Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Examination-Level Supervision for Deep Learning–based Intracranial Hemorrhage Detection on Head CT Scans

Jacopo Teneggi, MSE • *Paul H. Yi, MD* • *Jeremias Sulam, PhD*

From the Department of Computer Science (J.T.), Department of Biomedical Engineering (J.S.), and Mathematical Institute for Data Science (MINDS) (J.S., J.T.), Johns Hopkins University, 3400 N Charles St, Clark Hall, Suite 320, Baltimore, MD 21218; and University of Maryland Medical Intelligent Imaging Center (UM2ii), Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, Md (P.H.Y.). Received May 8, 2023; revision requested July 6; revision received November 2; accepted December 5. **Address correspondence to** J.T. (email: *jtenegg1@jhu.edu*).

This research was supported by National Science Foundation CAREER Award Computing and Communication Foundations 2239787.

Conflicts of interest are listed at the end of this article.

See also commentary by Wahid and Fuentes in this issue.

Radiology: Artificial Intelligence 2024; 6(1):e230159 • https://doi.org/10.1148/ryai.230159 • Content codes: AI NR

Purpose: To compare the effectiveness of weak supervision (ie, with examination-level labels only) and strong supervision (ie, with image-level labels) in training deep learning models for detection of intracranial hemorrhage (ICH) on head CT scans.

Materials and Methods: In this retrospective study, an attention-based convolutional neural network was trained with either local (ie, image level) or global (ie, examination level) binary labels on the Radiological Society of North America (RSNA) 2019 Brain CT Hemorrhage Challenge dataset of 21 736 examinations (8876 [40.8%] ICH) and 752 422 images (107 784 [14.3%] ICH). The CQ500 (436 examinations; 212 [48.6%] ICH) and CT-ICH (75 examinations; 36 [48.0%] ICH) datasets were employed for external testing. Performance in detecting ICH was compared between weak (examination-level labels) and strong (image-level labels) learners as a function of the number of labels available during training.

Results: On examination-level binary classification, strong and weak learners did not have different area under the receiver operating characteristic curve values on the internal validation split (0.96 vs 0.96; $P = .64$) and the CQ500 dataset (0.90 vs 0.92; $P = .15$). Weak learners outperformed strong ones on the CT-ICH dataset (0.95 vs 0.92; $P = .03$). Weak learners had better section-level ICH detection performance when more than 10 000 labels were available for training (average $f_1 = 0.73$ vs 0.65; $P < .001$). Weakly supervised models trained on the entire RSNA dataset required 35 times fewer labels than equivalent strong learners.

Conclusion: Strongly supervised models did not achieve better performance than weakly supervised ones, which could reduce radiologist labor requirements for prospective dataset curation.

Supplemental material is available for this article.

Intracranial hemorrhage (ICH) is a potentially life-threatening condition, accounting for approximately 10%–20% of all strokes (1). Expert radiologists can diagnose ICH from unenhanced head CT scans by analyzing the location, shape, and size of the lesions (2). The large number of CT scans produced daily and the importance of quick diagnosis make automated diagnosis and triage of ICH using deep learning (DL) a compelling application of this technology (3–7).

Supervised learning is the most common approach to training DL models for detection of disease or injury, whereby a collection of images with ground truth labels is used for model training. This framework requires radiologists to manually annotate hundreds or thousands of images, which is a time-consuming and expensive process. This disadvantage was made painfully clear during the 2019 Radiological Society of North America (RSNA) Brain CT Hemorrhage Challenge, which required 60 volunteer expert radiologists and thousands of hours to annotate 21 736 head CT examinations for ICH (2). An alternative approach to collecting image-level labels (ie, for each section) is to use weak labels obtained at the examination-level (ie, aggregated over all the sections), such as those that can be extracted from radiology reports, a method known as weakly supervised learning. That is, a CT examination should be flagged as with hemorrhage regardless of which sections show ICH. Preliminary research efforts have demonstrated potential utility of weakly supervised learning using radiology report–derived annotations for whole-body PET/CT and body CT (8,9), suggesting that similar applications in neuroimaging may be successful.

For ICH detection on head CT scans, it remains unclear what kind of labels are optimal. While ground truth labels for each image or section of a CT scan may provide granular pixel-level annotations, they are time-consuming and expensive to obtain. On the other hand, weak examination-level labels are quickly and cheaply obtained through automated extraction from radiology reports using natural language processing but are coarser and less informative. Importantly, they cannot be used to train two-dimensional DL models that consider a single image at a time. At the same time, weakly supervised learning could reduce the labor required to curate large medical imaging datasets, providing a scalable solution to the primary bottleneck in development of DL models in radiology.

Abbreviations

AUC = area under the receiver operating characteristic curve, DL = deep learning, ICH = intracranial hemorrhage, MIL = multiple instance learning, RSNA = Radiological Society of North America

Summary

Supervised learning with image-level labels did not show superior performance to multiple instance learning with examination-level binary labels for intracranial hemorrhage detection on head CT scans.

Key Points

- The proposed attention-based convolutional neural network predicted the presence of intracranial hemorrhage on CT volumes of any number of sections without needing section- or pixel-level annotations.
- The models trained on the Radiological Society of North America (RSNA) 2019 Brain CT Hemorrhage Challenge dataset with examination-level binary labels achieved better generalization performance (area under the receiver operating characteristic curve = 0.95; $P = .03$) compared with models trained with section-level binary labels on the task of examination-level binary classification.
- Weakly supervised models (examination-level labels only) trained on the entire RSNA dataset required 35 times fewer labels than equivalent strongly supervised models (image-level labels).

Keywords

CT, Head/Neck, Brain/Brain Stem, Hemorrhage

The purpose of this study was to compare the performance of weakly supervised and strongly supervised DL models for detection of ICH on head CT scans. Specifically, we evaluated whether DL models trained using multiple instance learning (MIL)—an instance of weakly supervised learning in which inputs are considered bags of instances, and the label of the bag is a known function of the labels of its instances (10–13)—underperformed compared with models using standard, strong supervision.

Materials and Methods

Our retrospective study used public data only and was acknowledged as nonhuman subjects research by the University of Maryland Baltimore institutional review board. All code will be released.

Datasets

Models were trained and validated (80/20 data split) on the RSNA 2019 Brain CT Hemorrhage Challenge dataset (Table 1) (2). The dataset comprises 21 736 examinations from three institutions (Stanford University, Thomas Jefferson University, and Universidade Federal de São Paulo), totaling 752 422 images labeled by a panel of board-certified radiologists with the types of hemorrhage present (epidural, intraparenchymal, intraventricular, subarachnoid, subdural). No pixel-level annotations or demographics are available. Data splits were created by random sampling at the examination level to guarantee a fair comparison. Weak learners have access to approximately 35 times fewer labels (ie, the mean number of sections in an examination).

Two datasets were used for external testing (Table 2): the CQ500 dataset (14) (436 examinations; 212 ICH; mean age, 48 years ± 29 [SD]; 158 of 436 [36.2%] women) and the CT-ICH

Table 1: Number of Positive and Negative Labels in the Training and Validation Splits of the RSNA 2019 Brain CT Hemorrhage Challenge Dataset

	Training Split		Validation Split	
Label Level	Positive Labels	Negative Labels	Positive Labels	Negative Labels
Image level	86 295 of 601 930 (14.3)	515 635 of 601 930 (85.7)	21 489 of 150 492 (14.3)	129 003 of 150 492 (85.7)
Examination level	7100 of 17 388 (40.8)	10 288 of 17 388 (59.2)	1776 of 4348 (40.8)	2572 of 4348 (59.2)

Note.—The RSNA dataset does not provide demographic information. For each dataset, data are presented as number of labels, with percentage of total image-level or examination-level labels in parentheses. RSNA = Radiological Society of North America.

Table 2: Demographics, Number of Positive and Negative Examinations, Number of Positive and Negative Images, and Total Number of Images in the CQ500 and CT-ICH Datasets

Dataset	Age (y)	Female	Positive Examinations	Negative Examinations	Positive Images	Negative Images
CQ500	48 ± 29 (7–95)	158 of 436 (36.2)	212 of 436 (48.6)	224 of 436 (51.4)	NA	NA
CT-ICH	28 ± 20 (0–72)	33 of 75 (44.0)	36 of 75 (48.0)	39 of 75 (52.0)	318 of 2814 (11.3)	2496 of 2814 (88.7)

Note.—For each dataset, data are presented as mean ages in years ± SDs with the ranges in parentheses, numbers and percentages of female patients, numbers of labels with percentage of total number of examinations in parentheses, and total numbers of images. ICH = intracranial hemorrhage, NA = not applicable.

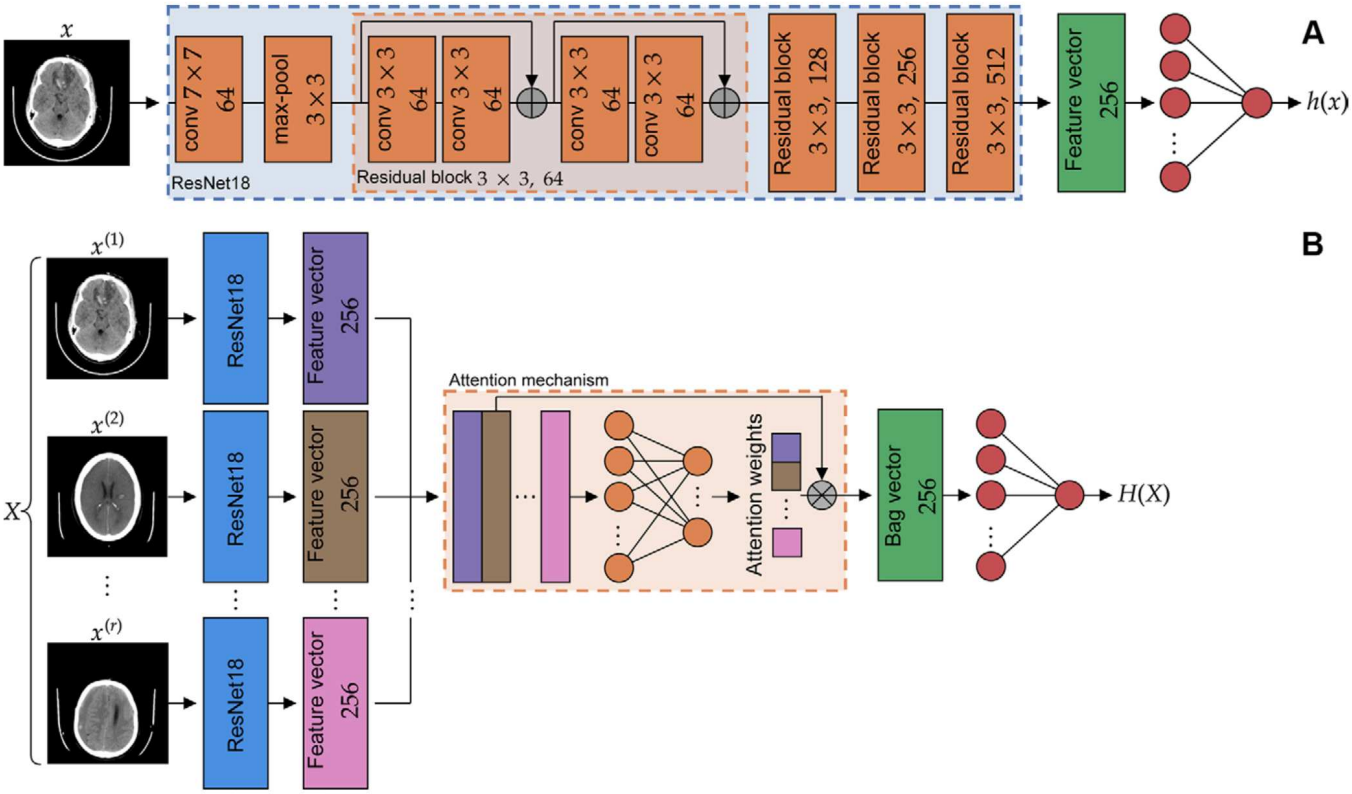

Figure 1: Schematic of model architectures. **(A)** The strong learner makes a prediction on every input image, *x*, and requires image-level labels for training. **(B)** The weak learner makes a prediction for an entire examination, $X = (x^{(1)},x^{(2)},...,x^{(r)})$, containing an arbitrary number of images, *r*. Thus, the weak learner requires only examination-level labels for training. Both models use the same convolutional neural network to encode images into feature representations and the same fully connected layer with sigmoid activation as output classifier. Conv = convolution, max = maximum.

dataset (15,16) (75 examinations; 36 ICH; mean age, 28 years ± 20; 33 of 75 [44.0%] women). The CQ500 dataset comprises scans from clinical centers in New Delhi, India, annotated with the types of hemorrhage present by three expert radiologists. For a subset of 196 scans, images were enhanced via the BHX dataset (17,16) with 6282 manual segmentations of bleeds performed by three other expert radiologists. The CT-ICH dataset was collected from Al Hilla Teaching Hospital, Iraq, from patients with traumatic brain injury with manual segmentations of bleeds performed by two expert radiologists. All images were windowed with the standard brain setting (window width = 80; window level = 40) and minimum and maximum normalized.

Model Architectures

Figure 1 depicts the strong and weak DL models. We considered examinations as bags of images (10–13) that were labeled as negative only if they did not contain any image with ICH. We used an attention-based MIL model (18) that can be trained with either image- or examination-level labels, thus enabling a controlled comparison. Details are provided in Appendix S1.

Training Procedure

Both learners are trained with focal loss (19) to account for label imbalances in the RSNA dataset and the gap in difficulty between predicting the presence or absence of ICH. Equations, data augmentation strategies, and hyperparameters are included in Appendix S1.

Statistical Analysis

Examination-level binary classification.— For the weak learner, the examination-level prediction is the output of the model. For the strong learner, the examination-level prediction is the maximum image-level prediction. Diagnostic performance was determined by each model's area under the receiver operating characteristic curve (AUC) and compared with the DeLong test.

Section-level hemorrhage detection.— Hemorrhage sequences are sets of consecutive images with ICH, and section-level hemorrhage detection is the task of retrieving the true hemorrhage sequences. This task is different from pixel-level binary classification and reflects how learners could be deployed in clinical settings to report actionable findings to radiologists. For strong learners, this is equivalent to selecting consecutive predicted positive images. For weak learners, two separate notions of image importance were used: *(a)* attention weights and *(b)* Shapley values (20,21). Positive images should receive larger attention weights because they contribute toward a positive prediction. This notion of importance is heuristic. In contrast, the Shapley value satisfies several desirable theoretical properties (20). h-Shap (22)—a hierarchical extension of the Shapley value—was modified to compute the attribution of every image in an examination. Images were selected by thresholding their importance. DL models were compared by means of

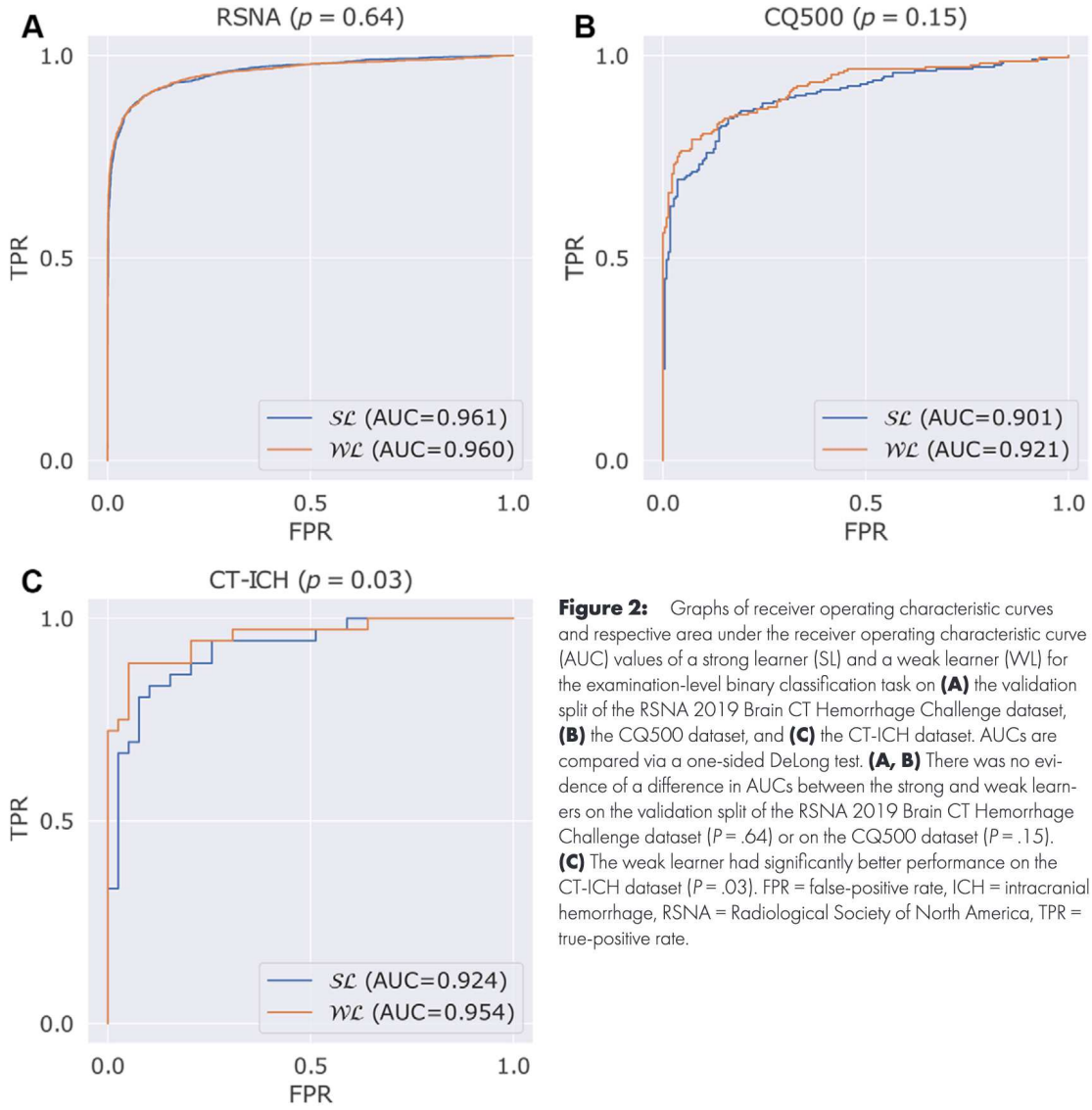

Figure 2: Graphs of receiver operating characteristic curves and respective area under the receiver operating characteristic curve (AUC) values of a strong learner (SL) and a weak learner (WL) for the examination-level binary classification task on **(A)** the validation split of the RSNA 2019 Brain CT Hemorrhage Challenge dataset, **(B)** the CQ500 dataset, and **(C)** the CT-ICH dataset. AUCs are compared via a one-sided DeLong test. **(A, B)** There was no evidence of a difference in AUCs between the strong and weak learners on the validation split of the RSNA 2019 Brain CT Hemorrhage Challenge dataset (P = .64) or on the CQ500 dataset (P = .15). **(C)** The weak learner had significantly better performance on the CT-ICH dataset (P = .03). FPR = false-positive rate, ICH = intracranial hemorrhage, RSNA = Radiological Society of North America, TPR = true-positive rate.

section-level f_1 score and recall on the validation split of the RSNA dataset (2413 true sequences [mean length = 9 images ± 7]) and the CT-ICH dataset (45 true sequences [mean length = 9 images ± 4]). The CQ500 dataset could not be used because it does not include image-level labels, and the BHX dataset does not provide segmentations for all positive scans. P values were computed via one-sided t tests. Details are provided in Appendix S1.

Pixel-level hemorrhage detection.— Pixel-level hemorrhage detection is the task of locating bleeds within positive images. We could not train segmentation models because the RSNA dataset does not provide pixel-level annotations. However, DL explainability methods can be used to compute saliency maps and augment classification models with detection ability (23,24). We compared two such methods: *(a)* Grad-CAM (25) and *(b)* h-Shap (22). The latter was extended with cycle spinning (26) to capture the complex shape of bleeds. Accuracy was computed via the Dice score between the saliency maps and the ground truth segmentations. Results were reported for true-positive images selected by both the strong learner and the weak learner with Shapley value thresholding; 885 images were compared from the CQ500 dataset and 119 images from the CT-ICH dataset. P values were computed by one-sided paired t tests on the aggregate distributions. Results were further stratified by hemorrhage type.

Performance as a function of training set size.— We studied how performance depends on training set size by training additional DL models while reducing the number of labeled data available for training (ie, number of labeled images or examinations). Training was repeated an increasing number of times as the number of labels decreased to account for variance. For the same number of labels, P values were computed with one-sided t tests across the training replicates. Details are provided in Appendix S1.

Results

Examination-Level Binary Classification

DL models were compared for the task of predicting whether a new examination contained at least one image with signs of ICH on all datasets included in this study. Respectively, the

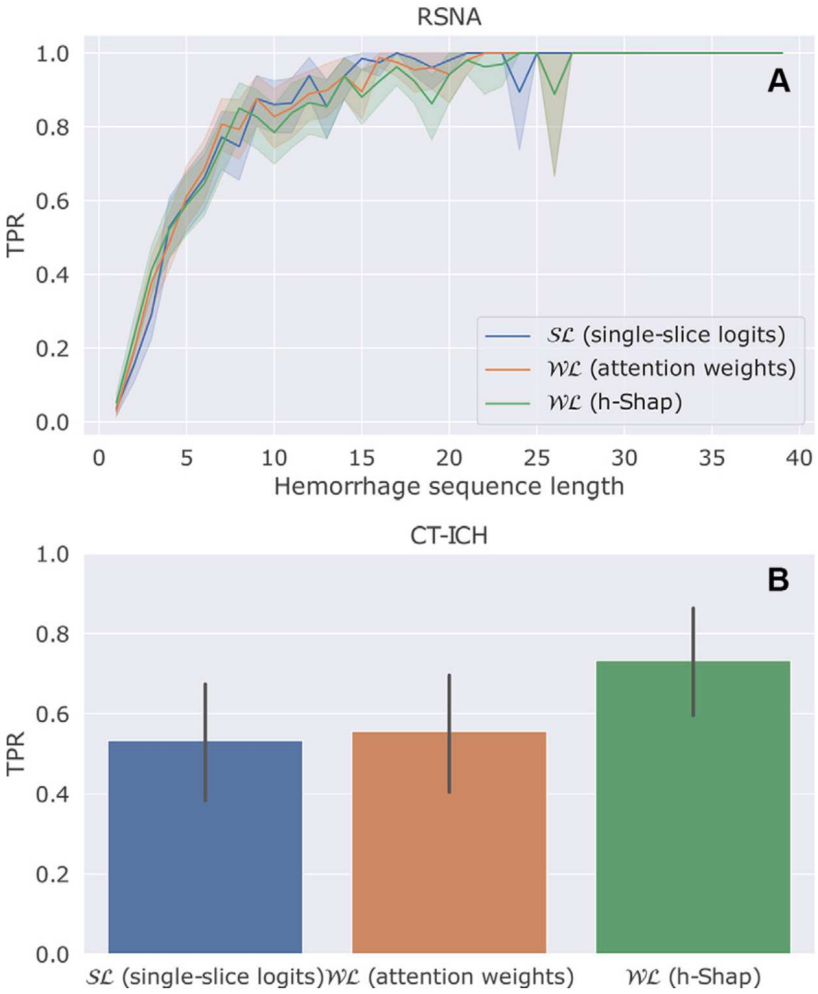

Figure 3: Graphs of section-level mean hemorrhage detection recall (ie, true-positive rate [TPR] = true positive divided by positive) and 95% CIs for a strong learner (SL) and a weak learner (WL) with attention weights and h-Shap on **(A)** the RSNA 2019 Brain CT Hemorrhage Challenge dataset and **(B)** the CT-ICH dataset. **(A)** Average recall stratified by the true hemorrhage sequence length (ie, the number of consecutive positive images in the true hemorrhage sequence). **(B)** Recall over true hemorrhage sequences of any type in the CT-ICH dataset. ICH = intracranial hemorrhage, RSNA = Radiological Society of North America.

strong and weak learners had AUCs of 0.96 versus 0.96 (P = .64) on the internal validation split of the RSNA dataset (4348 examinations; 40.8% ICH; 150 492 total images), 0.90 versus 0.92 (P = .15) on the CQ500 dataset (436 examinations; 48.6% ICH; 15 156 total images), and 0.92 versus 0.95 (P = .03) on the CT-ICH dataset (75 examinations; 48.0% ICH; 2814 total images). Figure 2 reports receiver operating characteristic curves and respective AUCs for the two models on all datasets.

Section-Level Hemorrhage Detection

Figure 3 shows mean section-level hemorrhage recall with 95% CIs for the strong and weak learners on the internal validation split of the RSNA dataset and the external CT-ICH dataset. Figure 3A displays results on the validation split of the RSNA dataset stratified by true hemorrhage sequence length (ie, the number of consecutive positive images in a true hemorrhage sequence). We found no evidence of a difference in performance between the strong learner (average recall = 0.62) and the weak learner with either notion of image importance: attention weights (average recall = 0.63 [P = .65]) and Shapley values (average recall = 0.62 [P = .51]). Figure 3B includes results for any type of hemorrhage in the CT-ICH dataset without stratification given the relatively limited number of examinations in the dataset. We found a statistically significant difference between the weak learner with h-Shap (average recall = 0.73) and the strong learner (average recall = 0.53 [P = .02]).

Pixel-Level Hemorrhage Detection

Figures 4A and 4B display representative saliency maps for a strong and a weak learner on the CQ500 and CT-ICH datasets for every type of hemorrhage. The saliency maps appeared qualitatively similar with appropriate localization of the various ICH types, and they aligned well with ground truth annotations. Other findings that may correlate with the presence of ICH, such as external hematomas due to injury, midline shift effects, or compression of the ventricles, were not highlighted by the saliency maps. Figures 4C and 4D report the distribution of the Dice scores between saliency maps and ground truth manual segmentations of the bleeds. The ranges of the median scores across hemorrhage types for each learner and explainer on the CQ500 dataset were 0.45–0.71 (strong learner, Grad-CAM), 0.48–0.69 (strong learner, h-Shap), 0.48–0.72 (weak learner, Grad-CAM), and 0.42–0.65 (weak learner, h-Shap). Similarly, results on the CT-ICH dataset were 0.09–0.46 (strong learner, Grad-CAM), 0.16–0.40 (strong learner, h-Shap), 0.12–0.40 (weak learner, Grad-CAM), and 0.17–0.44 (weak learner, h-Shap). Table 3 reports the complete distribution of median and IQR values for all learners and explainers across hemorrhage types. For the aggregate distributions, we found statistically significant differences on the CQ500 dataset between the strong learner with Grad-CAM (average Dice score = 0.54) and the weak learner with Grad-CAM (average Dice score = 0.52 [P < .001]) and between the strong learner with h-Shap (average Dice score = 0.54) and the weak learner with h-Shap (average Dice score = 0.51 [P < .001]). On the CT-ICH dataset, we found no evidence of a difference between the strong learner with Grad-CAM (average Dice score = 0.32) and the weak learner with Grad-CAM (average Dice score = 0.31 [P = .26]) or between the strong learner with h-Shap (average Dice score = 0.31) and the weak learner with h-Shap (average Dice score = 0.33 [P = .95]). No single combination of learner and explanation method provided the highest median across all types of hemorrhage and datasets.

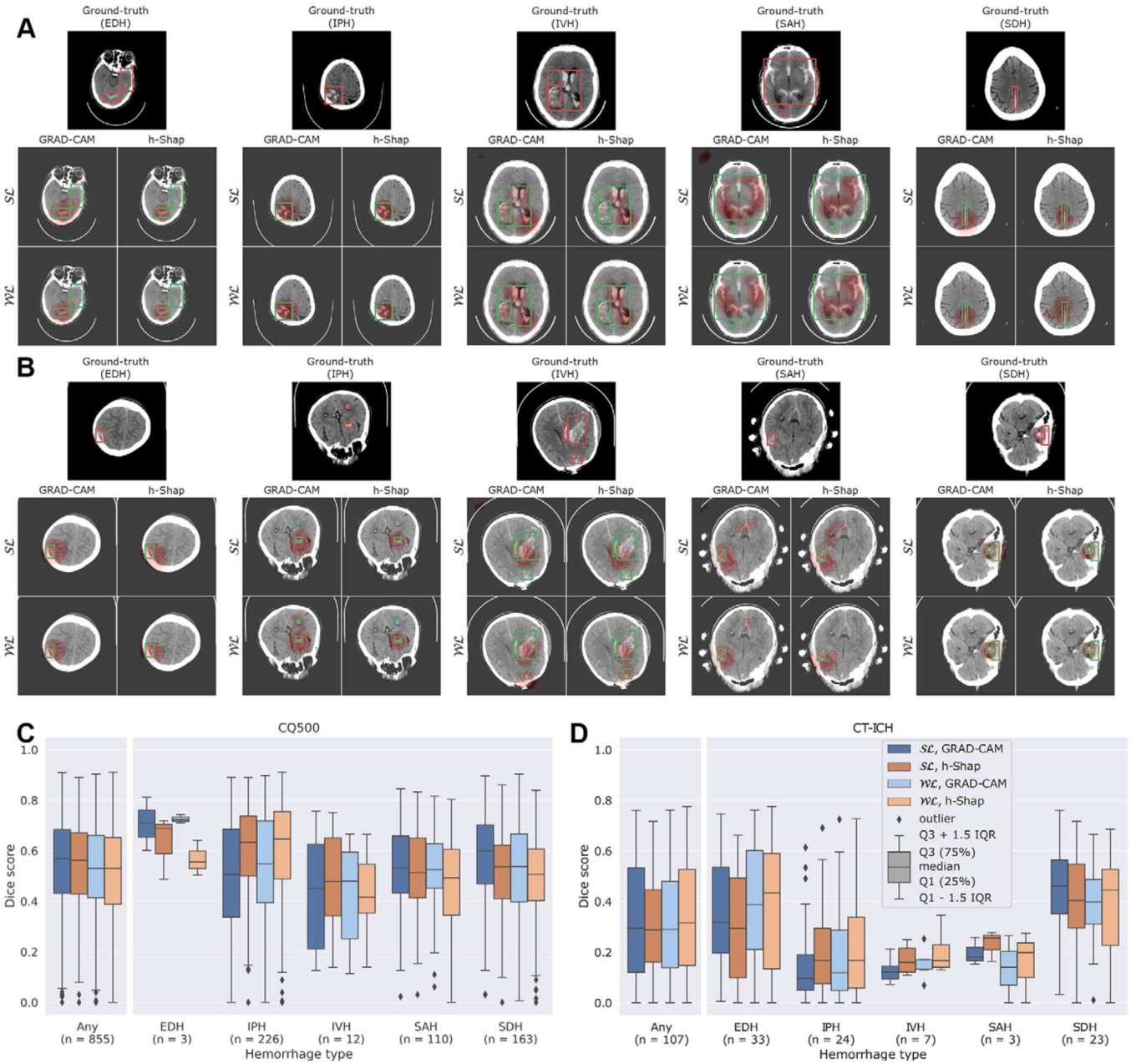

Figure 4: Comparison of a strong learner (SL) and weak learner (WL) on pixel-level hemorrhage detection. **(A, B)** Qualitative comparison of an example saliency map for every type of hemorrhage in the CQ500 dataset and the CT-ICH dataset, respectively. Saliency maps are obtained with Grad-CAM and h-Shap. **(C, D)** Quantitative comparison of the alignment of the saliency maps with the ground truth hemorrhage segmentations provided by radiologists by means of Dice scores. Results are stratified by hemorrhage type. EDH = epidural hemorrhage, ICH = intracranial hemorrhage, IPH = intraparenchymal hemorrhage, IVH = intraventricular hemorrhage, Q1 = 25th percentile, Q3 = 75th percentile, SAH = subarachnoid hemorrhage, SDH = subdural hemorrhage.

Performance as a Function of Training Set Size

Figure 5 compares strong and weak learners on the examination-level binary classification task and on section-level hemorrhage detection as a function of the number of labels available during training. Figures 5A–5C show mean AUCs and 95% CIs on examination-level binary classification. Weak learners had significantly better performance across all datasets for 10 000 labels. Respectively, the weak and strong learners had average AUCs of 0.95 versus 0.87 ($P < .001$) on the internal validation split of the RSNA dataset, 0.81 versus 0.77 ($P = .07$) on the CQ500 dataset, and 0.88 versus 0.75 ($P < .001$) on the CT-ICH dataset. Figure 5D showcases mean section-level hemorrhage detection f_1 score with 95% CIs. Examination-level supervision showed a sharp increase in performance between 1000 and 10 000 labels. For 1000 labels, strong learners (average $f_1 = 0.59$) had significantly better performance than weak learners (average $f_1 = 0.46$ [$P = .02$]). For 10 000 labels, weak learners (average $f_1 = 0.73$) had significantly better performance than strong learners (average $f_1 = 0.65$ [$P < .001$]).

Discussion

The collection and annotation of medical imaging datasets is a major bottleneck for the development of modern DL pipelines in radiology. For example, the 2019 RSNA Brain CT Hem-

Table 3: Dice Scores between Saliency Maps and Ground Truth Manual Segmentations on the CQ500 and CT-ICH Datasets for All Learners and Explanation Methods, Stratified by Hemorrhage Type

			Dice Score					
Dataset	Supervision	Explainer	All (855)	EDH (3)	IPH (226)	IVH (12)	SAH (110)	SDH (163)
CQ500	Strong	Grad-CAM	0.57 (0.25)	0.71 (0.11)	0.51 (0.35)	0.45 (0.41)	0.53 (0.23)	0.60 (0.24)
		h-Shap	0.56 (0.24)	0.69 (0.12)	0.63 (0.24)	0.48 (0.31)	0.51 (0.24)	0.54 (0.21)
	Weak	Grad-CAM	0.53 (0.25)	0.72 (0.02)	0.55 (0.33)	0.48 (0.34)	0.53 (0.18)	0.54 (0.27)
		h-Shap	0.53 (0.26)	0.56 (0.07)	0.65 (0.27)	0.42 (0.19)	0.49 (0.26)	0.51 (0.20)
			All (107)	EDH (33)	IPH (24)	IVH (7)	SAH (3)	SDH (23)
CT-ICH	Strong	Grad-CAM	0.29 (0.41)	0.32 (0.34)	0.09 (0.14)	0.12 (0.05)	0.18 (0.05)	0.46 (0.21)
		h-Shap	0.29 (0.28)	0.29 (0.39)	0.17 (0.22)	0.16 (0.09)	0.26 (0.06)	0.40 (0.25)
	Weak	Grad-CAM	0.29 (0.34)	0.39 (0.39)	0.12 (0.24)	0.17 (0.04)	0.14 (0.13)	0.40 (0.18)
		h-Shap	0.32 (0.38)	0.43 (0.46)	0.17 (0.28)	0.17 (0.09)	0.20 (0.14)	0.44 (0.30)

Note.—For each dataset, data are presented as medians and IQRs stratified by type of supervision, explainer, and hemorrhage type. Numbers in parentheses in column heads are numbers of images used for the comparison. EDH = epidural hemorrhage, ICH = intracranial hemorrhage, IPH = intraparenchymal hemorrhage, IVH = intraventricular hemorrhage, SAH = subarachnoid hemorrhage, SDH = subdural hemorrhage.

orrhage Challenge alone required 60 volunteer expert radiologists and thousands of hours (2). Nevertheless, it remains unclear whether expensive, image-level labels provide a clear advantage compared with cheap, examination-level labels for DL detection tasks. In this retrospective study, we investigated what kind of labels should be collected for ICH detection in head CT and, specifically, whether the classic supervised learning framework outperforms MIL. Attention-based MIL is desirable because it allows for a precise comparison with a strong learner implemented with the same feature extractor (ie, a ResNet-18), and the attention mechanism can be used to explain the model's prediction at the section level. We found that weakly supervised models maintained or improved performance across three different datasets while requiring drastically fewer labels. Our results suggest that MIL can facilitate the development of high-performing DL models with the use of annotations at the examination level compared with labor-intensive radiologist review of images.

On the examination-level binary classification task, the weak learner achieved high performance on the internal validation split, showing no evidence of a difference with the strong learner (AUC = 0.96 vs 0.96; P = .64). This agrees with other weakly supervised learning approaches that have been successfully applied to diagnosing abnormalities on whole-body fluorodeoxyglucose PET/CT using weak, examination-level labels (8). The examination-level prediction of the strong learner was defined as the maximum section-level prediction because it is the natural equivalent of the ground truth function. Alternatives exist (eg, quantiles, weighted sum), but they require tuning on an additional data split.

We also assessed generalizability of the strong and weak learners to external datasets, which is a critical feature for ensuring that artificial intelligence models will be suitable for clinical deployment. There were negligible drops in AUC (< 0.06) on external test sets for both learners, indicating good generalizability. Strong supervision did not have better performance for the CQ500 datasets (P = .15), and for the CT-ICH dataset, weak supervision achieved significantly better performance (P = .03). Generalizability is paramount to guarantee the safe deployment of DL models in real-world clinical scenarios, as these systems can often fail to generalize well to external data in radiology settings (27–29), including ICH detection on head CT scans (28). We speculate weak supervision may allow for more generalizable global features, though further study is required to confirm this.

Beyond examination-level classification, our results showed that strong supervision does not provide superior section-level hemorrhage-detection performance. Attention weights or Shapley values were used to select positive images within predicted positive examinations for the weak learner. For either notion of image importance, the strong learner did not achieve superior section-level recall (P = .65; P = .51). Our results confirmed that it is generally harder to detect shorter hemorrhage sequences. Recall was relatively stable for hemorrhage sequences longer than 10 images but dropped sharply with sequence lengths below 10. When both models were trained with 10 000 labels, weak supervision showed better section-level f_1 score (P < .001).

There may be several confounding factors that correlate with ICH but are not bleeds. Thus, gaining insights into the decision-making processes of DL models is important for medical applications to build trust with their end users. Certain laws require explanations of what led a DL model to recommend a specific treatment or diagnosis (30). In our study, Grad-CAM and h-Shap were used to verify whether both DL models learned to recognize signs of ICH. Saliency maps were compared both qualitatively and quantitatively in terms of their Dice scores with the ground truth segmentations of radiologists, and the results demonstrated strong supervision but did not guarantee improved performance on the CT-ICH dataset. These results suggest that pixel-level hemorrhage detection can be implemented without image-level labels by leveraging attention-based MIL and DL explanation methods. Neither learner was trained on ground truth segmentations, so we did not compare them with

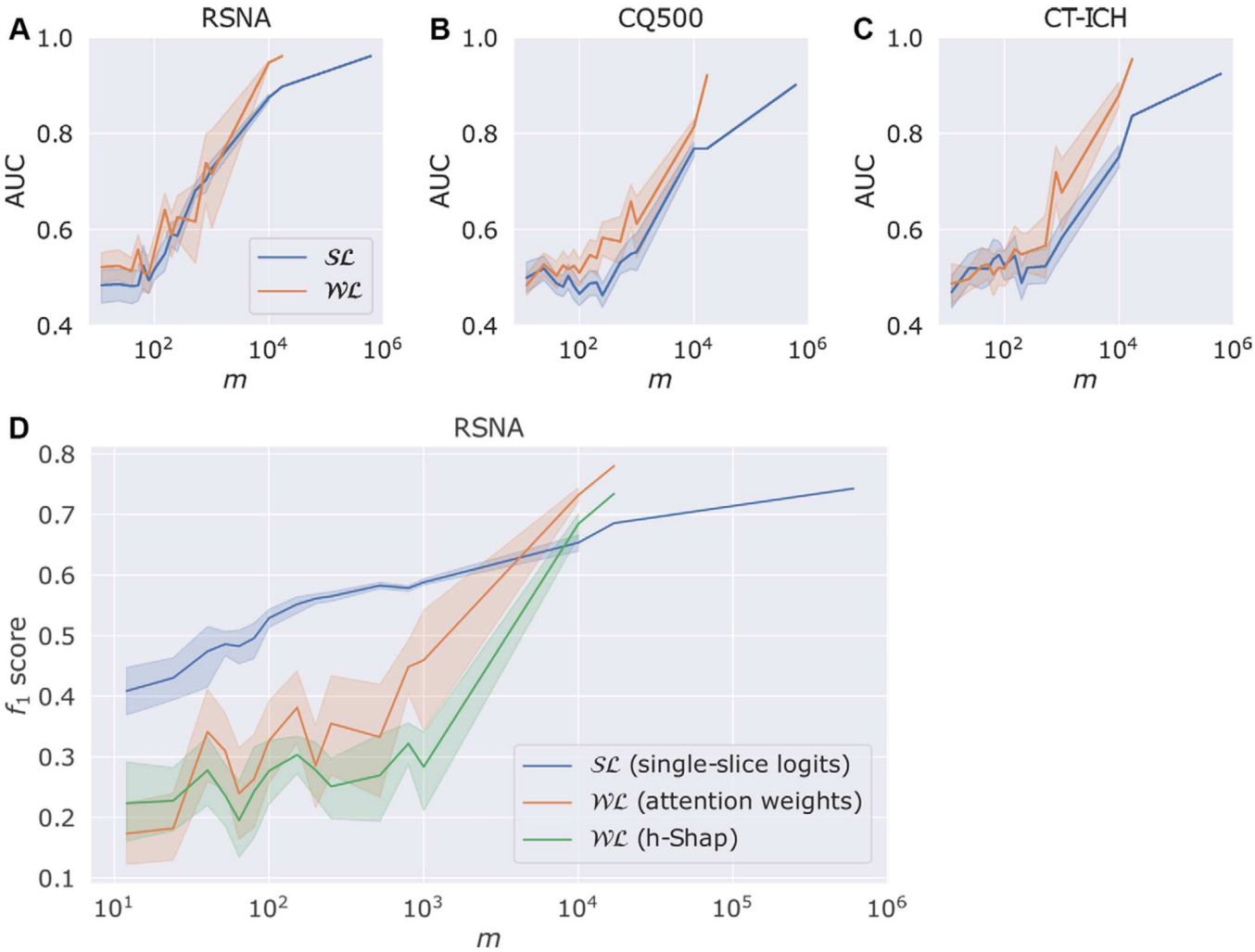

Figure 5: Graphs comparing several strong learners (SLs) and weak learners (WLs) on **(A-C)** examination-level binary classification and **(D)** section-level hemorrhage detection as a function of number of labels available during training, *m*. Results are reported across several training runs to account for variance. Solid lines represent the mean value across runs, and shaded regions represent 95% CIs. Data points with $m \geq 10000$ do not have 95% CIs because the training process is carried out only once. For the RSNA dataset, results were reported over a fixed subset of 1000 examinations from the validation split. AUC = area under the receiver operating characteristic curve, ICH = intracranial hemorrhage, RSNA = Radiological Society of North America.

segmentation models. We expand on the discussion of these results in Appendix S2 (Fig 6).

The ability to identify ICH at the section and pixel level is important in radiologists' workflows. DL models in triage use cases flag potentially actionable findings such as ICH (3–5). Radiologists then confirm whether they agree with the findings. The weaker sense of localization provided by explanation methods indicates weak learners could be deployed clinically to extend such flags beyond the presence of ICH and include specific images with saliency maps. This process is critical to allow radiologists to expeditiously confirm the diagnosis and to build trust with physician end users who have been shown to be wary of automated systems (31).

Our study had limitations. First, the scope of this study was limited to a single diagnostic use case of ICH detection on head CT scans. However, we note that the approach presented is applicable to other cross-sectional medical imaging methods, including CT on other body parts as well as other modalities like MRI. Future work includes validating this approach in other diagnostic scenarios. Second, some image-level annotations were still needed to validate our method on section- and pixel-level hemorrhage detection. This minimal amount of locally annotated data will also be necessary in future extensions. Only around 6% of the number of image-level labels otherwise necessary for strong supervision were employed in our study, which is feasible for future work. Other modern DL architectures, such as vision transformers, may provide additional advantage over our MIL weakly supervised methods and should be investigated to verify whether the findings of this work translate beyond convolutional neural networks. Finally, further studies should focus on quantifying the extent to which MIL can be advantageous to radiologists in clinical settings.

In conclusion, our results indicate that weak supervision may be sufficient for the task of ICH detection in head CT across three levels of granularity—*(a)* global binary classification, *(b)* section level, and *(c)* pixel level—if enough data are available ($m \gtrsim 5000$) while requiring approximately 35 times fewer labels. This approach could potentially save thousands of hours of annotation labor by radiologists, alleviating the

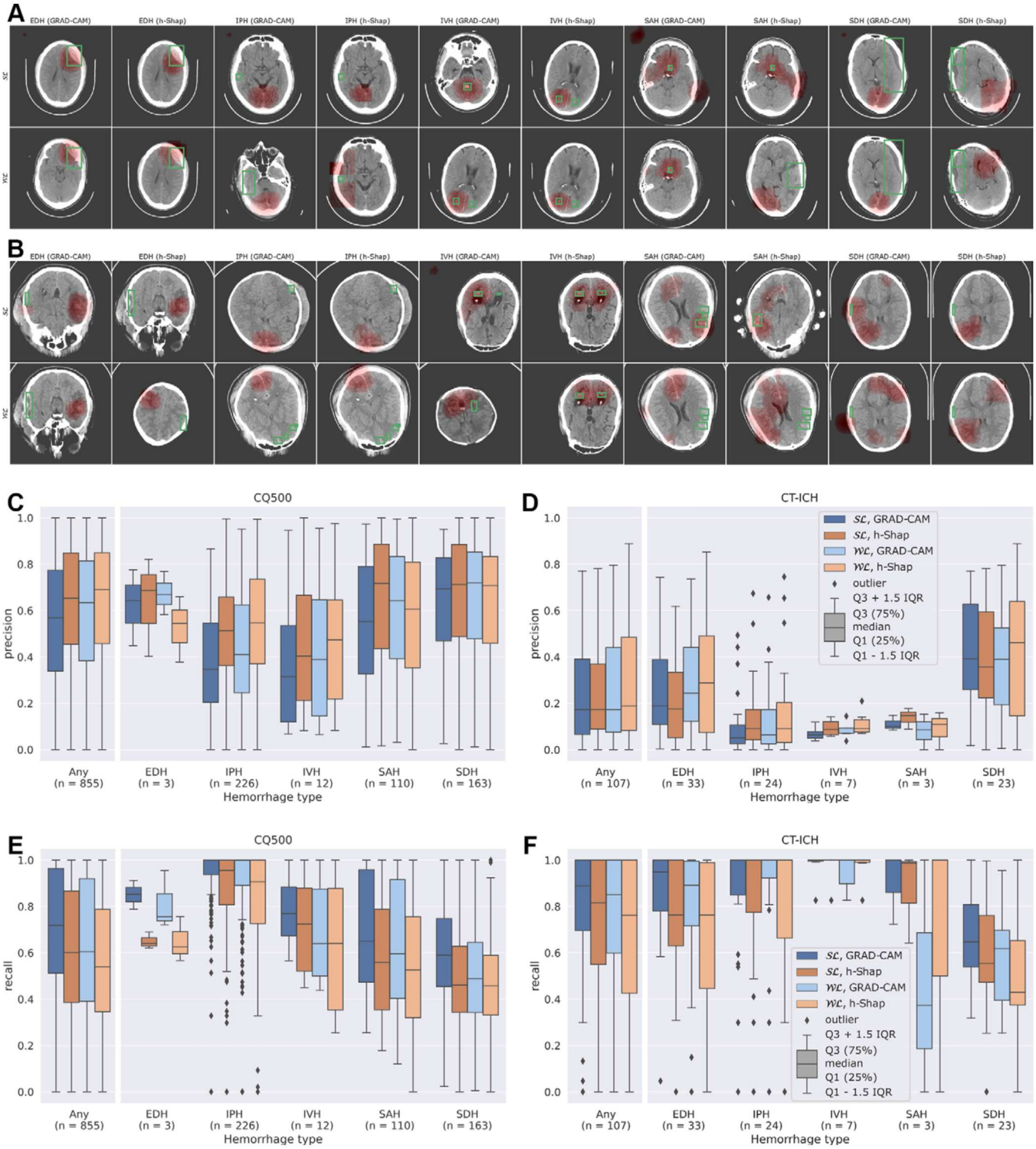

Figure 6: **(A, B)** Saliency maps with worst Dice score for each learner and explanation method across all types of hemorrhage in the CQ500 and CT-ICH datasets, respectively. **(C, D)** Graphs show precision of saliency maps. **(E, F)** Graphs show recall of saliency maps. EDH = epidural hemorrhage, ICH = intracranial hemorrhage, IPH = intraparenchymal hemorrhage, IVH = intraventricular hemorrhage, Q1 = 25th percentile, Q3 = 75th percentile, SAH = subarachnoid hemorrhage, SDH = subdural hemorrhage, SL = strong learner, WL = weak learner.

biggest bottleneck in developing high-performing DL models for medical imaging diagnosis.

Author contributions: Guarantor of integrity of entire study, **J.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **J.T., P.H.Y.**; experimental studies, all authors; statistical analysis, all authors; and manuscript editing, all authors

Disclosures of conflicts of interest: **J.T.** No relevant relationships. **P.H.Y.** Associate editor of *Radiology: Artificial Intelligence*; former trainee editorial board member of *Radiology: Artificial Intelligence*. **J.S.** No relevant relationships.

References

1. An SJ, Kim TJ, Yoon BW. Epidemiology, Risk Factors, and Clinical Features of Intracerebral Hemorrhage: An Update. J Stroke 2017;19(1):3–10.
2. Flanders AE, Prevedello LM, Shih G, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell 2020;2(3):e190211.
3. Buchlak QD, Milne MR, Seah J, et al. Charting the potential of brain computed tomography deep learning systems. J Clin Neurosci 2022;99:217–223.
4. Yeo M, Tahayori B, Kok HK, et al. Review of deep learning algorithms for the automatic detection of intracranial hemorrhages on computed tomography head imaging. J Neurointerv Surg 2021;13(4):369–378.
5. Kaka H, Zhang E, Khan N. Artificial Intelligence and Deep Learning in Neuroradiology: Exploring the New Frontier. Can Assoc Radiol J 2021;72(1):35–44.
6. 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(3):173–182.
7. Lee JY, Kim JS, Kim TY, Kim YS. Detection and classification of intracranial haemorrhage on CT images using a novel deep-learning algorithm. Sci Rep 2020;10(1):20546.
8. Eyuboglu S, Angus G, Patel BN, et al. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nat Commun 2021;12(1):1880.
9. Tushar FI, D'Anniballe VM, Hou R, et al. Classification of Multiple Diseases on Body CT Scans Using Weakly Supervised Deep Learning. Radiol Artif Intell 2021;4(1):e210026.
10. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 1997;89(1-2):31–71.
11. Quellec G, Cazuguel G, Cochener B, Lamard M. Multiple-Instance Learning for Medical Image and Video Analysis. IEEE Rev Biomed Eng 2017;10:213–234.
12. Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Adv Neural Inf Process Syst 1997;10:570–576. https://dl.acm.org/doi/10.5555/302528.302753.
13. Weidmann N, Frank E, Pfahringer B. A Two-Level Learning Method for Generalized Multi-instance Problems. In: Lavrač N, Gamberger D, Blockeel H, Todorovski L, eds. Machine Learning: ECML 2003. Vol 2837. Springer, 2003; 468–479.
14. 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(10162):2388–2396.
15. Hssayeni MD, Croock MS, Salman AD, Al-khafaji HF, Yahya ZA, Ghoraani B. Intracranial Hemorrhage Segmentation Using a Deep Convolutional Model. Data (Basel) 2020;5(1):14.
16. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215–E220.
17. Reis EP, Nascimento F, Aranha M, et al. Brain Hemorrhage Extended (BHX): Bounding box extrapolation from thick to thin slice CT images. PhysioNet.. Published July 29, 2020. Accessed November 21, 2021.
18. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. http://proceedings.mlr.press/v80/ilse18a.html.
19. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318–327.
20. Shapley L. 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307–317. In: Kuhn HW, ed. Classics in Game Theory. Princeton University Press, 1997; 69–79.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4768–4777. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
22. Teneggi J, Luster A, Sulam J. Fast Hierarchical Games for Image Explanations. IEEE Trans Pattern Anal Mach Intell 2023;45(4):4494–4503.
23. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. JAIR 2021;70:245–317.
24. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy (Basel) 2020;23(1):18.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017.
26. Coifman RR, Donoho DL. Translation-Invariant De-Noising. In: Antoniadis A, Oppenheim G, eds. Wavelets and Statistics. Lecture Notes in Statistics. Vol 103. Springer, 1995; 125–150.
27. Voter AF, Larson ME, Garrett JW, Yu JJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. AJNR Am J Neuroradiol 2021;42(8):1550–1556.
28. Voter AF, Meram E, Garrett JW, Yu JJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage. J Am Coll Radiol 2021;18(8):1143–1152.
29. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
30. Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". AI Mag 2017;38(3):50–57.
31. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31.