Statistical analysis was performed using SciPy.stats package in Python. To compute the 95% confidence intervals of the AUROCs we employed bootstrap with a resampling size of n = 1000. For the prioritization evaluation, we used a one-sided Wilcoxon rank-sum test to check if the rankings of CTs with a certain disease are higher than those without the disease, with a significance threshold of $p < 0.001$.

Cell Reports Medicine, Volume *4*

Supplemental information

Automatic intracranial abnormality detection and localization in head CT scans by learning from free-text reports

Aohan Liu, Yuchen Guo, Jinhao Lyu, Jing Xie, Feng Xu, Xin Lou, Jun-hai Yong, and Qionghai Dai

Supplemental Figures

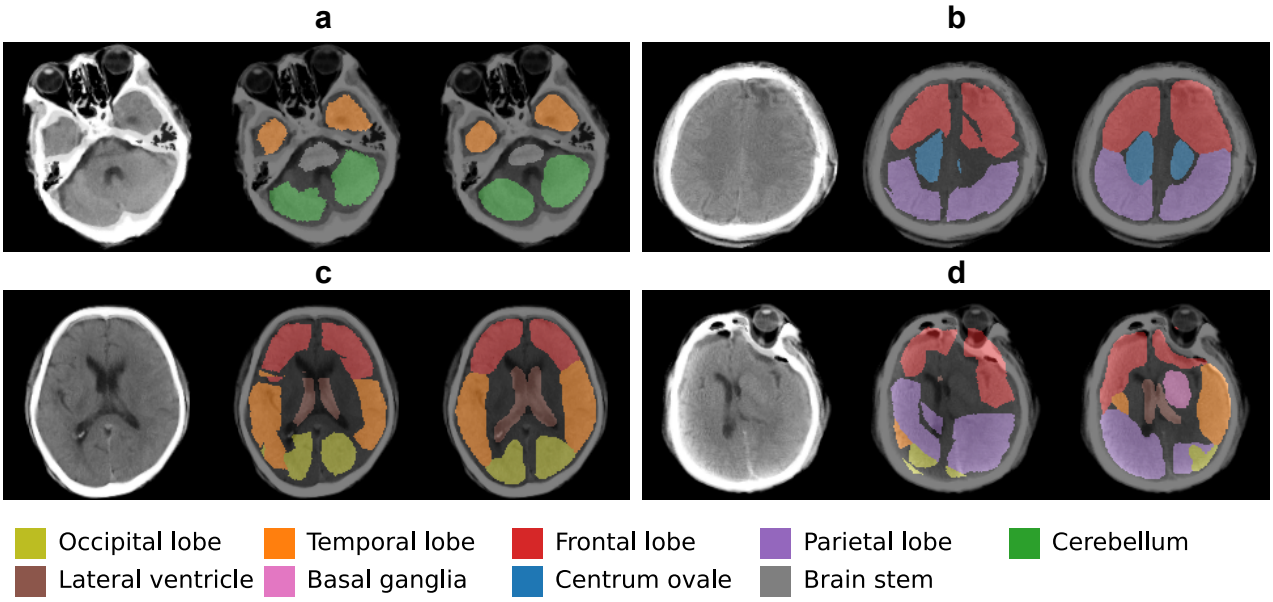


Figure S1: **Anatomical region segmentation results. Related to Figure 1 and STAR Methods. a-d**, 4 slice-level examples of the coarse segmentation masks and analyzer-predicted segmentation masks on the retrospective test set. Each example contains a slice of the input CT on the left, the corresponding coarse segmentation in the middle, and the prediction of the image analyzer on the right. Different colors denote different anatomical regions, as shown at the bottom. The coarse segmentation in **d** was seriously flawed due to registration failure in the atlas-based segmentation process, yet the analyzer predicted good segmentation results.

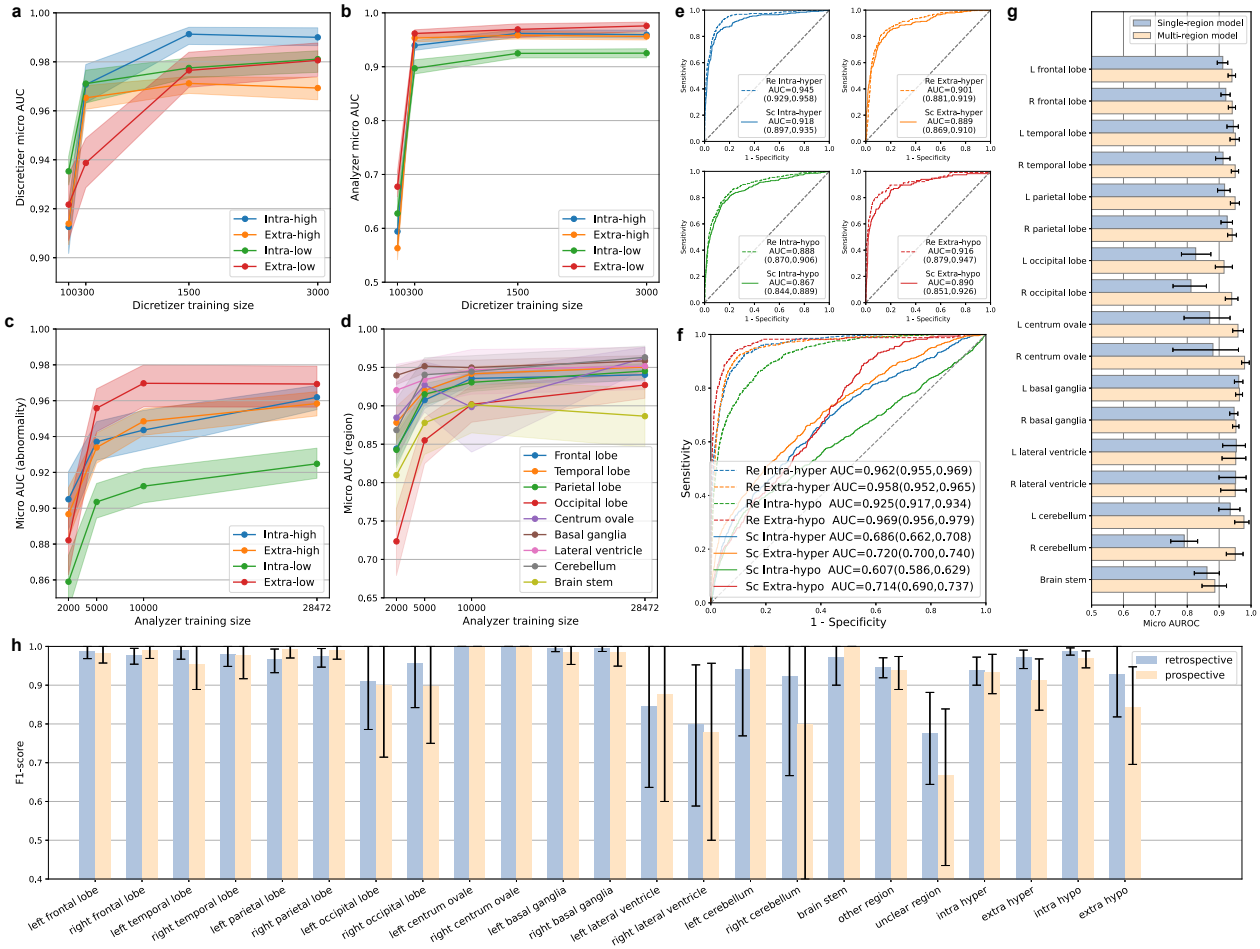


Figure S2: **Supplemental experiments. Related to Figure 2, Figure 3 and STAR Methods.** **a**, average AUROCs of the discretizer in detecting four abnormalities from reports under different training sizes. **b**, average AUROCs of the image analyzer in detecting four abnormalities from CT scans when trained with pseudo annotations generated by the discretizers trained by different sizes of training sets. **c**, change of average AUROC of the image analyzer in detecting four abnormalities in all regions as the size of the training set increases. **d**, change of average AUROC of the image analyzer in detecting abnormalities in different regions as the size of the training set increases. **e**, ROC curves of the scan-level analyzer (Sc) and region-level analyzer (Re) to detect scan-level abnormalities. **f**, average ROC curves of the scan-level analyzer and the region-level analyzer to detect four types of abnormalities in different regions. **g**, AUROC of single-region analyzers and the multi-region analyzer in detecting abnormalities on the 17 regions, respectively. All results in **a-g** were evaluated on the retrospective test set. **h**, F1-score of the discretizer to extract labels from report sentences in the retrospective report test set and prospective dataset.

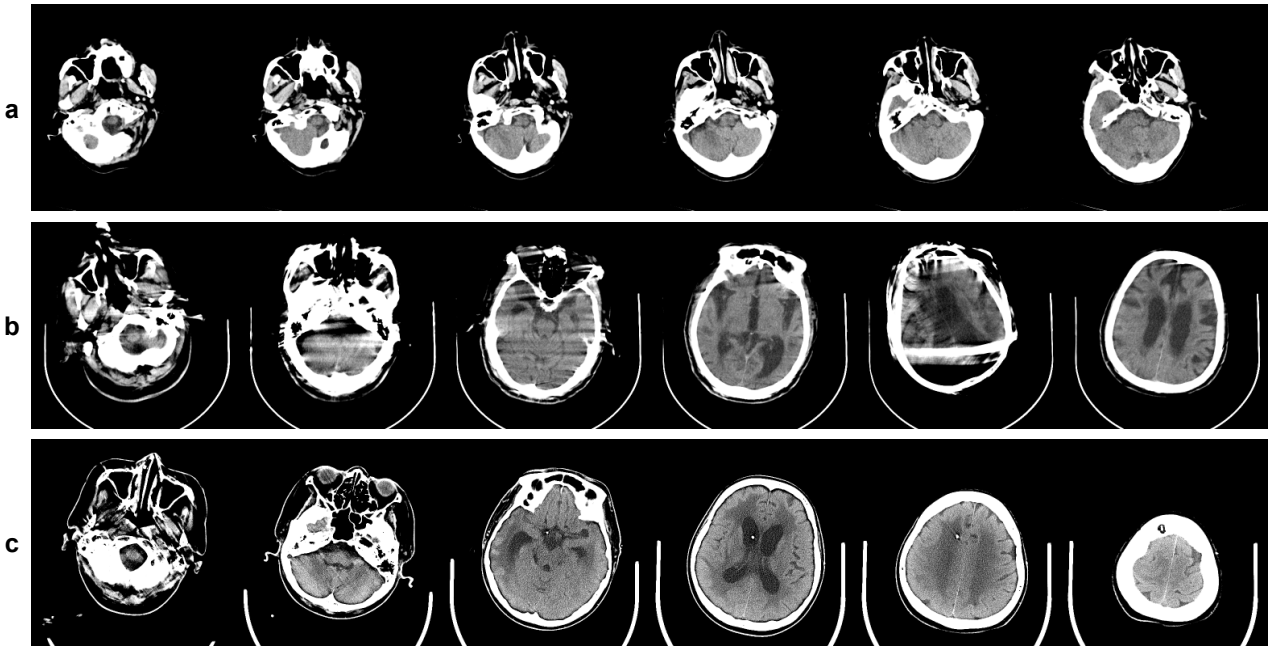


Figure S3: **Example of removed scans. Related to STAR Methods.**. **a**, The scan contains very few slices and does not cover the full brain area. **b**, there are extreme motion artifacts in the scan. **c**, the scan was reconstructed for bone window and has a low SNR under soft tissue window.

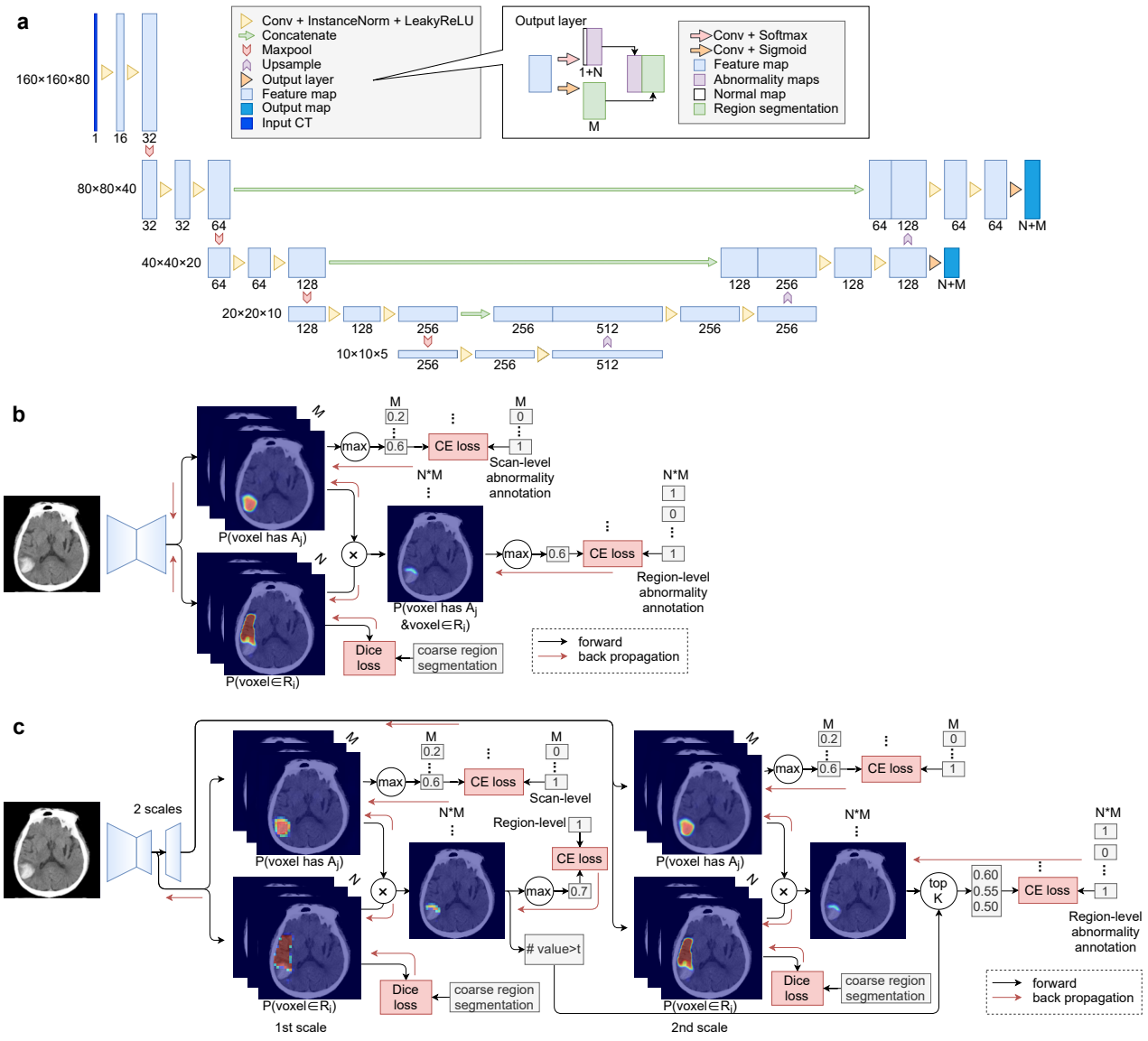


Figure S4: **Image analyzer illustration. Related to STAR Methods.** **a**, Image analyzer architecture. **b**, the single-scale training of the image analyzer. **c**, the multi-scale training of the image analyzer. For simplicity, we use 2D images in **b** and **c**, but the image analyzer actually processes 3D images.

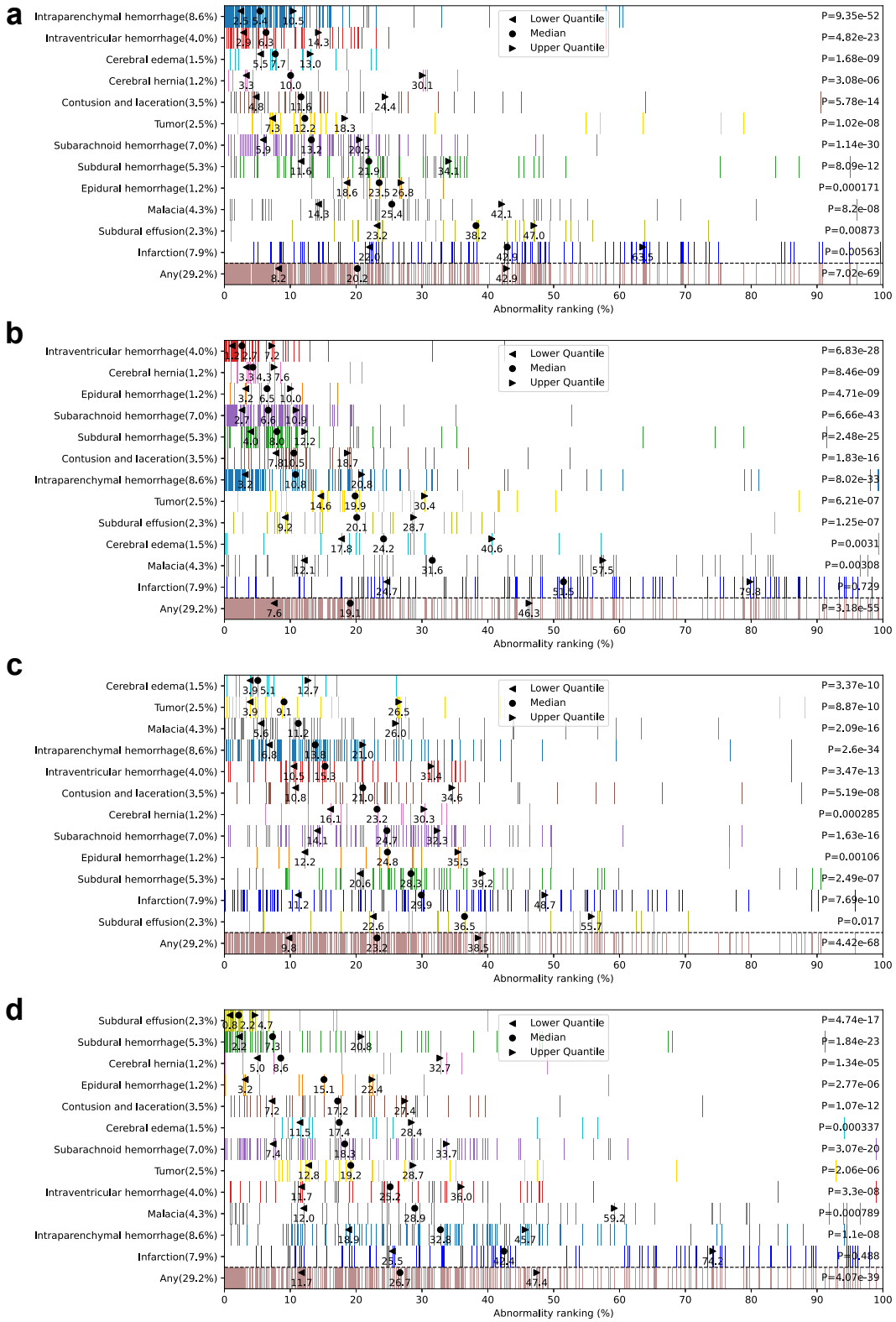


Figure S5: **Prioritization performance using scores of different abnormalities. Related to Figure 5. a**, intraparenchymal hyper-density. **b**, extraparenchymal hyper-density. **c**, intraparenchymal hypo-density. **d**, extraparenchymal hypo-density.

Supplemental Tables

Table S1: AUROC (%) of image analyzers trained using coarse segmentation generated by different reference CT scans to detect the four abnormality types over the 17 regions. Related to Figure 3.

Reference CT	Intra-hyper	Extra-hyper	Intra-hypo	Extra-hypo
Original	96.2 (95.5-96.9)	95.8 (95.2-96.5)	92.5 (91.7-93.4)	96.9 (95.6-97.9)
R1	95.5 (94.5-96.3)	94.9 (94.1-95.7)	92.4 (91.4-93.3)	96.6 (95.3-97.8)
R2	96.0 (95.3-96.7)	95.2 (94.4-96.0)	92.7 (91.9-93.6)	96.9 (95.6-98.1)
R3	96.1 (95.5-96.7)	95.6 (94.9-96.3)	92.0 (91.1-92.9)	96.8 (95.5-97.9)

Table S2: AUROC (%) of different methods to detect the four abnormality types over the 17 regions. Related to Figure 3.

Method	Intra-hyper	Extra-hyper	Intra-hypo	Extra-hypo
A	88.2 (86.8-89.5)	91.5 (90.4-92.4)	87.6 (96.5-88.7)	95.7 (94.6-96.8)
B	88.6 (87.3-90.0)	92.4 (91.5-93.2)	88.0 (86.8-89.0)	94.0 (92.7-95.2)
C	94.2 (93.2-95.1)	92.8 (91.9-93.7)	90.3 (89.3-91.4)	93.4 (91.8-94.9)
D	94.3 (93.3-95.3)	93.8 (92.9-94.6)	91.5 (90.6-92.5)	93.8 (92.0-95.4)
DaMIL	96.2 (95.5-96.9)	95.8 (95.2-96.5)	92.5 (91.7-93.4)	96.9 (95.6-97.9)

Table S3: Rules used during data collection and processing, written in Python-like pseudo code. Related to Table 1 and STAR Methods.

Process	Rule
a. CT collection	dicom.Modality==‘CT’ and dicom.SeriesDescription in [‘Recon 2:’,‘HeadSeq 4.8 H31s’, ‘HeadSeq 2.4 H30s’, ‘HEAD 5mm’, ‘Recon 3: 5mm/ 10mm’, ‘5mm/ 10mm’, ‘Add Scan 4.8 H30s’, ‘Add Scan 4.8 H31s’, ‘4.8 x 4.8’, ‘CTA 5mm’, ‘CerebrumSeq 6.0 H31s’, ‘CerebrumSeq 4.0 H31s’, ‘Add Scan 6.0 H31s’, ‘HeadSeq 2.4 H31s’, ‘3D_Batch1’, ‘Head.Seq 4.8 H31s’, ‘Head 4.8 H23s’, ‘Head 4.8 H31s’]
b. Report collection	EXAM_CLASS==‘CT’ and (EXAM_SUB_CLASS==‘head and neck’ or EXAM_SUB_CLASS==‘brain’) and not any(x in EXAM_PARA for x in [‘sinuses’, ‘maxillofacial region’, ‘orbital cavity’, ‘parotid gland’, ‘temporal bone’, ‘optic nerve’, ‘nasal bone’, ‘CTP’, ‘cervical vertebra’, ‘temporomandibular joint’, ‘vertebra’])

Table S4: Definition and typical report description of anatomical region labels. Related to STAR Methods.

Label	Definition	Typical description
Frontal lobe (L/R) Parietal lobe (L/R) Temporal lobe (L/R) Occipital lobe (L/R)	Anywhere within the parenchyma of these lobes or anywhere between these lobes and the cranial plate (including subdural, epidural, subarachnoid spaces).	(Take frontal lobe as an example) 'frontal lobe', 'frontal white matter', 'frontal cortex', 'under frontal cranial plate', 'frontal subdural space', 'frontal epidural space', 'frontal subarachnoid space', 'frontal sulcus'
Centrum ovale (L/R)	Anywhere within the centrum ovale region.	'centrum ovale'
Basal ganglia (L/R)	Anywhere within the basal ganglia region.	'basal ganglia region', 'external capsule', 'internal capsule', 'lenticular nucleus', 'caudate nucleus', 'globus pallidus'
Lateral ventricle (L/R)	Anywhere within the lateral ventricle.	'lateral ventricle', 'anterior horn of lateral ventricle', 'posterior horn of lateral ventricle'
Cerebellum (L/R)	Anywhere within the parenchyma of the cerebellum or anywhere between the cerebellum and the cranial plate.	'cerebellum', 'cerebellum sulcus'
Brain stem	Anywhere within the brain stem.	'brain stem', 'pons', 'midbrain', 'medulla oblongata'
Other region	Any intracranial region outside our defined anatomical regions.	'cerebral falx', 'lateral fissure cistern', 'third ventricle', 'fourth ventricle', 'corona radiate'
Unclear region	Any intracranial region that might contain our defined anatomical regions, but is not clearly described in the report.	'left hemisphere', 'surgery area'

Table S5: Definition and typical report description of abnormality labels. Related to STAR Methods.

Label	Definition	Typical description
Intraparenchymal	That is inside the parenchyma.	'in the left frontal lobe' (Take the left frontal lobe as an example)
Extraparenchymal	That is outside the parenchyma (in epidural, subdural, subarachnoid spaces, cistern or ventricle)	'under frontal cranial plate', 'in frontal subdural space', 'in frontal epidural space', 'in frontal subarachnoid space', 'in frontal sulcus', 'in the left ventricle', 'in lateral fissure cistern'
Hyper-density	Abnormalities that appear or partly appear as high density on CT, except for calcification and artifacts.	'hyper-density', 'increase in density', 'hemorrhage', 'hematocele', 'hematoma', 'mixed density'
Hypo-density	Abnormalities that appear or partly appear as low density on CT, except for those caused by gas, fat, or ischemia lesion.	'hypo-density', 'decrease in density', 'edema', 'fluid density', 'mixed density'

Table S6: AUROCs of the image analyzer in detecting abnormalities in the retrospective test dataset. Related to Figure 3.

AUROC (%)	Intra-hyper	Extra-hyper	Intra-hypo	Extra-hypo
L frontal lobe	93.7 (91.0-95.9)	94.9 (92.8-96.5)	89.1 (85.4-92.3)	98.1 (97.4-98.8)
R frontal lobe	95.8 (94.5-97.2)	92.8 (90.5-95.0)	93.3 (90.9-95.4)	94.5 (92.1-96.6)
L temporal lobe	96.3 (94.5-97.9)	95.6 (92.6-97.6)	91.4 (87.8-94.5)	96.5 (91.6-99.0)
R temporal lobe	94.6 (92.2-96.5)	95.6 (93.9-97.0)	96.3 (94.2-98.1)	94.2 (91.7-96.5)
L parietal lobe	95.4 (93.5-97.1)	95.1 (92.5-97.0)	88.1 (81.8-93.4)	98.6 (97.9-99.2)
R parietal lobe	94.6 (92.4-96.7)	95.3 (93.8-96.7)	89.6 (83.4-94.0)	95.0 (91.9-97.6)
L occipital lobe	91.4 (84.8-96.2)	83.8 (73.0-92.4)	90.1 (84.8-94.7)	98.2 (96.9-99.3)
R occipital lobe	93.6 (88.1-97.9)	92.8 (89.8-95.1)	92.3 (85.4-97.5)	96.5 (94.2-98.5)
L centrum ovale	97.9 (94.8-99.5)	-	86.7 (80.3-92.9)	-
R centrum ovale	-	-	92.1 (88.5-96.3)	-
L basal ganglia	97.2 (93.7-99.5)	-	86.5 (83.0-89.7)	-
R basal ganglia	94.6 (91.8-97.0)	-	86.8 (83.6-89.6)	-
L lateral ventricle	-	95.0 (92.7-97.0)	-	55.1 (32.1-86.5)
R lateral ventricle	-	94.0 (89.7-97.0)	-	50.3 (18.1-99.2)
L cerebellum	98.2 (96.1-99.6)	76.0 (51.9-98.9)	98.0 (96.5-99.3)	-
R cerebellum	96.9 (94.4-99.0)	86.7 (81.4-93.0)	93.6 (89.0-97.3)	-
Brain stem	88.7 (79.5-96.5)	-	72.1 (61.1-82.0)	-
Weighted macro	95.1 (94.1-95.9)	94.1 (93.1-95.1)	89.7 (88.4-91.1)	95.4 (93.9-96.7)
Micro	96.2 (95.5-96.9)	95.8 (95.2-96.5)	92.5 (91.7-93.4)	96.9 (95.6-97.9)

L: left. R: right. weighted macro: macro AUROC weighted by abnormality positive rates. All results were obtained on the retrospective test set. 95% confidence intervals were computed using bootstrapping over n=1,000 seeds. As by our definition some abnormalities would never occur in some specific regions (e.g., intraparenchymal abnormalities would not occur in ventricles), some cells remain empty.

Table S7: AUROC of the image analyzer in detecting abnormalities on the prospective set. Related to Figure 3.

AUROC (%)	Intra-hyper	Extra-hyper	Intra-hypo	Extra-hypo
L frontal lobe	91.8 (88.6-94.7)	92.8 (90.1-95.2)	90.7 (88.0-93.0)	93.9 (90.3-97.0)
R frontal lobe	90.3 (86.4-93.8)	89.3 (85.3-92.7)	89.7 (86.7-92.5)	94.0 (90.9-96.6)
L temporal lobe	93.0 (89.0-96.3)	93.9 (91.3-96.4)	91.5 (87.2-95.0)	93.2 (90.7-95.6)
R temporal lobe	94.1 (90.2-96.8)	92.9 (89.7-95.7)	93.2 (89.7-96.1)	95.3 (93.0-97.3)
L parietal lobe	90.7 (84.6-95.8)	91.8 (88.8-94.5)	88.6 (83.9-92.3)	93.4 (88.5-97.1)
R parietal lobe	90.8 (85.3-95.1)	90.7 (87.1-94.0)	88.8 (83.9-93.1)	94.9 (92.4-97.1)
L occipital lobe	92.7 (87.3-97.3)	92.5 (87.7-96.2)	91.0 (86.1-95.3)	81.1 (56.7-94.9)
R occipital lobe	88.0 (81.7-93.8)	87.2 (80.0-93.8)	87.3 (79.9-94.0)	86.0 (73.5-96.0)
L centrum ovale	84.1 (82.2-85.9)	-	79.6 (67.0-89.7)	-
R centrum ovale	80.1 (50.6-99.1)	-	75.3 (62.1-86.8)	-
L basal ganglia	88.6 (82.1-94.0)	-	78.3 (73.2-83.1)	-
R basal ganglia	88.0 (81.7-93.4)	-	78.3 (72.8-83.6)	-
L lateral ventricle	-	97.4 (95.5-98.9)	-	15.7 (13.0-18.4)
R lateral ventricle	-	97.2 (96.0-98.3)	-	20.8 (18.1-23.4)
L cerebellum	97.1 (93.8-99.2)	95.0 (93.9-96.1)	92.4 (87.1-97.0)	96.2 (91.5-100.0)
R cerebellum	99.2 (98.5-99.8)	-	94.3 (87.2-99.3)	-
Brain stem	75.2 (44.3-98.4)	-	80.1 (64.4-93.7)	94.1 (92.9-95.3)
Weighted macro	91.4 (89.8-92.9)	92.9 (91.2-94.1)	87.9 (86.1-89.4)	92.7 (90.5-94.2)
Micro	93.5 (92.4-94.5)	95.6 (95.0-96.2)	91.0 (90.0-91.9)	95.7 (94.5-96.8)