nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0324-9

An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets

Hyunkwang Lee[1,2,3], Sehyo Yune[1,3], Mohammad Mansouri[1], Myeongchan Kim[1], Shahein H. Tajmir[1], Claude E. Guerrier[1], Sarah A. Ebert[1], Stuart R. Pomerantz[1], Javier M. Romero[1], Shahmir Kamalian[1], Ramon G. Gonzalez[1], Michael H. Lev[1] and Synho Do[1]*

Owing to improvements in image recognition via deep learning, machine-learning algorithms could eventually be applied to automated medical diagnoses that can guide clinical decision-making. However, these algorithms remain a 'black box' in terms of how they generate the predictions from the input data. Also, high-performance deep learning requires large, high-quality training datasets. Here, we report the development of an understandable deep-learning system that detects acute intracranial haemorrhage (ICH) and classifies five ICH subtypes from unenhanced head computed-tomography scans. By using a dataset of only 904 cases for algorithm training, the system achieved a performance similar to that of expert radiologists in two independent test datasets containing 200 cases (sensitivity of 98% and specificity of 95%) and 196 cases (sensitivity of 92% and specificity of 95%). The system includes an attention map and a prediction basis retrieved from training data to enhance explainability, and an iterative process that mimics the workflow of radiologists. Our approach to algorithm development can facilitate the development of deep-learning systems for a variety of clinical applications and accelerate their adoption into clinical practice.

During the past decade, much progress has been made in machine learning, thanks to both increased computational power and the accumulation of big data. Advances in image recognition based on deep learning, a subset of machine learning, are changing the landscape of medicine by achieving physician-level performance in various tasks[1–3]. These breakthroughs can improve diagnostic accuracy, streamline physician workflow, provide expertise to underserved populations and even prompt the discovery of new biological insights.

Nevertheless, the need for big data and the black box problem represent substantial obstacles to the development and translation of medical deep-learning systems into clinical practice. The first important challenge in developing medical imaging deep-learning systems is access to large and well-annotated datasets. Deep learning is unique in its capability of recognizing meaningful patterns from raw data without explicit directions when provided with sufficient examples. To make the most of this capability, previously published medical imaging deep-learning studies that have achieved physician-level performance have used over 100,000 images to train their networks[1–3]. However, collecting and labelling such large datasets is onerous and often infeasible for many researchers. Moreover, institutions are often hesitant to share data with external collaborators because of patient privacy, as well as ethical and legal considerations. Even if a researcher manages to collect such large datasets, labelling and validating big data are expensive and time-consuming. The second important challenge is that the inner workings and decision-making processes of machine-learning algorithms remain opaque[4]. The US Food and Drug Administration requires any clinical decision support software to explain the rationale or support for its decisions to enable the users to independently review the basis of their recommendations[5]. To meet this requirement and gain trust from clinicians, medical deep-learning systems should provide explanations for their outputs. Overcoming these two major barriers to training and clinical implementation can facilitate the development and a broader clinician adoption of deep-learning technology into medical practice.

In this study, we address both these challenges by constructing understandable algorithms for rapid, accurate detection and classification of acute ICH on unenhanced brain computed-tomography (CT) scans, using a small and imbalanced dataset from fewer than 1,000 patients. Acute ICH is a potentially life-threatening condition that requires fast detection, and it must be rapidly distinguished from ischaemic stroke to prompt appropriate treatment and mitigate neurological deficit and mortality. As many facilities do not have subspecialty-trained neuroradiologists, especially at night and on weekends, non-expert healthcare providers are often required to make a decision to diagnose or exclude acute haemorrhage. A reliable second opinion that is trained by neuroradiologists can help make healthcare providers more efficient and confident to enhance patient care, empower patients and cut costs.

Results

Our proposed system for the detection and classification of ICH uses multiple ImageNet[6] pretrained deep convolutional neural networks (DCNNs), a preprocessing pipeline, an atlas creation module and a prediction-basis selection module (Fig. 1). The four DCNNs used for building our model are VGG16[7], ResNet-50[8], Inception-v3[9] and Inception-ResNet-v2[10]. The preprocessing pipeline was designed for developing a high-performance system from small and imbalanced data. The atlas is a set of training image patches and attention

[1]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA. [2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, USA. [3]These authors contributed equally: Hyunkwang Lee, Sehyo Yune. *e-mail: sdo@mgh.harvard.edu

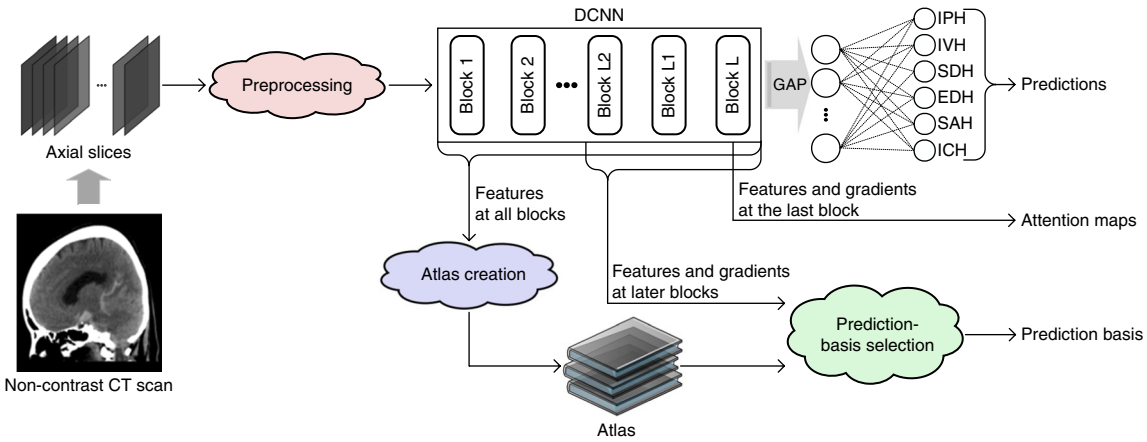

Fig. 1 | System overview. An illustration of the explainable deep-learning system for ICH detection and classification. The system includes ImageNet pretrained DCNNs, a pipeline of preprocessing techniques, an atlas creation module and a prediction-basis retrieval module. The system produces three types of outputs: predictions, attention maps and prediction basis. GAP, global average pooling.

maps that correspond to significant features of each of the five ICH subtypes. The prediction basis consists of training images retrieved from the atlas that are determined by the model to be relevant to a given image (Fig. 2).

Building a deep-learning model using small and imbalanced data. To develop the system, 904 non-contrast head CTs were retrieved from the picture archiving and communication system at our institution (Supplementary Table 1). By consensus, five subspecialty-trained US board-certified neuroradiologists labelled each of the two-dimensional (2D) axial images as one or more of the following: no haemorrhage, intraparenchymal haemorrhage (IPH), intraventricular haemorrhage (IVH), subdural haemorrhage (SDH), epidural haemorrhage (EDH) or subarachnoid haemorrhage (SAH). From the 904 cases, we randomly selected 100 cases with ICH ($n=42$ for IPH, $n=23$ for IVH, $n=18$ for SDH, $n=12$ for EDH and $n=51$ for SAH) and 100 cases without ICH as a validation dataset for hyperparameter tuning and model selection. The remaining 704 cases were used to train the model. The development dataset was intrinsically imbalanced due to the low incidence of EDH and SDH in our patient population. To counteract the skewed distribution of the training data, we introduced a preprocessing pipeline and network optimization techniques. First, we added an output that predicts the probability of any type of ICH, in addition to the five binary outputs to predict each type of haemorrhage—IPH, IVH, SDH, EDH and SAH. This allowed the model to learn the general representations of ICH, enhancing the overall sensitivity. Next, we imposed additional costs on the model for misclassifying positive instances by weighting the binary cross entropy (BCE) loss function by the ratios of positive and negative examples in each of the binary outputs[11,12] (Methods).

To address the relatively small size of the dataset, we introduced two other optimization steps inspired by the workflow of radiologists: multi-window conversion and slice interpolation. During clinical interpretation, radiologists adjust the window-width (WW) and window-level (WL) display settings to increase the conspicuity of subtle, potentially significant abnormalities (for example, 'haemorrhage-specific' WL display settings, with greyscale values preset to the typical CT density range of ICH). To mimic this approach, we utilized a multi-window conversion technique by generating three 8-bit greyscale images with different WL settings and encoding them into an RGB (red, green and blue) image: tissue window (WL = 40, WW = 40) for the red channel; brain window (WL = 50, WW = 100) for the green channel; and blood window (WL = 60, WW = 40) for the blue channel (Supplementary Fig. 3). Slice interpolation was introduced to mimic how radiologists integrate information from all adjacent images of a contiguous three-dimensional (3D) volume concurrently, rather than examine each single axial 2D slice in isolation. Interpolated images from adjacent slices were provided to the model with a modified loss function during training to imitate the 3D integration of image interpretation by radiologists. In addition, real-time data augmentation was performed by applying geometric transformations and a nonlinear denoising filtering step to improve the generalizability of the model to different rotations, scales, translations and noise (Methods). The incremental performance improvement achieved by each of these techniques is shown in Fig. 3. The ensemble model—defined as the unweighted average of probabilities predicted by all four optimized DCNNs—achieved the mean average precision (mAP) of 0.858.

Performance of the ICH detection and classification system compared to radiologists. To evaluate the perfomance of the model, two separate test datasets were collected retrospectively and prospectively after completion of the model development process. The retrospective test dataset included 100 cases with ICH and 100 cases without ICH, and the prospective test dataset included 79 cases with ICH and 117 cases without ICH. Among the 100 ICH cases in the retrospective test set, 79 had more than one type of ICH, with 53 labelled as IPH, 54 as IVH, 40 as SDH, 12 as EDH and 69 as SAH. Among the 79 ICH cases in the prospective test set, 20 had more than one subtype of ICH, with 16 labelled as IPH, 6 as IVH, 44 as SDH and 38 as SAH (Supplementary Table 1). The retrospective test dataset was enriched to include all subtypes of ICH (selected non-consecutively from our dataset). The prospective test dataset included all ICH cases consecutively collected from our emergency department over a 4-month period, reflecting the typical appearance of bleed subtypes that present in a tertiary emergency care setting. Testing was carried out on the case level, not the slice level. Figure 4 shows the receiver operating characteristic (ROC) curves for the performance of the ensemble model compared to that of five radiologists (three residents and two subspecialty-certified neuroradiologists, blinded to all data except the 5-mm axial CT slices) for the detection and classification of ICH on the two test datasets. The model achieved an area under the ROC curve (AUC) of 0.99 (95% confidence interval (CI) of 0.982–0.999) on the retrospective test set and an AUC of 0.96 (95% CI of 0.927–0.986) on the prospective test set. We selected the high sensitivity operating point from the validation subset of the development dataset given that maximizing

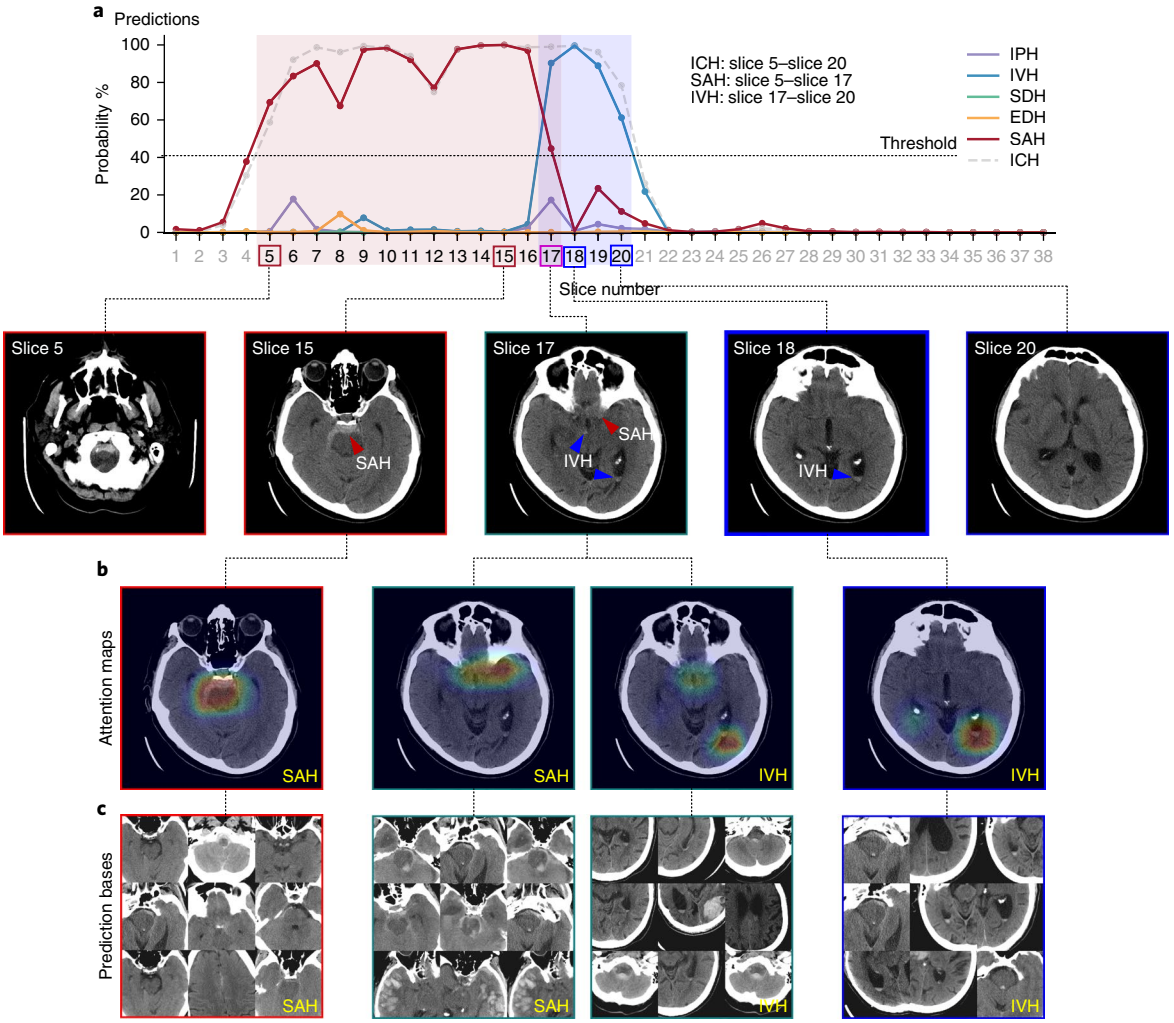

Fig. 2 | Summary of the system outputs. An example of a case with both SAH and IVH. **a**, Probabilities for the presence of each type of ICH by the DCNN on all slices in the test case are provided graphically, along with final predictions at a threshold. **b**, For each positive case, the system generates a colour-coded attention map (scale ranges from red, 'hot', to blue, 'cold') to indicate significant pixels in a test image. **c**, A set of prediction bases that are most relevant to each image.

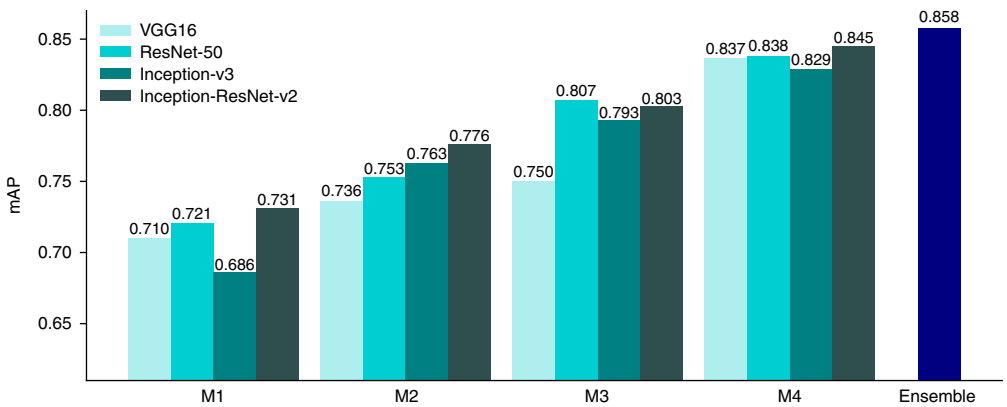

Fig. 3 | Iterative performance improvement by network optimization and preprocessing. Validation performance of the four DCNNs—VGG16, ResNet-50, Inception-v3 and Inception-ResNet-v2—as each technique was applied (M1, M2, M3 and M4). M1, DCNN baseline; M2, additional ICH output and weighted BCE loss; M3, M2 + multi-window conversion; M4, M3 + slice interpolation + real-time data augmentation. The mAP was used as a performance metric for comparisons. The ensemble of the four M4 models takes an unweighted average of the probabilities of the presence of ICH.

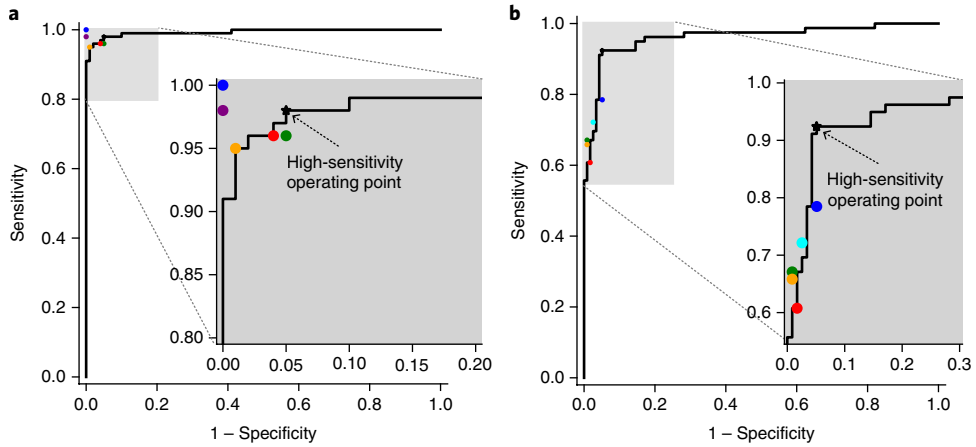

Fig. 4 | Test performance for ICH detection. ROC curves for ensemble model test performance (black lines) and radiologist performance (coloured circles) for the detection of ICH tested on the two separate datasets. **a**, The ensemble model tested with the retrospective dataset achieved an AUC value of 0.993 (95% CI of 0.982–0.999). Two radiologists outperformed the model, while three radiologists showed similar performance with that of the model. **b**, When tested with the prospective test dataset, the model achieved an AUC value of 0.961 (95% CI of 0.927–0.986) and showed higher sensitivity than any of the five radiologists at the predetermined operating point. Red, green and blue circles correspond to second-, third- and fourth-year radiology residents, respectively. Purple, cyan and orange circles correspond to attending radiologists with 9, 16 and 20 years of experience, respectively. ROC curves for each subtype of ICH are available in Supplementary Figs. 6 and 7.

Table 1 | Model performance on retrospective and prospective datasets in detecting and classifying ICH and its subtypes

	Retrospective			Prospective		
	AUC	Sensitivity (%)	Specificity (%)	AUC	Sensitivity (%)	Specificity (%)
ICH	0.993 (0.982, 0.999)	98.0 (95.3, 100)	95.0 (90.7, 99.3)	0.961 (0.927, 0.986)	92.4 (86.6, 98.2)	94.9 (90.9, 98.9)
IPH	0.980 (0.963, 0.993)	92.5 (85.4, 99.6)	91.8 (87.4, 96.2)	0.921 (0.843, 0.983)	68.8 (46.1, 91.5)	95.0 (91.8, 98.2)
IVH	0.979 (0.961, 0.992)	87.0 (78.0, 96.0)	95.9 (92.7, 99.1)	0.973 (0.910, 1.000)	83.3 (53.5, 100)	99.5 (98.5, 100)
SDH	0.959 (0.929, 0.983)	87.5 (77.3, 97.7)	86.9 (81.7, 92.1)	0.881 (0.812, 0.943)	70.5 (57.0, 84.0)	92.8 (88.7, 96.9)
EDH	0.922 (0.851, 0.978)	58.3 (30.4, 86.2)	95.2 (92.1, 98.3)	NA	NA	NA
SAH	0.960 (0.933, 0.980)	84.1 (75.7, 92.7)	88.5 (83.0, 94.0)	0.926 (0.883, 0.962)	76.3 (62.8, 89.8)	89.9 (85.2, 94.6)

The 95% CIs on the metrics are provided in parentheses. No cases of EDH were included in the prospective dataset.

the sensitivity for ICH detection is the primary clinical motivation for the development of this tool. Using this threshold, the sensitivity and specificity for ICH detection were 98.0% (95% CI of 95.3–100%) and 95.0% (95% CI of 90.7–99.3%), respectively, for the retrospective test set. For the prospective test set, the sensitivity was 92.4% (95% CI of 86.6–98.2%) and the specificity was 94.9% (95% CI of 90.9–98.9%). The model showed a comparable performance to that of radiologists in the retrospective test set, but achieved consistently higher sensitivity than that of the human experts on the prospective test set. For ICH classification, the algorithm achieved AUC values between 0.92 (for EDH) and 0.98 (for IPH) in the retrospective test set, and between 0.88 (for SDH) and 0.97 (for IVH) in the prospective test set (Table 1; Supplementary Figs. 6 and 7). Table 1 summarizes the model performance for both test datasets, showing higher specificity than sensitivity in subtype classification. In detecting the presence of ICH, regardless of subtype, both sensitivity and specificity were over 92% in both datasets.

Localization accuracy of attention maps. The attention maps generated by the CNN were evaluated for accuracy by calculating the proportion of 'bleeding points' (selected by neuroradiologists to indicate the centre of haemorrhagic lesions) that overlapped with segmentation maps generated from the attention maps (Methods; Supplementary Fig. 2). The mean area of the segmentation maps was 8.9 ± 2.1% of the total head area. By consensus, two radiologists annotated bleeding points for 100 randomly selected slices from the test dataset that were predicted as haemorrhage-positive by the model. Of these, 98 slices contained one or more bleeding points that overlapped with those of the segmentation maps; in two slices, the radiologists did not find evidence of bleeding. Of the 40 slices that contained 1 or 2 bleeding points, the overlap rate was 91.1% (51 out of 56 bleeding points), and of the 58 slices that contained 3 or more bleeding points, the overlap rate was 75.5% (209 out of 277 bleeding points). Overall, 260 out of all 333 bleeding points (78.1%) annotated by the neuroradiologists overlapped with the segmentation maps.

Prediction basis for understanding model decisions. To make model decisions understandable to human users, we developed a visualization tool to generate an atlas and display the basis of the predictions of the model (Fig. 5). Our system first creates an ICH atlas from the training dataset, which contains various representations of each ICH subtype. This process feeds all training images through the trained network, sorting the feature maps from all convolutional layers by their maximum activation values from highest to lowest, and tracking their relevance to each of the five ICH subtypes. The most relevant features for each subtype were chosen as components of the atlas (Methods). Figures 2 and 5 present

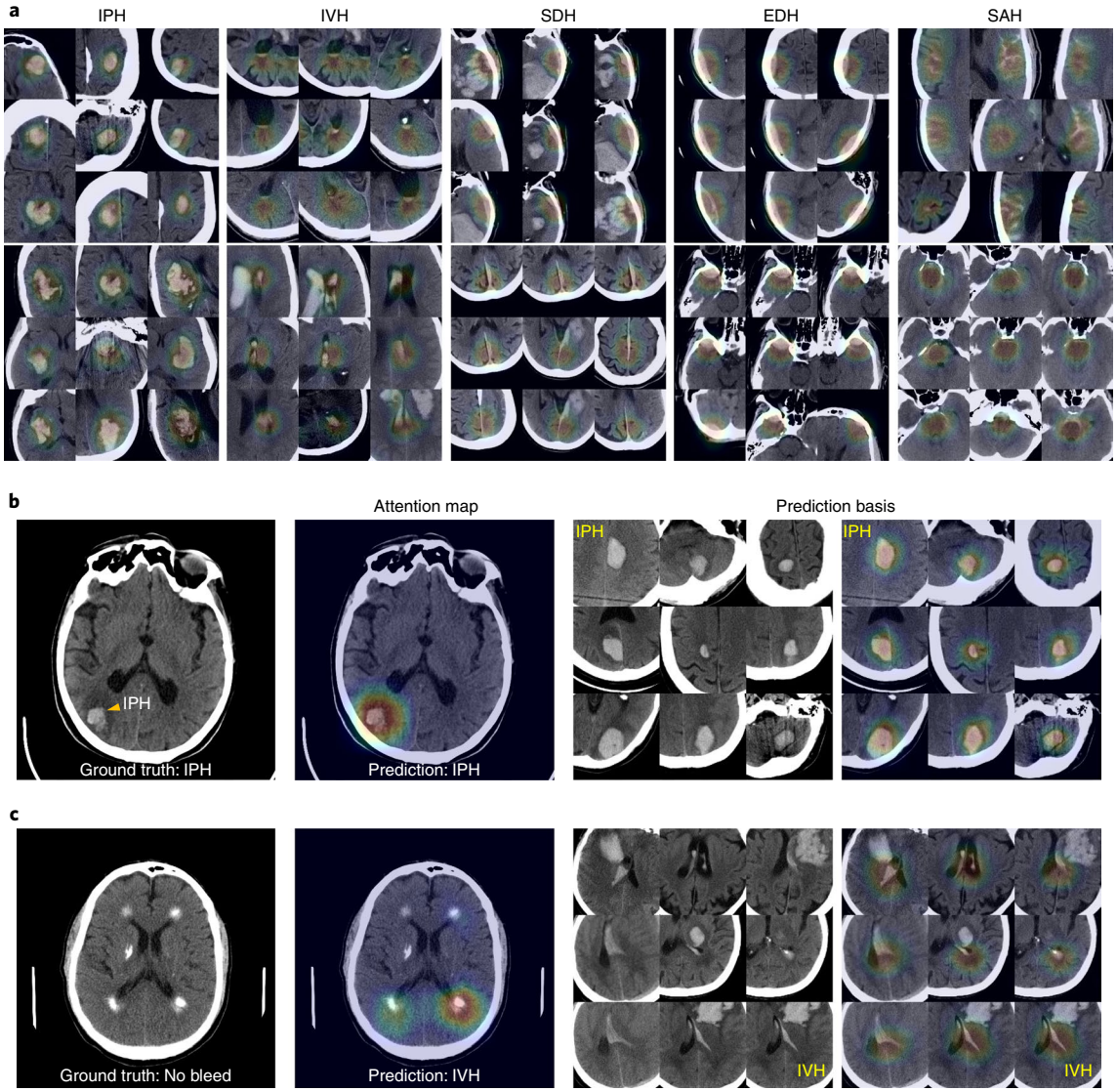

Fig. 5 | Examples of ICH atlas and prediction basis. a, Colour-coded activation maps (that is, heatmaps) overlaid with the corresponding image patches were generated for each ICH subtype from the training images and collected to make an ICH atlas. **b**, A true-positive test CT image with IPH (left) and corresponding attention map (middle left). IPH example images that are determined to be similar to the test case are retrieved from the atlas as the prediction basis (middle right and right). **c**, Example of a false-positive, haemorrhage-negative case with confounding periventricular calcifications that was predicted as IVH. The prediction basis shows retrieved atlas images of IVH, which human users can use to understand the justification of the model for its prediction by comparing the morphological similarities between the CT image and the prediction basis in the regions identified by the attention maps.

examples of atlas images that are considered significant for each subtype. During classification, the class activation mapping (CAM)[13,14] technique was applied to features from the final convolutional layer to generate an attention map that highlights the important regions relevant to the predictions of the model. Next, we computed the L2 distances of individual attention maps generated from features using CAM from the overall attention map and sorted the features by distance from lowest to highest. We also retrieved the training images and activation maps relevant to the selected features from the atlas to provide the prediction basis for a given case. The agreements between 3 independent neuroradiologists and the algorithm for selecting morphologically similar prediction-basis images were 93%, 94% and 95%, with Cohen's kappa values of 0.91, 0.92 and 0.93, respectively. The agreements between the radiologists were 94% (radiologists A and B), 93% (radiologists B and C) and 90% (radiologists A and C).

Discussion

We integrated several methods to achieve explainable, radiologist-level performance by using deep learning from a dataset of fewer than 1,000 cases to develop our model. Our platform detects ICH on unenhanced head CT scans in an independent test set of mixed, representative cases with and without haemorrhage. Moreover, over 90% sensitivity and specificity were achieved in two test datasets collected retrospectively and prospectively. Overall accuracy compared favourably with that of blinded reads from five radiologists for the retrospective test set, and the model sensitivity exceeded that of radiologists for the prospective test set. It is noteworthy that our system is distinct from previous work in its classification of five subtypes of ICH with radiologist-level performance, and provision of reliable localization and prediction basis. We achieved this by using a relatively small dataset and without manual pixel-level segmentation of the lesions. Our rigorous study design—including

high-quality data that contains all ICH subtypes, performance comparison with radiologists and quantification of haemorrhage localization—sets us apart from other studies. In addition, to our knowledge, the radiologist-mimicking techniques and provision of prediction basis has not been attempted by previous work.

Despite recent efforts in applying deep-learning techniques to diagnostic imaging, large datasets have remained a requirement for achieving physician-level performance[1–3]. Moreover, lack of precise measurement and meticulous validation of data reliability can result in failure of promising models that use big data[15]. Our training dataset was labelled by five expert neuroradiology subspecialists with 9–34 years of experience, providing reliable ground truth. This high-quality dataset, combined with our preprocessing and network optimization techniques, enabled the successful development of our deep-learning model. Our literature review identified only two studies that used deep learning for ICH detection and compared the performance of the model to that of radiologists[16,17]. One of these studies[16] used 4,304 head CT scans and 165,809 slices, each manually labelled with the ICH subtype. The other study[17] used a much smaller dataset, containing 252 scans, and although ICH subtypes were not classified, each slice was manually segmented to delineate the haemorrhagic regions. Our model showed a performance comparable to radiologists not only for ICH detection but also for classification of ICH subtypes, and was developed using a dataset containing only 904 CT scans.

Another notable feature of this study is the inclusion of ICH subtypes. A major challenge in medical image analysis is to recognize and characterize the wide variety of appearances of certain conditions. Brain haemorrhage may occur infrequently even in large datasets, but is nonetheless clinically important to detect and classify accurately. Although a few other studies that developed ICH detection models used smaller-sized datasets than ours[18–20], none of these specified haemorrhage subtypes or compared model performance against that of radiologists. Thus, the morphological diversity and subtlety of the lesions used as test datasets for evaluating those models is unknown. For example, one study achieved an AUC of 1.00 for detecting basal ganglia haemorrhage on unenhanced head CT scans. However, detection of only a single specific type and location of haemorrhage, although an important step forward, is impractical for clinical practice[18]. Another study also reported that their model achieved an AUC of 1.0 using 100 cases, but their processes for collection, labelling and splitting of data for development and validation were not clearly described[19]. Another group aggregated five imaging findings from unenhanced head CT examinations into two categories to develop a binary, automated critical findings identification and online notification system[20]. Their algorithms showed an AUC of 0.91 for detecting haemorrhage, mass effect or hydrocephalus, but the performance for detecting ICH alone—essential in certain clinical scenarios such as triaging patients with acute stroke—was not reported.

We achieved high performance from a relatively small-sized dataset by mimicking the workflow of radiologists. We encoded an image into a multichannel composite with standard and customized WL settings to allow the system to consider multiple versions of the input image, which improved performance, especially for detecting subtle lesions (Fig. 3, M3). Other studies used fixed WL settings to detect haemorrhage[16,17,20,21] or multiple WL settings for data augmentation[18,19]. There have been a few studies that used full-range 12-bit or 16-bit images to develop deep-learning models for plain radiographs[22,23]. However, plain radiographs have no absolute pixel values, whereas CT images can be quantified in absolute Hounsfield unit values to be used for comparison across different studies. During our experiment, simulating standard radiologist practices to set a narrow WW greyscale display to detect subtle haemorrhage led to a higher performance than using full-range 12-bit images. We believe that filtering out irrelevant information from the image, by setting a few prespecified display settings, helped the deep-learning model to recognize subtle lesions, as it does for human radiologists. Another radiologist-mimicking approach we used was slice interpolation. This technique reduced the false positive interpretation by the model of partial volume effects and intracranial calcifications (Fig. 3, M4). Although a long short-term memory network was previously used to consider inter-slice dependency, the network required manually segmented data from thin-slice head CT images[17]. Long short-term memory could also adversely affect the final output by unnecessarily incorporating distant lesions in the examination into the prediction.

Another approach to address inter-slice dependency is to build a 3D network that directly inputs the voxel data from the entire imaging volume into a 3D format rather than as pixel-data from discrete axial slices in a 2D format. To compare the 3D versus 2D approaches, we trained a 3D model using previously described methodology[21] by using case-level labels aggregated from slice-level labels, as well as volume data with a standardized dimensionality ($24 \times 512 \times 512$ voxels) generated using 2D slices. The resulting 3D model, however, achieved a mAP of only 0.328 for the multi-label classification of our five ICH subtypes, which is substantially inferior to the mAP we obtained with our existing 2D model (mAP of 0.686). This finding is consistent with the 'curse of dimensionality' reported in a previous study[24], which noted that the amount of data required to train a deep-learning model scales exponentially with the dimensionality of the data.

Precise localization is important not only for understanding the output of a model but also for rapidly confirming the reliability of the output, especially for potentially false positive cases. Although manual segmentation of regions of interest might enable more accurate localization and an increased focus on salient features[17], this requires an enormous—and possibly unsustainable—time and effort commitment by expert radiologists. Some previous studies also proposed algorithms that provide attention maps[1,3,11,25]; however, most either did not confirm the accuracy of localization compared to radiologists[1,3,16,25] or failed to show reliable accuracy[11]. Haemorrhagic lesions on head CT images can be heterogeneous and scattered across broad regions (Supplementary Fig. 2), and their description is often variable, even among subspecialty-trained neuroradiologists[26]. We achieved overlap between the bleeding points selected by neuroradiologists and the segmentation maps in 98 of 100 test slices, without the need for tedious and costly manual pixel-level segmentation.

Another noteworthy feature of our approach is the creation of prediction basis. Although attention maps are useful for end users to find the regions of interest identified by the model, they do not fully explain the reasons for the decisions made by the model. To better understand the logic behind model predictions, we propose a method that includes the automatic selection of significant and invariant features for detecting the presence of the ICH subtypes. We found that our models learnt invariant features of various appearances and locations of ICH (Figs. 2 and 5), a finding similar to that observed in previous works that utilized natural images[27,28]. We additionally incorporated retrieval of representative training images containing significant features relevant to each test image (that is, a representative feature atlas), which provide an explanation for the predictions of the model as a supplement to the attention maps. The degree of agreement between the prediction bases and the radiologist assessments for haemorrhage characterization were as high as the inter-observer agreement between the individual neuroradiologists. Showing morphologically similar, prelabelled cases of specific haemorrhage subtypes compared to a test case can be especially helpful for users with insufficient experience with ICH. By looking at the localization map and prediction basis, a user can understand the model prediction. It may not only increase confidence levels of clinicians for making or excluding a diagnosis but

also provide educational feedback at the point-of-care that will benefit non-experts such as residents, general radiologists and other non-radiologist clinicians.

Through this study, we also gained insights into approaches for optimizing deep-learning model performance. Previous work has demonstrated that deeper neural networks deliver better visual recognition performance than shallower networks when the training dataset is kept constant[29]. Simply choosing the deepest network, however, was not the answer in our study. We achieved much greater performance gains by iteratively adding ICH-specific preprocessing and network optimization techniques (Fig. 3, M1 to M4) compared to the small incremental improvements obtained utilizing deeper and more complex neural networks. These results imply that application-specific customization techniques are more effective to improve performance than the choice of the underlying CNN architecture. Additionally, combining the four M4 DCNNs into an ensemble model improved performance, possibly by way of consensus agreement, in much the same way that having multiple human raters can improve accuracy.

A limitation of our study is that all cases were collected from a single institution while excluding several conditions that are present in practice, raising concerns about the generalizability of the system. The distant time periods between when the normal and haemorrhage cases in the test dataset were collected could have also introduced bias. However, the datasets were carefully selected to reflect all types of ICH that are present in a tertiary care emergency setting. Future validation of our system should include data from several institutions, scanner manufacturers and acquisition protocols. Further testing on a prospective, consecutive dataset of patients presenting to the emergency department is in progress for improving the performance and generalizability of the system. Additional optimization and integration of the prediction-basis justification modules are also needed for these to be seamlessly embedded into the clinical workflow. Continual user feedback to understand unmet needs, by working closely with physicians and other domain experts, will be critical in the further development of practical deep-learning tools that can expand our diagnostic imaging armamentarium and help make clinicians more efficient.

Our explainable system offers a practical tool that could stimulate greater physician adoption. Currently, there is widespread belief that the answers to many crucial questions can be found from big data by using deep learning. However, a large portion of healthcare big data is unstructured and equivocal, making it unsuitable for building a deep-learning model. The balance between data size and quality plus careful data processing tailored to each application is the key to developing high-performance deep-learning algorithms.

Methods

This study was compliant with the Health Insurance Portability and Accountability Act and was approved by the Institutional Review Board of the Massachusetts General Hospital.

Retrospective collection of the development and test datasets. All DICOM (digital imaging and communications in medicine) images were de-identified before data analyses. From our institutional research database, we screened reports of non-contrast head CT scans to identify cases with or without acute ICH. Cases with a history of brain surgery, skull fracture, intracranial tumour, intracranial device placement, cerebral infarct or non-acute ICH were excluded. If a patient had multiple examinations, only the index study was included. A subspecialty board-certified neuroradiologist reviewed the initially identified cases and confirmed 625 ICH-positive cases and 279 ICH-negative cases to be included in the development dataset. Within the development dataset, 100 ICH-positive and 100 ICH-negative cases were randomly selected to be set aside for use as validation data, while 525 ICH-positive and 179 ICH-negative cases were used to train the models. For testing of the model, we collected an additional 100 ICH-positive and 100 ICH-negative, non-consecutive cases from the same database after completing the model development (the retrospective test set). The development dataset contained CT scans acquired between June 2003 and July 2017, and the retrospective test dataset contained cases from February 2005 to August 2017. The retrospective test set was enriched to include all subtypes of ICH (selected non-consecutively to test a broad range of bleed-subtype appearances).

Labelling of the development and test datasets. Five US subspecialty board-certified neuroradiologists (9–34 years of experience) annotated every slice in the development dataset based on consensus. Each slice was labelled with the presence or absence of IPH, IVH, SDH, EDH or SAH. For the retrospective test dataset, one of the five neuroradiologists with 24 years of experience (radiologist A) reviewed the clinical reports generated after consensus of two radiologists and labelled each case as positive or negative for each subtype. For annotation, the neuroradiologists incorporated additional data, including thin slices, sagittal and coronal reformatted images, and subsequent imaging tests such as CT and magnetic resonance imaging.

Prospective test data collection and labelling. To ensure generalizability of the test data, we additionally collected a prospective test dataset. All consecutive non-contrast head CT reports acquired for 4 months (from September to December 2017) from our hospital emergency department were reviewed by a physician. Based on the same exclusion criteria used for the retrospective dataset, 107 ICH-positive cases were identified. For compiling a balanced dataset, 130 consecutive ICH-negative cases were identified from the same consecutive CT studies. The prospective test dataset included all ICH cases identified from our emergency department over a 4-month period of data collection, reflecting the typical appearance of bleed subtypes (collected consecutively for each subtype) that presents to a tertiary emergency care setting. For labelling of this dataset, two neuroradiologists (radiologists A and D with 24 and 34 years of experience, respectively) independently annotated all 237 cases. In 31 out of the 237 cases, the annotation between the two neuroradiologists was discordant. A third neuroradiologist with 9 years of experience (radiologist B) reviewed these 31 cases, blinded to the reader identity, and provided the final annotation. After this image review, an additional 28 ICH-positive and 14 ICH-negative cases were excluded based on the exclusion criteria described above.

Standard and denoised non-contrast CT scans. In our institutional database, denoised images generated by commercial CT scanners using their own patented noise reduction algorithms were available in addition to corresponding standard images for 5-mm non-contrast head CT scans. In this study, the development dataset contained 650 cases with standard images only, 22 cases with denoised images only and 232 cases with both. The retrospective test dataset contained 92 cases with standard images only, 10 cases with denoised images only and 98 cases with both. The prospective test dataset contained 8 cases with standard images only, 10 cases with denoised images only and 178 cases with both. Standard images were used if standard images exist, while denoised images were used only when standard images were not available. Eventually, the development dataset included 882 standard and 22 denoised cases, the retrospective test dataset included 190 standard and 10 denoised cases, and the prospective test dataset included 186 standard and 10 denoised cases for the study.

Radiologist performance assessment. For comparison with the system, five radiologists with various levels of experience independently interpreted the two test datasets on the case level, based on the axial 5-mm series only, and blinded to clinical information and model output. The radiologists who interpreted the retrospective test dataset included first-, second- and third-year residents and two subspecialty board-certified neuroradiologists with 9 and 20 years of experience (radiologists B and C, respectively). For the prospective test, another board-certified neuroradiologist with 16 years of experience (radiologist E) replaced radiologist B, because radiologist B annotated the prospective dataset.

Problem formulation. ICH classification is a multi-label classification problem with five binary outputs predicting the presence of each ICH subtype. Each input image can be labelled as one or more ICH subtype[30]. The input is a 2D axial slice from a head CT scan x and its output is $y=\{y_1, y_2, y_3, y_4, y_5\}$ that indicates the probability of the presence of IPH, IVH, SDH, EDH and SAH, respectively. The sum of the BCE losses of the five outputs for a single instance is given by equation (1):

$$L_{\mathrm{BCE}}(x,y) = -\sum_{c=1}^{5} y^c \ln \hat{y}^c + (1-y^c)\ln(1-\hat{y}^c) \quad (1)$$

where y^c indicates the probability of the presence of a class label c that an algorithm predicts given an input image x. During deployment, slice level outputs were aggregated to a case level output by defining the presence of a positive ICH subtype on any slice.

Network training. Four validated DCNNs—VGG16[7], ResNet-50[8], Inception-v3[9] and Inception-ResNet-v2[10]—were selected to develop the ICH detection and classification system because they showed excellent classification performance

in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[31] (Supplementary Fig. 1). The DCNNs were pretrained on a subset of ImageNet used for the classification task in ILSVRC 2012—a 1.28 M natural image training dataset with 1,000 categories. The pretrained DCNN models, available from the official repository in Keras[32]—were then fine-tuned with the ICH training dataset after the last fully connected layers were replaced with three consecutive layers containing a global average pooling[33] layer, a fully connected layer and an element-wise sigmoid layer. All models were optimized using a mini-batch stochastic gradient descent with Nesterov momentum[34] with a batch size of 64 to maximize GPU (graphics processing unit) utilization. We used a weight decay of 5×10^{-5} and a base learning rate of 0.001, decayed by 0.1 three times when the validation loss plateaus. All experiments were conducted on a NVIDIA DevBox equipped with four TITAN X GPUs with 12 GB of memory per GPU, and all deep-learning models were implemented using Keras[32] (v.2.1.2) with a Tensorflow[35] backend (v.1.3.0).

Additional binary output and weighted loss function. We introduced the ICH binary value (Fig. 3, M2) to update the output of the network to $y=\{y_1, y_2, y_3, y_4, y_5, y_6\}$, giving the probabilities of IPH, IVH, SDH, EDH, SAH and ICH, respectively. The multi-label classification task was then reformulated into a binary classification, with ICH as positive if one or more of the subtype outputs were positive and negative if not. The BCE loss function was weighted by the ratios of positive and negative instances for each class label, in a similar fashion as described previously[11,12]. The modified BCE loss function is given by equations (2) and (3):

$$L_{\mathrm{W-BCE}}(x,y) = -\sum_{c=1}^{6} \alpha_P^c y^c \ln \hat{y}^c + \alpha_N^c (1-y^c) \ln(1-\hat{y}^c) \quad (2)$$

$$\alpha_P^c = \frac{|P^c| + |N^c|}{2\,|P^c|}, \alpha_N^c = \frac{|P^c| + |N^c|}{2\,|N^c|} \quad (3)$$

where $|P^c|$, $|N^c|$ are the total numbers of positive and negative instances of a class label c, and α_P^c, α_N^c are the corresponding loss weights. The balancing factors were calculated on the basis of heuristics[36].

Multi-window conversion. Medical images can range from 12-bits to 16-bits with corresponding greyscale resolutions of 4,096 to 65,536 shades of grey per pixel. This level of data is beyond the limits of human vision, and most medical displays support only 8-bit resolution[37]. For these reasons, CT images presented on digital displays are viewed applying predefined window levels and widths. Different window settings can increase the conspicuity of certain pathologies. For example, a 'haemorrhage' window setting is often used for the detection of ICH as it enhances the appearance of acute haemorrhage against brain parenchyma[38]. Additionally, other window settings, such as the 'stroke' window preset, have been demonstrated to help highlight subtle, potentially significant abnormalities[39]. Therefore, inspired by the way radiologists interpret CT images, we propose a multi-window conversion technique (Fig. 3, M3) by generating three 8-bit greyscale images with different window settings and encoding them into the following RGB images: tissue window (WL = 40, WW = 40) for the red channel; brain window (WL = 50, WW = 100) for the green channel; and blood window (WL = 60, WW = 40) for the blue channel (Supplementary Fig. 3a). This multi-window conversion is not specific to intracranial pathology and can be optimized for other clinical applications.

Slice interpolation. The slice interpolation technique (Fig. 3, M4) was implemented to have a model take into account neighbouring slices and reduce false positive rates. During training, input image x_s and label y_s were fed through a neural network together with interpolated data with adjacent slices to optimize the modified BCE loss function. The equations to compute interpolated images and labels are given by equation (4):

$$\begin{aligned} x_{s,s-1} &= \beta x_s + (1-\beta) x_{s-1}, y_{s,\,s-1} = \beta y_s + (1-\beta) y_{s-1} \\ x_{s,s+1} &= \beta x_s + (1-\beta) x_{s+1}, y_{s,\,s+1} = \beta y_s + (1-\beta) y_{s+1} \end{aligned} \quad (4)$$

where (x_s, y_s) correspond to the image and the label of the sth slice, $(x_{s,s-1}, y_{s,s-1})$ and $(x_{s,s+1}, y_{s,s+1})$ are images and labels of interpolated data from one slice lower (s − 1th) and upper (s + 1th), respectively (Supplementary Fig. 3b). The interpolation ratio β is randomly selected from the range of $0.5 < \beta < 1.0$ on every training batch. If an input image is the first or the last slice of the 3D scan, two interpolated slices are created from the existing neighbour using two different interpolation ratios β_1 and β_2. After training, we utilized varying degrees of interpolation, including 3, 5, 7 and 9, for classification. An interpolation degree n indicates the number of images to be used for predicting a single image. For degree 3, one original image and two interpolated images were generated using 0.67 as β. For degree 5, one original image and four interpolated images were created with 0.8 and 0.6 as β. For degree 7, one original image and six interpolated images were generated with 0.86, 0.72 and 0.58 as β. For degree 9, one original image and eight interpolated images were created using 0.89, 0.78, 0.67 and 0.56 as β. The final prediction of the target slice was constructed by taking a linear combination of multiple probabilities from the generated images with the corresponding interpolation ratios. The best interpolation degree was selected for each of the M4 models based on the performance using the validation dataset, with $n = 5$ for VGG16, $n = 3$ for ResNet-50, $n = 5$ for Inception-v3 and $n = 3$ for Inception-ResNet-v2.

Training data augmentation. Real-time data augmentation (Fig. 3, M4) was performed by applying geometric transformations (rotation, scaling and translation) to make models learn invariant features to geometric perturbations. In addition, to improve invariance of the model to noise, either standard or denoised images was randomly selected to be used. We generated denoised images for standard cases by applying a median filter with a window size of 3 and used the scanner-generated denoised images if they already existed in the datasets. For the cases only with scanner-generated denoised images, only the denoised images were used as we were concerned about a bias that might be produced by reversing the denoising processes that are unknown to us. Rotation angles ranging from −10° to 10° with an interval of 1°, scaling ratios of heights and widths ranging from 90% to 100% with an interval of 1%, translation parameters ranging from −12 to 12 pixels in x and y directions with an interval of 1 pixel, and a median filter with a window size of 3 were used for augmentation. All these parameters were randomly selected in the predefined ranges.

Model ensemble. To improve the performance of the model, an ensemble of the four M4 models was created using unweighted averaging such that the final probability is defined as an average of probabilities predicted by the four models. We tested different ensemble methods with the four models, including weighted averaging, majority voting and unweighted averaging. We selected unweighted averaging as our preferred ensemble method because its performance (mAP of 0.858 for multi-label classification of ICH subtypes) was similar to or nominally better than that of weighted averaging (mAP of 0.857) and majority voting (mAP of 0.855). Previous work has also demonstrated that simple averaging is effective to boost model performance[40].

Automatic atlas creation and prediction-basis justification. To generate the representative ICH atlas, all training images were presented again to the fully trained neural network to obtain the corresponding filter responses at each block. For each filter from each block, N activation maps were generated. These activation maps were sorted by their maximum activation values from highest to lowest. Then we quantified the relevance of each activation map to a particular label by counting the number of activation maps assigned to the label consecutively from the top (Supplementary Fig. 4a). For example, if the activation map with the highest maximum value is labelled as SAH and IPH, the second highest is labelled as SAH and the third is labelled as SDH, the relevance count is 2 for SAH, 1 for IPH and 0 for all other labels. Then all activation maps at all blocks for each label were sorted from the highest to lowest relevance count, and the top 5% of them were selected as having salient features to best represent the image assigned with the label (Supplementary Table 2). This served as the basis to create a radiology atlas of ICH within the training dataset with image patches and the corresponding activation maps for the selected features.

While performing inference on a test image, feature maps generated by the selected atlas filters and the corresponding gradients were extracted. The CAM technique[14] generates attention maps highlighting the important regions in a test image for the model prediction to a target label c. The mapping equation is given by equation (5):

$$M^c = \mathrm{ReLU}\left(\sum_k A_{\mathrm{L,k}} \cdot \frac{1}{S_{\mathrm{L}}} \sum_i \sum_j \frac{\partial z^c}{\partial A_{\mathrm{L,k}}}\right) \quad (5)$$

where z^c is a class score (logit) before being fed through the sigmoid function, S_{L} is the size of each feature map from the last block L, and ReLU corresponds to a rectified linear unit nonlinear function[41] (Supplementary Fig. 4b). M^c is the attention map generated using CAM with all feature maps A_{L} from the last block L. In addition, CAM can be utilized to generate individual attention maps $M_{l,k}^c$ for the features using the equation (6):

$$M_{\mathrm{l,k}}^c = \mathrm{ReLU}\left(A_{\mathrm{l,k}} \cdot \frac{1}{S_{\mathrm{l}}} \sum_i \sum_j \frac{\partial z^c}{\partial A_{\mathrm{l,k}}}\right) \quad (6)$$

where S_{l} is the size of each feature map from the block l. These individual attention maps $M_{\mathrm{l,k}}^c$ and the overall one M^c are L2-normalized and the L2 distances between them are then computed using the equation (7):

$$d_{\mathrm{l,k}}^c = \left\| \frac{M_{\mathrm{l,k}}^c}{\|M_{\mathrm{l,k}}^c\|_2} - \frac{M^c}{\|M^c\|_2} \right\|_2 \quad (7)$$

The attention maps with lowest L2 distances are utilized as index to retrieve representative training images from the atlas with similar significant features to justify the prediction of the network (Supplementary Fig. 4b).

Localization performance evaluation of attention maps. We performed quantitative evaluation of the accuracy of the attention maps in localizing haemorrhagic lesions through the following process. From the validation dataset, we randomly selected 100 cases with haemorrhage lesions and selected a slice containing haemorrhage lesions from each case based on the labels provided by the radiologists. Because the test dataset was labelled at the case level, we randomly selected one of the slices predicted as haemorrhage-positive by the model from each of the 100 ICH-positive test cases. Two board-certified neuroradiologists independently reviewed these 200 cases and marked the centre of haemorrhagic lesions with arrows, followed by consensus review. The radiologists did not have access to the attention maps, but they had access to the whole examinations that contain the reviewed slice. In the selected validation dataset, 74 cases contained single-type haemorrhage and 26 cases contained 2 or more types of bleeding in an image. In the selected test dataset, 64 cases were predicted as containing single-type haemorrhage and 36 cases as containing 2 or more types in an image. Segmentation maps were generated from the attention maps generated by the CNN using a simple thresholding technique by selecting pixels of which activation values were higher than 20% of the highest value in the attention maps, as described in earlier work[13]. The localization performance of the attention maps was then evaluated by calculating the proportion of bleeding points that overlapped with the segmentation maps. The mean area of the segmentation maps was also calculated and presented as a proportion to the total head area to determine how focused the attention maps are.

Evaluation of prediction-basis reliability. The 100 slices in the test dataset used for the accuracy evaluation of attention maps were also used for prediction-basis assessment. We developed a multiple-choice questionnaire (a–d) that presented readers with four sets of possible prediction-bases images corresponding to each case. Among the 787 image sets included in the atlas from the selected ResNet-50 model (Supplementary Table 2), we included the following in the questionnaire: (a) 1 set with the lowest L2 distance (most relevant to the case); (b, c) 2 sets with the 393th and 394th lowest L2 distance (moderately relevant); and (d) 1 set with the highest L2 distance (least relevant). The questionnaire displayed both the original image and the colour-coded attention map overlaid on this image, together with the choice of four different prediction-basis images (Supplementary Fig. 5). To avoid potential bias for the localization process by being exposed to the attention map results, the prediction-basis reliability multiple choice questionnaire was only administered to readers after completion of the localization process. Three radiologists selected the single set of prediction-basis images (a–d) that they rated as most similar morphologically to each test image. The percentage agreement and Cohen's kappa value were calculated between each radiologist and the model, as well as between the three radiologists.

Model selection. Model performance for the multi-label classification of ICH was defined as mean average precision, which is the mean of the areas under the precision-recall curves, for the multiple binary outputs. Average precision is known to be a single informative metric when there is a large skewness in the class distribution, such as our data[42]. The best models and hyperparameters were selected based on mAP values calculated from the validation dataset. ResNet-50 was selected to generate attention maps and the prediction basis based on its superior performance of localization in the validation dataset.

Statistical analyses. To assess the statistical significance of AUCs, we calculated 95% CIs using a non-parametric bootstrap approach via the following process. First, n ($n=200$) cases were randomly sampled from the test dataset of n cases with replacement, and the DCNN models were evaluated on the sampled test set. After running this process 2,000 times, 95% CIs were obtained by using the interval between 2.5 and 97.5 percentiles from the distribution of AUCs. The 95% CIs of percentage accuracy, sensitivity and specificity of the models at the selected operating point were calculated using binomial proportion CIs.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The deep-learning models were developed using standard libraries and scripts available in Keras and TensorFlow. Custom codes for the deployment of the system are available for research purposes from the corresponding author upon reasonable request.

Data availability

The training, validation and test datasets generated for this study are protected patient information. Some data may be available for research purposes from the corresponding author upon reasonable request.

Received: 14 March 2018; Accepted: 12 November 2018;
Published online: 17 December 2018

References

1. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
2. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
3. Rajpurkar, P. et al. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
4. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
5. *Clinical and Patient Decision Support Software. Draft Guidance for Industry and Food and Drug Administration Staff* (US FDA, 2017).
6. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
7. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/abs/1409.1556 (2014).
8. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
9. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (IEEE, 2016).
10. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proc. 31st AAAI Conference on Artificial Intelligence* 4278–4284 (AAAI, 2017).
11. Wang, X. *et al.* ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 3462–3471 (IEEE, 2017).
12. Sozykin, K., Khan, A. M., Protasov, S. & Hussain, R. Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks. Preprint at https://arxiv.org/abs/1709.01421 (2017).
13. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929 (IEEE, 2016).
14. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. Preprint at https://arxiv.org/abs/1610.02391v3 (2016).
15. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
16. Chilamkurthy, S. et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* https://doi.org/10.1016/S0140-6736(18)31645-3 (2018).
17. Grewal, M., Srivastava, M. M., Kumar, P. & Varadarajan, S. RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. In *IEEE International Symposium on Biomedical Imaging* 281–284 (IEEE, 2018).
18. Desai, V., Flanders, A. E. & Lakhani, P. Application of deep learning in neuroradiology: automated detection of basal ganglia hemorrhage using 2D-convolutional neural networks. Preprint at https://arxiv.org/abs/1710.03823 (2017).
19. Phong, T. D. et al. Brain hemorrhage diagnosis by using deep learning. In *Proc. 2017 International Conference on Machine Learning and Soft Computing* 34–39 (ACM, 2017).
20. Prevedello, L. M. et al. Automated critical test Findings identification and online notification system using artificial ntelligence in imaging. *Radiology* **285**, 923–931 (2017).
21. Arbabshirani, M. R. et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digit. Med.* **1**, 9 (2018).
22. Rubin, J. et al. Large scale automated reading of frontal and lateral chest X-rays using dual convolutional neural networks. Preprint at https://arxiv.org/abs/1804.07839 (2018).
23. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
24. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
25. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158 (2018).
26. Brinjikji, W. et al. Inter- and intraobserver agreement in CT characterization of nonaneurysmal perimesencephalic subarachnoid hemorrhage. *AJNR Am. J. Neuroradiol.* **31**, 1103–1105 (2010).
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Object detectors emerge in deep scene cnns. Preprint at https://arxiv.org/abs/1412.6856 (2014).
28. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. Preprint at https://arxiv.org/abs/1506.06579 (2015).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).
30. Tsoumakas, G. & Katakis, I. Multi-label classification: an overview. *Int. J. Data Warehousing Mining* **3**, 1–13 (2007).

31. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comp. Vision* **115**, 211–252 (2015).
32. Chollet, F. et al. Keras (2015); http://keras.io
33. Lin, M., Chen, Q. & Yan, S. Network in network. Preprint at https://arxiv.org/abs/1312.4400 (2013).
34. Nesterov, Y. A method of solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk USSR* **269**, 543–547 (1983).
35. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation* **16**, 265–283 (2016).
36. King, G. & Zeng, L. Logistic regression in rare events data. *Political Anal.* **9**, 137–163 (2001).
37. Kimpe, T. & Tuytschaever, T. Increasing the number of gray shades in medical display systems—how much is enough? *J. Digit. Imaging* **20**, 422–432 (2007).
38. Xue, Z., Antani, S., Long, L. R., Demner-Fushman, D. & Thoma, G. R. Window classification of brain CT images in biomedical articles. In *AMIA Annual Symposium Proceedings* 1023 (American Medical Informatics Association, 2012).
39. Turner, P. & Holdsworth, G. CT stroke window settings: an unfortunate misleading misnomer? *Br. J. Radiol.* **84**, 1061–1066 (2011).
40. Ju, C., Bibaut, A. & van der Laan, A. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Stat.* **45**, 2800–2818 (2018).
41. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th International Conference on Machine Learning* 807–814 (ICML, 2010).
42. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. *Proc. 23rd International Conference on Machine Learning* 233–240 (ACM, 2006).

Acknowledgements

The authors would like to acknowledge NVIDIA for the use of a DevBox and providing feedback and support, which made this work possible. R.G.G. is funded in part by an NIH U01 grant under the grant number 5U01EB025153.

Author contributions

H.L., S.Y., M.M., R.G.G., M.H.L. and S.D. initiated and designed the research. H.L., S.Y. and M.K. executed the research. M.M., S.H.T., C.E.G., S.A.E., S.R.P., J.M.R., S.K., R.G.G. and M.H.L. acquired and/or interpreted the data. R.G.G. and M.H.L. supervised the data collection. H.L., S.Y., M.H.L. and S.D. analysed and interpreted the data. H.L. and M.K. developed the algorithms and software tools necessary for the experiments. H.L., S.Y., S.H.T. and M.H.L. wrote the manuscript.

Competing interests

M.H.L. is a consultant of GE Healthcare and Takeda Pharmaceutical Company and receives an institutional research support from Siemens Healthcare. S.R.P. is a consultant of GE Healthcare. S.D. is a consultant of Nulogix and Doai and receives research supports from ZCAI, Tplus and MediBloc. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41551-018-0324-9.

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

Correspondence and requests for materials should be addressed to S.D.

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

nature research

Corresponding author(s): Synho Do

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	We used the Keras (v2.1.2) deep-learning framework with Tensorflow backend (v1.3.0) and python scripts.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated or analysed in this study are not publicly available because of institutional policy restricting public disclosure of patient-derived data. The datasets are available from the corresponding author upon reasonable request, for academic purposes.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	It is generally agreed that deep learning requires at least tens of thousands of examples. Because of resource constraints, we were only able to collect 904 cases containing 14,758 images. For the testing of the model, the calculated minimum sample size was 70 when the expected sensitivity/specificity was 0.9 and the margin of error was 0.05. We collected 200 cases for better estimation of the performance. We did not consider the prevalence of the condition in determining the sample size, as this study was focused on the technical ability of detecting brain haemorrhage. Additionally, we evaluated the model on the 196 cases acquired from the 4-month period after we developed the model to ensure that there was no bias in the test dataset.
Data exclusions	Exclusion criteria were pre-established as any history of brain surgery, skull fracture, intracranial tumor, intracranial device placement, or cerebral infarct. The exclusion criteria were determined on the basis of the intended clinical indication of the developed application.
Replication	Following the standard procedure in deep learning, we tested the algorithms using a distinct dataset that was not used for training and validation.
Randomization	Samples were randomly allocated to training, test and validation datasets.
Blinding	The radiologists evaluated for performance comparison were blind to the clinical information of the test dataset. The developers were blinded to clinical information when running the algorithms on the test dataset. The radiologists who annotated haemorrhagic foci were blinded to the attention maps. The radiologists who evaluated the prediction-bases were blinded to the model-determined prediction-bases.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Head CT images were obtained from an adult population who visited the emergency department of a tertiary urban teaching institution in the U.S. between 2005 and 2017.
Recruitment	The institutional review board exempted informed consent on the basis of minimal harm to the study participants. We collected radiographic images to be included in the study by using predetermined criteria.