Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study

Yuchen Guo*, Yuwei He*, Jinhao Lyu*, Zhanping Zhou*, Dong Yang, Liangdi Ma, Hao-tian Tan, Changjian Chen, Wei Zhang, Jianxing Hu, Dongshan Han, Guiguang Ding, Shixia Liu, Hui Qiao, Feng Xu, Xin Lou, Qionghai Dai

Lancet Digit Health 2022; 4: e584–93

Published Online June 17, 2022 https://doi.org/10.1016/S2589-7500(22)00090-5

This online publication has been corrected. The corrected version first appeared at thelancet.com/digital-health on June 30, 2022

*Co-first authors

Institute for Brain and Cognitive Sciences, BNRist (Y Guo PhD, Y He PhD, Z Zhou BE, D Yang ME, L Ma MS, H-t Tan BE, C Chen BE, Prof G Ding PhD, Prof S Liu PhD, H Qiao PhD, Prof F Xu PhD, Prof Q Dai PhD), **School of Software** (Y He, Z Zhou, D Yang, L Ma, H-t Tan, C Chen, Prof G Ding, Prof S Liu, Prof F Xu), **and Department of Automation, BLBCI** (H Qiao, Prof Q Dai), **Tsinghua University, Beijing, China; Department of Radiology, Chinese PLA General Hospital, Beijing, China** (J Lyu MS, J Hu BS, D Han MS, Prof X Lou MD); **Department of Radiology, Brain Hospital of Hunan Province, Hunan, China** (W Zhang BS)

Correspondence to:
Prof Feng Xu, School of Software, Tsinghua University, Beijing 100084, China
feng-xu@tsinghua.edu.cn

or

Prof Xin Lou, Department of Radiology, Chinese PLA General Hospital, Beijing 100853, China
louxin@301hospital.com.cn

or

Prof Qionghai Dai, Department of Automation, BLBCI, Tsinghua University, Beijing 100084, China
daiqh@tsinghua.edu.cn

Summary

Background A large training dataset with high-quality annotations is necessary for building an accurate and generalisable deep learning system, which can be difficult and expensive to prepare in medical applications. We present a novel deep-learning-based system, requiring no annotator but weak annotation from a diagnosis report, for accurate and generalisable performance in detecting multiple head disorders from CT scans, including ischaemia, haemorrhage, tumours, and skull fractures.

Methods Our system was developed on 104 597 head CT scans from the Chinese PLA General Hospital, with associated textual diagnosis reports. Without expert annotation, we used keyword matching on the reports to automatically generate disorder labels for each scan. The labels were inaccurate because of the unreliable annotator-free strategy and inexact because of scan-level annotation. We proposed RoLo, a novel weakly supervised learning algorithm, with a noise-tolerant mechanism and a multi-instance learning strategy to address these issues. RoLo was tested on retrospective (2357 scans from the Chinese PLA General Hospital), prospective (650 scans from the Chinese PLA General Hospital), cross-centre (1525 scans from the Brain Hospital of Hunan Province), cross-equipment (1484 scans from the Chinese PLA General Hospital), and cross-nation (CQ500 public dataset from India) test datasets. Four radiologists were tested on the prospective test dataset before and after viewing system recommendations to assess whether the system could improve diagnostic performance.

Findings The area under the receiver operating characteristic curve for detecting the four disorder types was 0·976 (95% CI 0·976–0·976) for retrospective, 0·975 (0·974–0·976) for prospective, 0·965 (0·964–0·966) for cross-centre, and 0·971 (0·971–0·972) for cross-equipment test datasets, and 0·964 (0·964–0·966) for CQ500 (with only haemorrhage and fracture). The system achieved similar performance to four radiologists and helped to improve sensitivity and specificity by 0·109 (95% CI 0·086–0·131) and 0·022 (0·017–0·026), respectively.

Interpretation Without expert annotated data, our system achieved accurate and generalisable performance for head disorder detection. The system improved the diagnostic performance of radiologists. Because of its accuracy and generalisability, our computer-aided diganostic system could be used in clinical practice to improve the accuracy and efficiency of radiologists in different hospitals.

Funding National Key R&D Program of China, National Natural Science Foundation of China, and Beijing Natural Science Foundation.

Introduction

Head disorders, such as brain ischaemia, haemorrhage, tumours, and skull fracture, greatly affect the structure and function of the head and brain, causing high morbidity and mortality.[1] CT scanning has been serving as the frontline diagnostic modality to assess head abnormalities.[2] However, this technique is challenging and labour-intensive for radiologists, and many low-income and middle-income regions are short of experienced radiologists. Therefore, developing computer-aided diagnosis systems to detect multiple disorder types is of practical significance in medical applications.

A clinically applicable computer-aided diagnosis system should be accurate and generalise well across different centres and different CT equipment. To enable these features by deep learning, a large-scale training dataset is needed to cover sufficient diversity.[3] Moreover, the training dataset should have high-quality annotations so that deep networks can learn correct knowledge by supervised learning and generate accurate predictions.[4] Collecting a large medical image dataset for common diseases can be cost free, for example, by retrieving CT scans from picture archiving and communication systems. Unfortunately, it is difficult and expensive to precisely annotate a large-scale training dataset as this

Research in context

Evidence before this study

We searched PubMed and Google Scholar for studies related to the use of artificial intelligence (AI) and deep learning on CT for the diagnosis of head disorders published between Jan 1, 2017, and March 15, 2022, using the search terms "brain disorder", "head disorder", "brain injury", "hemorrhages", "fractures", "ischemia", "stroke", "deep learning", "weakly supervised learning", or "artificial intelligence", with no language restrictions. One study reported an area under the curve (AUC) of 0·991 (SD 0·006) for acute intracranial haemorrhage in an intra-centre test, and no cross-centre test was done. The model in this previous study was trained with 4396 expert-annotated CT scans. Another study reported an AUC of 0·942 (95% CI 0·919–0·965) for intracranial haemorrhage detection and 0·962 (0·920–1·000) for fracture detection in a cross-centre test done on the CQ500 dataset (in India), in which the model was trained with 5423 expert-annotated CT scans. A third study reported an AUC of 0·961 (95% CI 0·927–0·986) for intracranial haemorrhage detection in a prospective test set, in which the model was trained on 904 expert-annotated CT scans. The remaining studies identified by our search did not achieve the results observed in the research mentioned above using CT with AI or deep learning. All studies we identified focused on no more than two types of disorder, and required heavy expert annotation. We found no publications that reported how AI systems could assist and improve radiologist performance.

Added value of this study

In this study, we reported a novel annotator-free deep learning system using weak annotation from diagnosis reports for accurate and generalisable head disorder detection. To our knowledge, this system is the first to simultaneously require no expert-annotated CT scans for model training, cover four types of head disorder at the same time, achieve accurate performance for multiple disorder types in an intra-centre test, generalise well for different centres, different CT equipment, and different countries, and improve the performance of radiologists in clinical practice. This study also proposed a novel deep learning algorithm with the following unique features compared with existing AI and deep learning models for head disorder detection from CT scans. Our system used keyword matching on the textual diagnosis reports to generate disorder labels for each CT scan, leading to no required expert efforts. Therefore, it requires less effort to construct a large dataset for model training, leading to accurate and generalisable performance. We also proposed RoLo, a novel weakly supervised learning algorithm, with a noise-tolerant mechanism for robust learning and a multi-instance learning strategy with an attention module to localise lesions, even without accurate and exact labels. Finally, the learning framework is task-agnostic, such that it enables the flexibility to involve new types of disorders, and the principles of this system could be applied to building computer-aided diagnosis systems for a range of other diseases.

Implications of all the available evidence

Head disorders, such as brain ischaemia, haemorrhage, tumours, and skull fractures, greatly affect the structure and function of the head and brain, causing high morbidity and mortality. Our system could be a useful aid for radiologists to diagnose multiple head disorder types efficiently and accurately. The generalisability of our system ensured that it had stable performance in different institutions and different countries, indicating potentially broad deployment of our system worldwide.

requires considerable efforts and medical expertise. Investigating new algorithms to build accurate and generalisable deep-learning systems in an efficient way is highly necessary and will facilitate the development and application of deep learning in medical practice.

In this study, we present an annotator-free deep-learning system for detecting multiple head disorders, including ischaemia, haemorrhage, tumours, and skull fracture, with the advantage of accuracy, generalisability, efficiency, and explainability (appendix p 1). The system was built in an annotator-free manner, and tested on different datasets from multiple independent centres. The system provided slice and region-level localisation of lesions for explaining the decision. We developed computer-aided diagnosis software, with which the diagnosis performance of radiologists was improved.

See Online for appendix

Methods

Overview of annotator-free deep-learning system

The annotator-free deep-leaning system was based on a large but coarse training dataset retrieved from the picture archiving and communication systems of the Chinese PLA General Hospital. We collected a dataset with 104 597 CT scans and labelled scans by keyword-based retrieving and matching from a picture archiving and communication system. The labels are inaccurate because of the annotator-free keyword matching, and inexact because of scan-level annotation. To address these issues, we proposed RoLo, a novel weakly supervised learning method, with a noise-tolerant mechanism for robust learning from inaccurate labels and a multi-instance learning[5] strategy with an attention module to localise lesions from inexact labels (appendix p 1). Scan-level labels were assigned by an automatic keyword matching strategy. The system was trained in a weakly supervised manner to address the inaccuracy and inexactness of the scan-level annotations. The system generates both disorder types and lesion localisation for decision understanding.

Data collection

We retrieved 630 992 CT scans from the picture archiving and communication systems of the Chinese PLA General Hospital (Beijing, China), a leading national hospital serving patients all over China, with the inspection time (when the scan was taken) from March 1, 2012, to Sept 30, 2018. All scans were stored in a digital imaging and communications in medicine (DICOM) format. To collect positive and negative samples of the four disorder types (ischaemia, haemorrhage, tumours, and skull fracture), we retrieved 121 576 electronic diagnosis reports by keyword retrieval (appendix p 3). After matching CT scans and diagnosis reports by patient identification number and inspection time, and excluding the non-axial section scans, the non-head scans, and all reconstructed scans (appendix p 4), 70 898 reports were matched to 107 754 CT scans post-filtering, with 21 067 haemorrhage scans, 25 446 ischaemia scans, 19 804 skull fracture scans, 3634 tumour scans, and 46 646 other scans (normal head scans). These scans were randomly divided into the training dataset (104 597 scans), the validation dataset (800 scans; appendix p 5), and the retrospective test dataset (2357 scans; 41% with disorders), with no patients shared between datasets. The retrospective test dataset was further divided into ten groups based on the maximum lesion size in each scan to evaluate the performance of our system on different lesion sizes. We also used this dataset to analyse some unconfident and incorrect predictions of the system. Based on the retrieval keywords and diagnosis report, each scan was automatically labelled by disorder type. We did not involve any expert annotators to label any CT scan in the training and validation datasets, but we used labels extracted from the historical free-text reports as the annotation of each scan.

Between Oct 1, 2018, and July 31, 2019, we further prospectively collected 17 937 scans from the Chinese PLA General Hospital to construct our prospective test dataset. Between April 1, 2018, and May 31, 2019, we collected 29 805 scans from the Brain Hospital of Hunan Province (Hunan, China). We used the same filtering strategy as for the retrospective test dataset to obtain valid CT scans, resulting in 650 scans (45% with disorders) in the prospective test dataset, of which 200 were randomly selected for the reader study for four radiologists (with 5, 10, 10, and 11 years experience, respectively) from the Chinese PLA General Hospital, and 1525 scans (35% with disorders) in the cross-centre test dataset (appendix p 1). We also used CQ500,[6] a public dataset from India with 491 scans taken between Jan 1, 2012, and Feb 1, 2018, to do a cross-nation test.

The CT scans from the Chinese PLA General Hospital were from four different manufacturers: Siemens (Munich, Germany), GE Healthcare (Chicago, IL, USA), United Imaging Healthcare (UIH; Shanghai, China), and Philips (Amsterdam, Netherlands). We selected 64 025 CT scans from Siemens and GE Healthcare in the training dataset as the cross-equipment training dataset, 440 CT scans from Siemens and GE Healthcare in the validation dataset as the cross-equipment validation dataset, and 1484 CT scans from UIH and Philips in the retrospective test dataset as the cross-equipment test dataset (appendix pp 1, 6). We used these datasets to train and evaluate the cross-equipment generalisability of the system.

This study was approved by the Chinese PLA General Hospital. Informed consent was waived for retrospectively collected CT scans, which were anonymised. Written informed consent was obtained from patients whose CT scans were prospectively collected.

Dataset labelling

To precisely evaluate our system, 20 radiologists (five of whom were senior radiologists) annotated the ground truth of retrospective, prospective, and cross-centre test datasets at the lesion level, where the lesions were marked by rectangular bounding boxes on each slice (appendix p 7). Each scan was labelled by at least one attending radiologist and one senior radiologist. We used the information from clinical diagnosis and surgery as supplementary information for radiologist annotations to ensure the correctness of the annotations (appendix p 8). During construction of the test datasets, we considered three aspects, as follows: the numbers of patients being male or female are similar; the numbers of patients are similar for three age groups (<44 years, 44–59 years, and >59 years); and the normal CT scans cover 55–65% of the whole dataset. These distributions are consistent with real-world distributions in the Chinese PLA General Hospital and the Brain Hospital of Hunan Province (appendix p 9).

CT slice conversion

We transferred a CT slice to a three-channel 8-bit image from its original format, following the common image format that is suitable for display. Radiologists typically use different window locations and window widths to identify different disorder types.[7,8] Inspired by this, we applied different settings of window locations and window widths for three channels (appendix p 11).

Model learning

One CT scan might contain multiple disorder types, and thus we formulated head disorder detection as a multi-class, multi-label classification problem.[9] A CT scan is judged against each disorder type. Specifically, we adopted the one-versus-all strategy.[10] For each disorder type, we built a classification network containing a convolutional neural network-based backbone g (ResNet–18[11]) and a classifier f (a two-layered fully-connected network). Given a slice x, g generates a feature h=g(x). Then h is fed into f and a single confidence value f(h) in [0,1] is obtained, indicating the probability that a slice contains the corresponding disorder type.

As the scan-level labels were inaccurate and inexact, we proposed RoLo, a weakly supervised strategy. We used a gated attention mechanism[5] to identify abnormal slices in the scan in a multi-instance learning manner. Given the hidden feature h* for a scan and its corresponding label $y\varepsilon(0,1)$, we proposed a novel noise-tolerant loss function:

$$l_{scan}(\mathbf{h}^*,y)=\beta^n \cdot (1-y) \cdot l_{cross}\,(\mathbf{h}^*,y) + \beta^p \cdot y \cdot l_{trunc}\,(\mathbf{h}^*,y)$$

This loss function used β^p and β^n to address the class imbalance by class-balanced loss.[12] For negative scans, the cross-entropy loss is:

$$l_{cross}(\mathbf{h}^*,y=0)=-\log(1-f(\mathbf{h}^*))$$

For positive scans, we used the truncated loss[13]:

$$l_{trunc}(\mathbf{h}^*,y=1)=\begin{cases}(1-f(\mathbf{h}^*)^q)/q & \text{if } f(\mathbf{h}^*)\leq \mathrm{t}\\(1-t^q)/q & \text{if } f(\mathbf{h}^*)>\mathrm{t}\text{'}\end{cases}$$

where q and t are hyper-parameters. This noise-tolerant loss downweights wrongly labelled samples. In this way, our model can learn correct information from inaccurate labels, and learn to localise lesions from inexact scan-level annotations (appendix p 13).

Disorder visualisation

An explainable model should be able to provide understandable clues to support its prediction.[14,15] Our system can identify key regions that might contain lesions by segmentation maps, which can be viewed by radiologists to understand the predictions. We employed the Grad-CAM technique[16] and an attention module to develop the segmentation maps to identify these regions (appendix p 15).

System development

To develop the system, we used the training dataset from the Chinese PLA General Hospital, which was built in an annotator-free manner. We changed the size of the training dataset by random sampling, from 10% to the whole dataset, to evaluate the performance of our system with different training dataset sizes. The non-expert automatic annotation contains wrong labels (inaccurate), which provides wrong supervision to mislead the model, and the scan-level annotation does not specify the location of lesions in a three-dimensional scan that contains many slices (inexact), such that it is difficult to identify key features from the small lesions in a large CT scan, making if difficult for conventional supervised training to reach high accuracy.[17] Based on the comparison between the automatically generated labels and expert annotated labels on the test data, the simple annotation strategy leads to 329 (14%) of 2357 wrong labels in datasets (appendix p 16).

The system for head disorder detection is shown in the appendix (p 1). Automatic scan annotation created a large-scale labelled training dataset in an annotator-free way. We used RoLo (appendix p 1) to train a sub-network for each disorder type from inaccurate and inexact annotations, providing accurate predictions and disorder visualisation for decision understanding. Training details are in the appendix (p 17).

Model pretraining

A pretrained deep model is important for deep learning. Many previous works used ImageNet[3] pretrained models. However, ImageNet images are different from medical images. We used momentum contrast loss[18] to pretrain the model with our training dataset in an unsupervised way.

Model selection and statistical analysis

The model was dynamically updated during training. We selected the optimal model by using the validation dataset. Specifically, for each type of disorder, the model with the largest area under the curve (AUC; computed on the validation dataset) was selected. To select a cutoff threshold for each model, we used the validation dataset to compute the sensitivity and specificity under different thresholds (appendix p 18).

All statistical analyses were done using Python (version 3.6.9). We used the sklearn package (version 0.20.3) to analyse receiver operating characteristic (ROC) curves and compute AUCs, sensitivity, specificity, positive predictive value, and negative predictive value (appendix p 19). For each test task, a ROC curve was created by plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) by varying the predicted probability threshold. Subsequently, the AUC values were calculated accordingly. For statistical significance, we applied 95% CIs calculated with the scipy package (version 1.3.2).

Ablation study

In our system, multi-instance learning, noise-tolerant loss, and network pretraining by medical images are the main technical innovations. To quantify their effect, we did an ablation study on the retrospective dataset by replacing multi-instance learning with simple two-dimensional or three-dimensional deep networks, noise-tolerant loss by cross-entropy loss or label-smoothing loss,[19] and medical images by ImageNet for pretraining.

Role of the funding source

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

Results

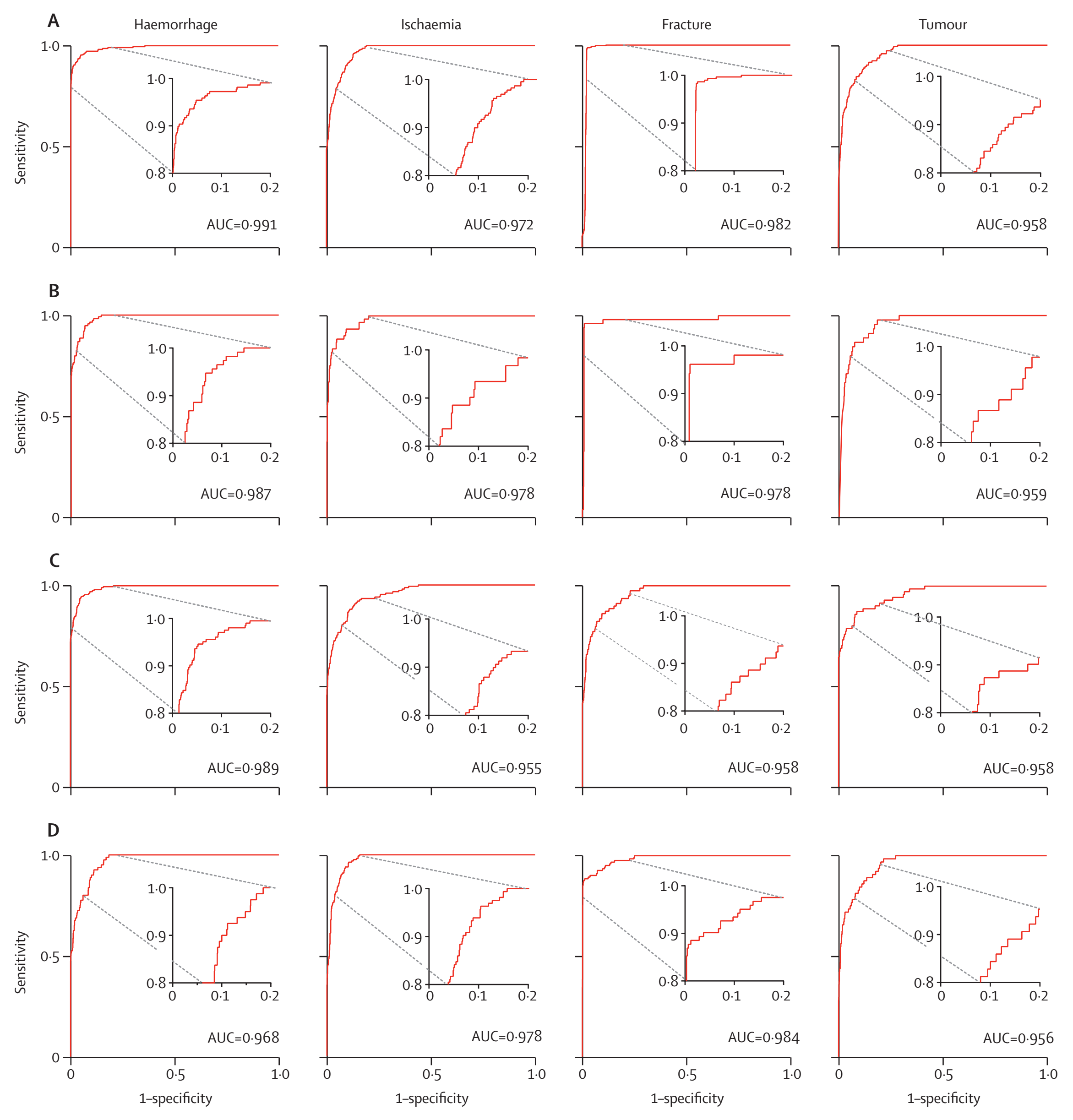
In this study, we first used the retrospective dataset and the prospective dataset to evaluate our system. Disorder detection was evaluated at the scan level.[6]

The ROC curves for detecting the four disorder types on the two test datasets are shown in figure 1A and 1B. The mean AUC for the retrospective dataset was 0·976 (95% CI 0·976–0·976), and the mean AUC for the prospective dataset was 0·975 (0·974–0·976; table). The table shows the detailed results for the four disorder types (appendix p 20).

We also compared our system with four independent radiologists at the Chinese PLA General Hospital for 200 randomly selected scans from the prospective test dataset (figure 1E). The independent radiologists were masked to both the clinical and the model-predicted labels. The working experiences of the radiologists were 5, 10, 10, and 11 years, respectively. Our system achieved superior performance to the attending radiologist (5 years) and was similar to the senior radiologists (10, 10, and 11 years, respectively; figure 1E).

We investigated the performance of our system in handling lesions of different sizes. We divided the retrospective test sample into ten groups based on the maximum lesion size in each scan and then tested the system on each group. Our system yielded a stable performance for various lesion sizes (figure 1F). For the smallest 10% of test data lesions (about 7-mm diameter), our system showed an AUC of 0·957 (95% CI 0·953–0·960).

In the ablation study (appendix p 21), the AUCs were 0·904 (95% CI 0·903–0·905) when multi-instance learning was replaced with a two-dimensional network, 0·898 (0·897–0·899) when multi-instance learning was replaced with a three-dimensional network, 0·872



(Figure 1 continues on next page)

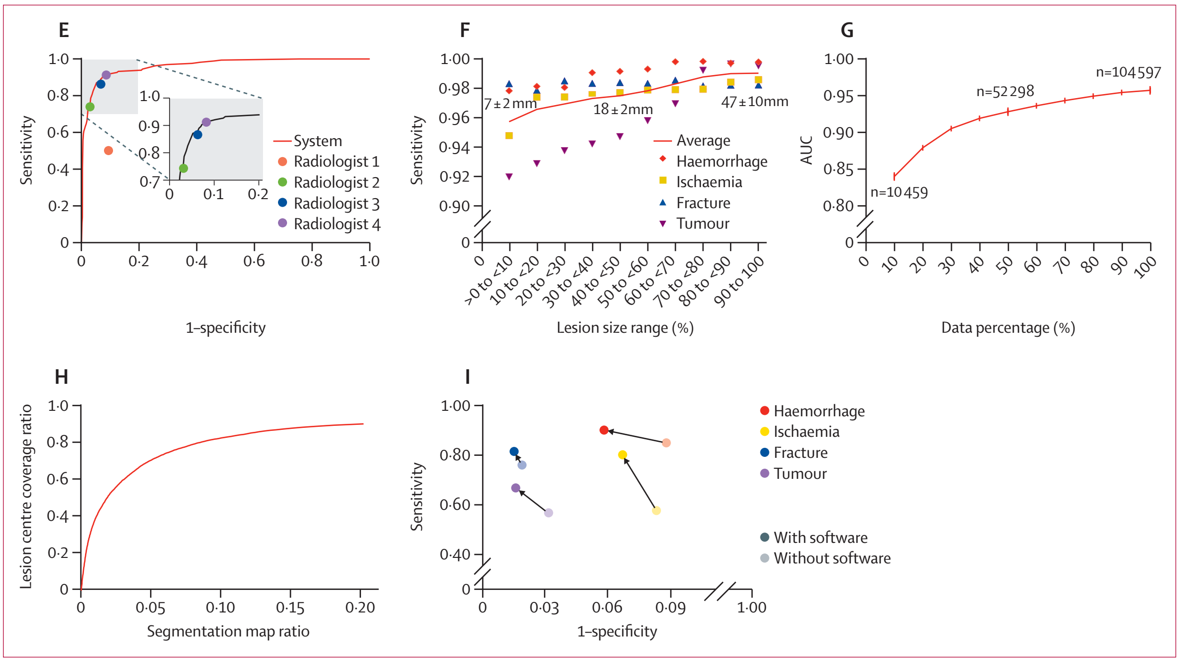


Figure 1: **Performance and comparative analyses of the deep-learning system**

(A–D) Receiver operating characteristic curves on the retrospective (2357 scans), prospective (650 scans), cross-centre (1525 scans), and cross-equipment (1484 scans) test datasets. (E) Performance comparison between our system and four radiologists on a subset of the prospective test dataset. The radiologists have 5 (orange), 10 (green), 10 (blue), and 11 (purple) years of working experience. (F) Performance with respect to different lesion size on the retrospective test dataset. Based on the ranking of the bounding box size of each lesion, we divided the retrospective test dataset into ten groups. The mean diameters of some groups are shown. (G) Performance on the CQ500 test dataset for detecting haemorrhage with respect to the size of training dataset. The number of scans used for training is denoted as n. (H) Evaluation of lesion localisation. The x-axis represents the mean area ratio of segmentation maps in the whole CT scans, and the y-axis is the ratio of lesion centre points covered by the maps. (I) The performance of four radiologists before and after considering the system recommendation. The mean improvement of these four radiologists on the detection of haemorrhage, ischaemia, fracture, and tumours is shown. AUC=area under the curve.

(0·869–0·875) when noise-tolerant loss was replaced with cross-entropy loss, 0·908 (0·906–0·909) when noise-tolerant loss was replaced with label-smoothing loss, and 0·948 (0·947–9·949) when medical images were replaced by ImageNet for pretraining.

Besides accuracy, a practical deep-learning model should generalise to new data from different centres and equipment. We used the cross-centre test dataset from the Brain Hospital of Hunan Province (figure 1C; table). The mean AUC dropped 0·011 from the retrospective intra-centre test set to the cross-centre test set, showing the generalisability of the system for different centres. We further tested our system on the CQ500 dataset from India[6] (table). For haemorrhage and fracture, the AUCs were 0·957 (95% CI 0·957–0·959) and 0·971 (0·970–0·973), respectively. For cross-equipment evaluation (Figure 1D; table), the mean AUC was 0·971 (95% CI 0·971–0·972).

The size of training dataset is important for generalisability. We evaluated the effect of the training dataset size. We randomly sampled a subset from the training dataset to develop our system and tested it on the CQ500 (figure 1G). The AUC gradually increased with more training data, showing the importance of a large training dataset for generalisability.

Our system could generate a segmentation map to highlight the possible lesion region in a scan, which showed the decision clues that could be used by radiologists (figure 2A). We quantitatively evaluated the segmantaion maps by calculating the localisation accuracy[7] (appendix p 26). Even though the mean highlighted area was only 9% of the total head area, 81% of all the lesion centres were correctly covered (figure 1H).

We used average drop (lower is better) and increase in confidence (higher is better) for further evaluation[20] (appendix p 27). Our system showed a 0·94% average drop and 49·77% increase in confidence. These results show that our system can localise the key regions and features of lesions for detecting disorders.

Our system can generate a confidence value for each prediction. We collected the test samples with unconfident results and discussed the results with five expert radiologists (with 8, 10, 12, 13, and 15 years of experience, respectively) from the Chinese PLA General Hospital. Some observations warrant further comment. First, some cases were confusing even for experts, and might need additional information to distinguish the disorder type(s). For example, the cerebral falx normally shows hyperdensity, which is occasionally very similar to subarachnoid haemorrhage on CT scans. Suspected haemorrhage identified by our model in these particular regions could help alert radiologists (figure 2B). Second, sometimes multiple entities overlay. For example,

	Area under the curve	Positive predictive value	Negative predictive value	Sensitivity	Specificity
Retrospective					
Haemorrhage	0·991 (0·990–0·991)	0·954 (0·953–0·955)	0·935 (0·934–0·937)	0·970 (0·969–0·971)	0·935 (0·934–0·937)
Ischaemia	0·972 (0·972–0·973)	0·949 (0·947–0·950)	0·877 (0·876–0·879)	0·963 (0·962–0·965)	0·877 (0·876–0·879)
Fracture	0·982 (0·982–0·983)	0·979 (0·978–0·979)	0·971 (0·970–0·971)	0·990 (0·990–0·991)	0·971 (0·970–0·971)
Tumour	0·958 (0·958–0·959)	0·910 (0·908–0·911)	0·852 (0·849–0·854)	0·934 (0·932–0·935)	0·852 (0·849–0·854)
Mean	0·976 (0·976–0·976)	0·948 (0·947–0·949)	0·909 (0·907–0·910)	0·964 (0·963–0·965)	0·909 (0·907–0·910)
Prospective					
Haemorrhage	0·987 (0·987–0·987)	0·950 (0·949–0·951)	0·915 (0·913–0·917)	0·979 (0·978–0·981)	0·915 (0·913–0·917)
Ischaemia	0·978 (0·978–0·979)	0·918 (0·916–0·919)	0·883 (0·880–0·886)	0·973 (0·971–0·975)	0·883 (0·880–0·886)
Fracture	0·978 (0·976–0·979)	0·913 (0·911–0·915)	0·966 (0·963–0·968)	0·979 (0·978–0·981)	0·966 (0·963–0·968)
Tumour	0·959 (0·957–0·959)	0·885 (0·882–0·888)	0·864 (0·861–0·866)	0·962 (0·959–0·965)	0·864 (0·861–0·866)
Mean	0·975 (0·974–0·976)	0·916 (0·914–0·918)	0·907 (0·904–0·909)	0·973 (0·971–0·975)	0·907 (0·904–0·909)
Cross-centre					
Haemorrhage	0·989 (0·989–0·989)	0·949 (0·949–0·950)	0·930 (0·928–0·931)	0·966 (0·965–0·967)	0·930 (0·928–0·931)
Ischaemia	0·955 (0·954–0·956)	0·908 (0·906–0·909)	0·849 (0·846–0·853)	0·930 (0·929–0·932)	0·849 (0·846–0·853)
Fracture	0·958 (0·957–0·958)	0·903 (0·901–0·905)	0·835 (0·831–0·838)	0·947 (0·945–0·948)	0·835 (0·831–0·838)
Tumour	0·958 (0·958–0·960)	0·890 (0·888–0·892)	0·817 (0·813–0·822)	0·937 (0·935–0·939)	0·817 (0·813–0·822)
Mean	0·965 (0·964–0·966)	0·912 (0·911–0·914)	0·858 (0·855–0·861)	0·945 (0·944–0·947)	0·858 (0·855–0·861)
CQ500					
Haemorrhage	0·957 (0·957–0·959)	0·925 (0·923–0·926)	0·856 (0·854–0·858)	0·942 (0·940–0·943)	0·856 (0·854–0·858)
Fracture	0·971 (0·970–0·973)	0·876 (0·873–0·879)	0·920 (0·918– 0·923)	0·959 (0·956–0·962)	0·920 (0·918–0·923)
Mean	0·964 (0·964–0·966)	0·900 (0·898–0·903)	0·888 (0·886–0·891)	0·950 (0·948–0·953)	0·888 (0·886–0·891)
Cross-equipment					
Haemorrhage	0·968 (0·967–0·968)	0·926 (0·924–0·928)	0·864 (0·862–0·866)	0·969 (0·967–0·971)	0·864 (0·862–0·866)
Ischaemia	0·978 (0·978–0·978)	0·951 (0·950–0·952)	0·903 (0·901–0·904)	0·971 (0·970–0·973)	0·903 (0·901–0·904)
Fracture	0·984 (0·983–0·984)	0·932 (0·930–0·933)	0·895 (0·893–0·897)	0·959 (0·958–0·961)	0·895 (0·893–0·897)
Tumour	0·956 (0·954–0·956)	0·918 (0·917–0·920)	0·815 (0·811–0·819)	0·974 (0·972–0·975)	0·815 (0·811–0·819)
Mean	0·971 (0·971–0·972)	0·932 (0·930–0·933)	0·869 (0·867–0·872)	0·968 (0·967–0·970)	0·869 (0·867–0·872)

Data in parentheses are 95% CIs.

***Table:* Performance of our system on the retrospective (2357 scans), prospective (650 scans), cross-centre (1525 scans), CQ500 (491 scans), and cross-equipment (1484 scans) test datasets for detecting multiple head disorders**

tumour apoplexy refers to haemorrhage secondary to tumour growth. In such cases, tumours are labelled but haemorrhage is not (figure 2B). Third, some output labels should be regarded as false positives for the four disorder types, and indicate disorder types that were not considered in this paper. For example, a hyperdense middle cerebral artery sign indicates fresh occlusive thrombus located in the middle cerebral artery. Despite being misrecognised as haemorrhage by our model, this imaging feature is indicative for early territorial infarction and is important for making a rapid and appropriate treatment decision (figure 2B).

In some cases, the system yielded a large probability (very confident) for wrong disorder types, for example, misclassification of calcification as haemorrhage. Sometimes, calcification with low attenuation on CT occurs, and discrimination from calcification and acute haemorrhage is challenging for even expert radiologists (figure 2C).

We developed computer-aided diagnosis software for clinical use (appendix p 28). The software takes a CT scan as an input, and outputs the probability for each type of disorder and the possible lesion regions. To quantitatively evaluate whether this software could improve the diagnostic performance of radiologists, we evaluated the decisions of four radiologists immediately before and after viewing the information from the computer-aided diagnosis system (figure 1I). 200 samples randomly selected from the prospective test dataset were chosen. These four radiologists are the same as the four in the experiment shown in figure 1E. After viewing the prediction of the system, 24·5% of decisions across all four radiologists were changed, and the sensitivity and specificity of radiologists were improved by a mean of 0·109 (95% CI 0·086–0·131) and 0·022 (0·017–0·026), respectively.

Discussion

Annotation is a key component in machine learning and deep learning. However, annotating medical data requires expert knowledge and is time consuming. Region-level annotation, indicating the exact region of lesions,

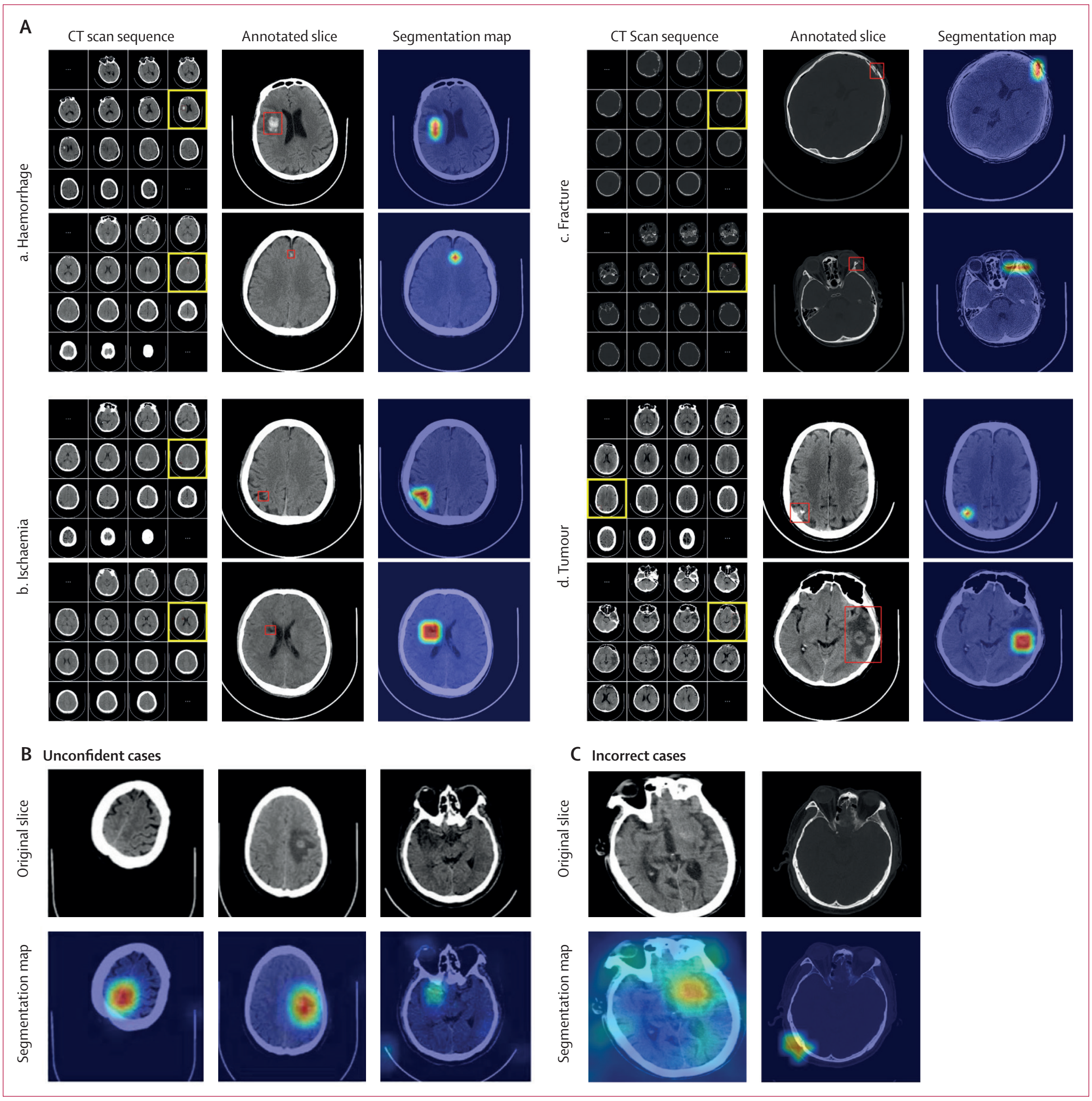


Figure 2: **Lesion visualisation analysis**

(A) Examples of our system for lesion detection. The CT scan sequence is the sequence of slices in a CT scan that contains the corresponding head disorder. The yellow boxes indicate the slices containing lesions. Annotated slice shows annotations by the expert radiologists, shown by red rectangles. Segmentation maps show the probability that each pixel belongs to a lesion; a warmer colour indicates a higher probability of lesions. The segmentation maps provide visual clues of the system decision. (B) Unconfident cases analysis. Our system (left) reported cerebral falx as haemorrhage, (middle) detected haemorrhage within the brain tumour, and (right) treated a hyperdense middlecerebral artery sign as a haemorrhage. (C) Incorrect cases analysis. Our system (left) reported calcification as haemorrhage and (right) detected the widening of the cranial sutures as bone fracture.

provides strong supervision to the system such that it has sufficient and reliable information to learn from. However, it can take several minutes for an expert to annotate one scan; therefore, much research[21] with region-level annotations might only have small numbers of labelled scans. A small training dataset limits the system generalisability such that performance drops in cross-centre tests. Slice-level annotation identifies whether each slice contains lesions or not. With slice-level annotation, a two-dimensional network can be trained by each annotated slice. In this case, the supervision is also strong. However, annotating each slice is, again, time consuming. Scan-level annotation gives a label for the whole scan, not specifying the locations of lesions, with a low annotation time-cost. However, the challenge is that its supervision is weak. Furthermore, without using an elaborate algorithm design, it is difficult to achieve high accuracy. Some previous research used weakly supervised learning[22,23] to reduce annotation effort. However, such research typically focused on single-slice images instead of multi-slice CT scans, and annotations were perfect, making the problem easier. Some previous research considered the label noise.[24] However, such research typically focused on real-life images, such a photographs, in which the key object was large, and assumed that it was easy to collect a reliable training dataset as a reference to correct incorrect labels, which is difficult to satisfy in medical applications. This study aimed to complete a more challenging task, where the scan-level annotation contains mistakes. With the proposed RoLo architecture, our system showed more stable performance than systems based on region-level or slice-level annotations. To our knowledge, this has not been shown in previous works (appendix p 29).

The proposed system generalises well across different domains, shown by its small performance drop in cross-centre, cross-equipment, and cross-nation experiments. For haemorrhage detection, the AUC of the proposed system dropped 0·004 from the retrospective dataset to the prospective dataset, 0·002 to the cross-centre dataset, 0·023 to the cross-equipment dataset, and 0·034 to the cross-nation dataset (CQ500). The system showed stable performance on cross-domain datasets. The cross-nation test AUC was higher than the intra-nation test AUC reported in the original paper.[6] Although our system was developed based on data from China with no expert annotation, we found higher AUCs than those reported in the original literature (0·942 and 0·964 and p=0·0010 and p=0·049, respectively,[6] which used expert annotated intra-nation training data from India (appendix p 25). Computer-aided diagnostic systems are in demand in low-income and middle-income regions, which have a shortage of experts, such that cross-centre and cross-equipment performances are more important than intra-domain performance in practice. Our annotator-free system benefits from the large-scale dataset and avoids the heavy annotation burden, making it easy to develop and deploy in practice. For cross-equipment evaluation the mean AUC was only 0·005 lower than the non-cross-equipment setting. These results show that our system generalises well across different centres and equipment, making it practical for clinical use.

Our system can provide visual clues to localise lesions and support decisions. This feature is important in practice as visual clues explain system predictions, which is especially important in medical applications. Previous studies have also adopted visualisation methods, such as attention maps[8,23] and Grad-CAM.[10] Our visualisation method combines an attention module and Grad-CAM with MinMaxScaler,[25] making the presentation of the lesion regions more obvious. Since identification of haemorrhage, which usually suggests an acute condition and is related to clinical symptoms, is critical for timely clinical diagnosis and treatment, the increased sensitivity of our model would be helpful to decrease false negatives. An additional issue is misidentification of the widening of the cranial sutures as a bone fracture. In the presence of secondary signs, including subdural or epidural haematoma or soft tissue injury in head trauma, the incidence of this type of mistake would increase. Our results show the advantages of clinical use of the proposed system, particularly to improve the diagnostic performance of radiologists.

Alongside the good performance for multiple head disorder detection, our learning framework and algorithm are task agnostic. Therefore, it is easy to extend our system to include more head disorder types, without involving any expert labelling effort. Additionally, the learning framework and algorithm can be applied to many other situations with multiple disorder types, such as chest CT.

Our study has some limitations. Our system required a large-scale CT dataset (over 100 000 CT scans post filtering in this study) with clinical reports, which can be difficult to collect even at a large hospital. Here, we should note that a large-scale training dataset is essential for high accuracy and generalisability, especially in practical applications. We are currently developing our system in a federated learning framework,[26] which shares data in a secure way among multiple hospitals. In this way, it should be easy to build a large training dataset by combining data from many hospitals.

In this study we used only non-contrast CT. Some types of disorders, for example hyperacute ischemia, are barely detectable on non-contrast CT. We plan to extend our system to include more modalities, including contrast-enhanced CT and MRI. Another limitation is the large hardware consumption required for training. Since our algorithm uses all slices in each scan to compute the loss function, it takes more than 10GB video random access memory (VRAM) for training. Thus, a powerful graphics processing unit with large VRAM is required. We aim to optimise the training strategy and make the system more hardware efficient. Additionally, this system is supportive for qualitative diagnosis, but it is difficult to yield a precise quantitative evaluation, for example, the volume of haematoma or ischaemia. This limitation could be addressed by extending the weakly supervised classification algorithm into a segmentation task.

In conclusion, the proposed system achieved accurate and generalisable performance for multiple head disorder detection, without involving expert annotated data during system development. The system improved the diagnostic performance of radiologists. Because of its accuracy, generalisability, and explainability, the system could be used in clinical practice to improve the accuracy and efficiency of radiologists in different hospitals.

Contributors

FX, XL, QD, and YG designed the research. XL, JL, DY, JH, DH, and WZ collected the data. YG, YH, ZZ, JL, FX, and XL verified the raw underlying data, which was accessed by all authors. YH, ZZ, LM, HT, and CC developed the computer-aided diagnosis software. YH, YG, GD, JL, DH, ZZ, and LM analysed the results. YG, YH, and JL co-wrote the manuscript. FX, SL, HQ, and YG critically revised the manuscript, and all authors

discussed the results and provided feedback regarding the manuscript. All authors had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The system was developed using standard libraries and scripts available in PyTorch. The full code, including the training code, test code, and the computer-aided detection software, are available at http://xufeng.site/AutoHeadCAD. The datasets to develop our system are not publicly available because of the privacy requirement of the Chinese PLA General Hospital (Beijing, China). The anonymised cross-centre dataset (1525 scans and expert annotations) from the Brain Hospital of Hunan Province (Hunan, China) is available for research purposes from the corresponding authors upon reasonable request.
The evaluated CQ500 dataset can be downloaded from http://headctstudy.qure.ai/dataset.

Acknowledgments

We thank the radiologists from the Chinese PLA General Hospital (Beijing, China) for their efforts in labelling the test data. We thank National Key R&D Program of China (2020AAA0105500 to YG, GD, HQ, QD, and SL, 2018YFA0704000 to FX), National Natural Science Foundation of China (61822111 to FX, 81825012 to XL, 81730048 to XL, 61727808 to FX, 62021002 to FX, 61971260 to YG, GD, HQ, QD, and SL, U21B2013 to YG and HQ, 81901708 to JL), and Beijing Natural Science Foundation (JQ19015 to FX) for supporting this study.

References

1 Feigin VL, Abajobir AA, Abate KH, et al. Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Neurol* 2017; **16:** 877–97.

2 Larson DB, Johnson LW, Schnell BM, Salisbury SR, Forman HP. National trends in CT use in the emergency department: 1995–2007. *Radiology* 2011; **258:** 164–73.

3 Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009.

4 Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* 2017; published online Dec 25. https://doi.org/10.48550/arXiv.1711.05225 (preprint).

5 Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. *arXiv* 2018; published online June 28. https://doi.org/10.48550/arXiv.1802.04712 (preprint).

6 Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 2018; **392:** 2388–96.

7 Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng* 2019; **3:** 173–82.

8 Brant WE, Helms CA. Fundamentals of diagnostic radiology. Philadelphia: Lippincott Williams & Wilkins, 2012.

9 Zhou ZH, Zhang ML, Huang SJ, Li YF. Multi-instance multi-label learning. *Artif Intell* 2012; **176:** 2291–320.

10 Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008; **9:** 1871–74.

11 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27–30, 2016.

12 Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 15–20, 2019.

13 Zhang Z, Sabuncu MR. Generalized cross entropy loss for training deep neural networks with noisy labels. 32nd Conference on Neural Information Processing Systems; Dec 3–8, 2018.

14 Muelly MC, Peng L. Spotting brain bleeding after sparse training. *Nat Biomed Eng* 2019; **3:** 161–62.

15 Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2:** 749–60.

16 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision ICCV 2017. p 618–26.

17 Ding G, Guo Y, Chen K, Chu C, Han J, Dai Q. DECODE: deep confidence network for robust image classification. *IEEE Trans Image Process* 2019; **28:** 3752–65.

18 He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. *arXiv* 2019; published online Nov 13. https://doi.org/10.48550/arxiv.1911.05722 (preprint).

19 Müller R, Kornblith S, Hinton GE. When does label smoothing help? *Adv Neural Inf Process Syst* 2019; **32:** 4694–703.

20 Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. 2018 IEE Winter Conference on Applications of Computer Vision; March 12–14, 2018.

21 Monteiro M, Newcombe VFJ, Mathieu F, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. *Lancet Digit Health* 2020; **2:** e314–22.

22 Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 14–19, 2020.

23 Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* 2021; **5:** 555–70.

24 Han B, Yao Q, Yu X, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. 32nd Conference on Neural Information Processing Systems; Dec 3–8, 2018.

25 Patro S, Sahu KK. Normalization: a preprocessing stage. *arXiv* 2015; published online March 15. https://doi.org/https://doi.org/10.17148/IARJSET.2015.2305 (preprint).

26 Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res* 2021; **5:** 1–19.