Published July 26, 2018 as 10.3174/ajnr.A5742

Hybrid 3D/2D Convolutional Neural Network for Hemorrhage Evaluation on Head CT

P.D. Chang, E. Kuoy, J. Grinband, B.D. Weinberg, M. Thompson, R. Homo, J. Chen, H. Abcede, M. Shafie, L. Sugrue, C.G. Filippi, M.-Y. Su, W. Yu, C. Hess, and D. Chow

ABSTRACT

BACKGROUND AND PURPOSE: Convolutional neural networks are a powerful technology for image recognition. This study evaluates a convolutional neural network optimized for the detection and quantification of intraparenchymal, epidural/subdural, and subarachnoid hemorrhages on noncontrast CT.

MATERIALS AND METHODS: This study was performed in 2 phases. First, a training cohort of all NCCTs acquired at a single institution between January 1, 2017, and July 31, 2017, was used to develop and cross-validate a custom hybrid 3D/2D mask ROI-based convolutional neural network architecture for hemorrhage evaluation. Second, the trained network was applied prospectively to all NCCTs ordered from the emergency department between February 1, 2018, and February 28, 2018, in an automated inference pipeline. Hemorrhage-detection accuracy, area under the curve, sensitivity, specificity, positive predictive value, and negative predictive value were assessed for full and balanced datasets and were further stratified by hemorrhage type and size. Quantification was assessed by the Dice score coefficient and the Pearson correlation.

RESULTS: A 10,159-examination training cohort (512,598 images; 901/8.1% hemorrhages) and an 862-examination test cohort (23,668 images; 82/12% hemorrhages) were used in this study. Accuracy, area under the curve, sensitivity, specificity, positive predictive value, and negative-predictive value for hemorrhage detection were 0.975, 0.983, 0.971, 0.975, 0.793, and 0.997 on training cohort cross-validation and 0.970, 0.981, 0.951, 0.973, 0.829, and 0.993 for the prospective test set. Dice scores for intraparenchymal hemorrhage, epidural/subdural hemorrhage, and SAH were 0.931, 0.863, and 0.772, respectively.

CONCLUSIONS: A customized deep learning tool is accurate in the detection and quantification of hemorrhage on NCCT. Demonstrated high performance on prospective NCCTs ordered from the emergency department suggests the clinical viability of the proposed deep learning tool.

ABBREVIATIONS: CNN = convolutional neural networks; EDH/SDH = epidural/subdural hemorrhage; GPU = graphics processing unit; ICH = intracranial hemorrhage; IPH = intraparenchymal hemorrhage; mask R-CNN = mask ROI-based CNN

Intracranial hemorrhages (ICHs) represent a critical medical event that results in 40% patient mortality despite aggressive care.[1] Early and accurate diagnosis is necessary for the management of acute ICHs.[2,3] However, increasing imaging use and distractions from noninterpretive tasks are known to cause delays in diagnosis[4] with turn-around time for noncontrast CT head examinations reported up to 1.5–4 hours in the emergency department.[4] These delays impact patient care because acute deterioration from hemorrhage expansion often results early, within the initial 3–4.5 hours of symptom onset.[5-7] Therefore, a tool for expeditious and accurate diagnosis of ICHs may facilitate a prompt therapeutic response and ultimately improved outcomes.

In addition to ICH detection, a tool for automated quantification of hemorrhage volume may provide a useful metric for patient monitoring and prognostication.[8,9] For intraparenchymal hemorrhage (IPH) specifically, the current clinical standard for quantification relies on a simplified formula (ABC/2) calculation

Received April 10, 2018; accepted after revision June 6.

From the Departments of Radiology (P.D.C., E.K., M.T., R.H., M.-Y.S., D.C.), Neurosurgery (J.C.), and Neurology (H.A., M.S., W.Y.), University of California Irvine; Departments of Radiology (P.D.C., L.S., C.H.), University of California, San Francisco, California; Department of Radiology (J.G.), Columbia University, New York, New York; Department of Radiology (B.D.W.), Emory University School of Medicine, Atlanta, Georgia; and Department of Radiology (C.G.F.), North Shore University Hospital, Long Island, New York.

The work of P.C. was, in part, supported by grant T32EB001631 from the National Institutes of Health (National Institute of Biomedical Imaging and Bioengineering). The work of D.C. was, in part, supported by Canon Medical Systems.

Please address correspondence to Daniel Chow, MD, University of California, Irvine Medical Center, 101 The City Drive South, Douglas Hospital, Route 140, Room 0115, Orange, CA 92868-3201; e-mail: chowd3@uci.edu; @TheCAIDM

Indicates open access to non-subscribers at www.ajnr.org
http://dx.doi.org/10.3174/ajnr.A5742

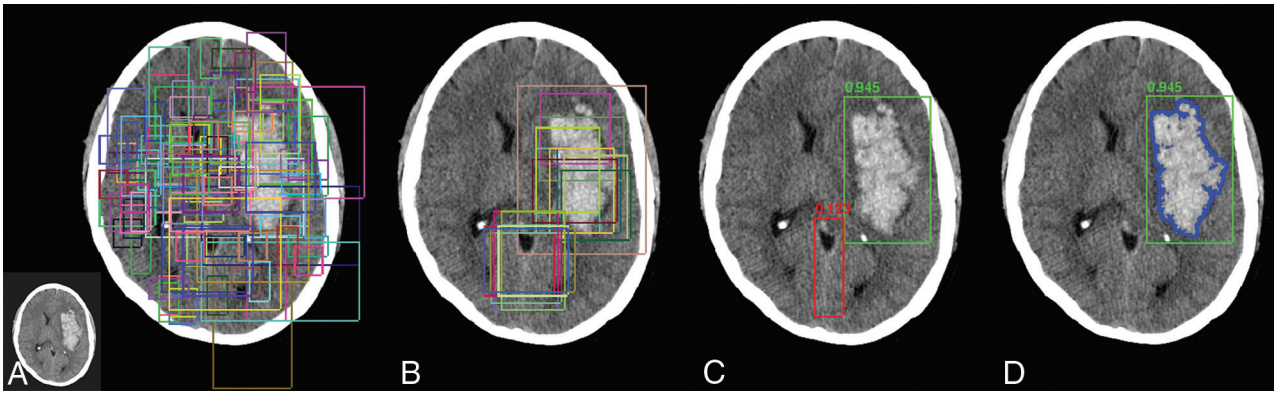

FIG 1. Overview of the mask R-CNN approach. Mask R-CNN architectures provide a flexible and efficient framework for parallel evaluation of region proposal (attention), object detection (classification), and instance segmentation. *A*, Preconfigured bounding boxes at various shapes and resolutions are tested for the presence of a potential abnormality. *B*, The highest ranking bounding boxes are identified and used to generate region proposals that focus algorithm attention. *C*, Composite region proposals are pruned using nonmaximum suppression and are used as input into a classifier to determine the presence or absence of hemorrhage. *D*, Segmentation masks are generated for cases positive for hemorrhage.

that commonly overestimates true IPH volumes by up to 30%.[10] Alternatively, while manual delineation of hemorrhage may provide accurate volume estimates, time constraints make this impractical in the emergency setting. Accordingly, a fully automated and objective tool for rapid quantification of ICH volume may be a compelling alternative to current approaches, offering more accurate, detailed information to guide clinical decision-making.

In this study, we propose a tool based on deep learning convolutional neural networks (CNN), an emerging technology now capable of image interpretation tasks that were once thought to require human intelligence.[11] The effectiveness of CNNs is based on the capacity of the algorithm for self-organization and pattern recognition without explicit human programming. Using a deep learning approach, Prevedello et al[12] previously described a generic algorithm for broad screening of various acute NCCT findings (hemorrhage, mass effect, hydrocephalus) with an overall sensitivity and specificity of 90% and 85%, respectively. We extend this preliminary work by customizing a new mask ROI-based CNN (mask R-CNN) architecture optimized specifically for ICH evaluation and training the network on an expanded cohort of NCCT head examinations. In addition to validation on a retrospective cohort, the trained algorithm will be tested for real-time interpretation of new, prospectively acquired NCCT examinations as part of an automated inference pipeline. By testing performance in a realistic environment of consecutive NCCT examinations, we hope to assess the feasibility of future implementation in clinical practice.

In summary, the 3 key objectives of this study include deep learning algorithm development and assessment of final trained CNN performance in the following: 1) detection of ICH including intraparenchymal, epidural/subdural (EDH/SDH), and subarachnoid hemorrhages; 2) quantification of ICH volume; and 3) prospective, real-time inference on an independent test set as part of an automated pipeline.

MATERIALS AND METHODS

Patient Selection

After approval of the institutional review board of the University of California, Irvine Medical Center, 2 separate cohorts were identified for this study: one cohort for training (combined with cross-validation) and a second cohort as an independent test set. The initial retrospectively defined training cohort consisted of every NCCT examination acquired at the study institution between January 1, 2017, and July 31, 2017. The subsequent prospectively acquired independent test set cohort consisted of every NCCT examination ordered from the emergency department between February 1, 2018, and February 28, 2018. For both cohorts, cases positive for hemorrhage (IPH, EDH/SDH, and SAH) were identified from clinical reports and confirmed with visual inspection by a board-certified radiologist. 3D ground truth masks were generated for all cases positive for hemorrhage using a custom semiautomated Web-based annotation platform developed at our institution, implementing a variety of tools for level-set segmentation and morphologic operations. All masks were visually inspected for accuracy by a board-certified radiologist.

Convolutional Neural Network

A custom architecture derived from the mask R-CNN algorithm was developed for detection and segmentation of hemorrhage.[13] In brief, the mask R-CNN architecture provides a flexible and efficient framework for parallel evaluation of region proposal (attention), object detection (classification), and instance segmentation (Fig 1). In the first step, a preconfigured distribution of bounding boxes at various shapes and resolutions is tested for the presence of a potential abnormality. Next, the highest ranking bounding boxes are identified and used to generate region proposals, thus focusing algorithm attention on specific regions of the image. These composite region proposals are pruned using nonmaximum suppression and are used as input into a classifier to determine the presence or absence of hemorrhage. In the case of detection positive for hemorrhage, a final segmentation branch of the network is used to generate binary masks.

The efficiency of a mask R-CNN architecture arises from a common backbone network that generates a shared set of image features for the various parallel detection, classification, and segmentation tasks (Fig 2). The backbone network used in this article is a custom hybrid 3D/2D variant of the feature pyramid network.[14] This custom backbone network was constructed using standard residual bottleneck blocks[15] without iterative tuning,

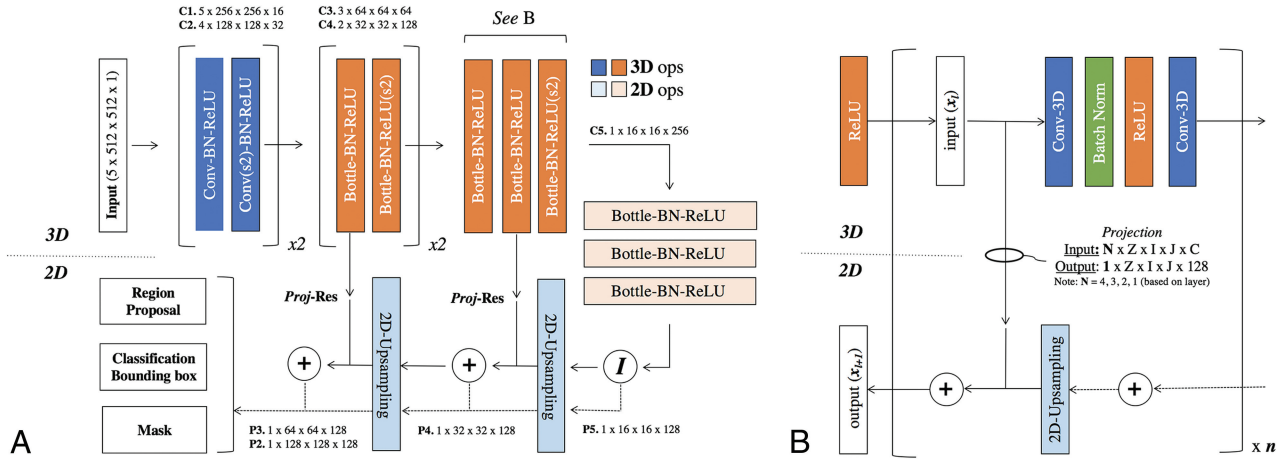

FIG 2. Convolutional neural network architecture. *A*, Hybrid 3D-contracting (bottom-up) and 2D-expanding (top-down) fully convolutional feature-pyramid network architecture used for the mask R-CNN backbone. The architecture incorporates both traditional 3 × 3 filters (blue) as well as bottleneck 1 × 1–3 × 3–1 × 1 modules (orange). The contracting arm is composed of 3D operations and convolutional kernels. Subsampling in the x- and y-directions is implemented via 1 × 2 × 2 strided convolutions (marked by s2). Subsampling in the *z*-direction is mediated by a 2 × 1 x 1 convolutional kernel with valid padding. The expanding arm is composed entirely of 2D operations. *B*, Connections between the contracting and expanding arms are facilitated by residual addition operations between corresponding layers. 3D layers in the contracting arm are mapped to 2D layers in the expanding arm by projection operations, which are designed both to match in the input (N) and output (I) *z*-dimension shape in addition to input (C) and output (128) feature map sizes. Ops indicates operations; Conv, convolutions; BN-ReLU, Batch Normalization Rectified Linear Unit; Proj-Res, Projection-Residual; Z, Z-axis; I, In plane axis; J, In plane axis.

given the observation that mask R-CNN architectures, particularly those based on pyramid networks, are robust to many design choices. In this implementation, a 3D input matrix of 5 × 512 × 512 is mapped to 2D output feature maps at various resolutions, with 3D input from the pyramid network bottom-up pathway added to the 2D feature maps of the top-down pathway using a projection operation to match the matrix dimensions. Thus, the network can use contextual information from the 5 slices immediately surrounding the ROI to predict the presence and location of hemorrhage.

Implementation

The approximate joint training method as described in the original faster mask R-CNN implementation[16] was used for parallel optimization of the region-proposal network classifier and segmentation heads. The mask R-CNN architecture was trained using 128 sampled ROIs per image, with a ratio of positive-to-negative samples fixed at 1:3. During inference, the top 256 proposals by the region-proposal network are pruned using nonmaximum suppression and are used to generate detection boxes for classification. The region-proposal network anchors span 4 scales (128 × 128, 64 × 64, 32 × 32, 16 × 16) and 3 aspect ratios (1:1, 1:2, 2:1).

Network weights were initialized using the heuristic described by He et al.[17] The final loss function included a term for L2 regularization of the network parameters. Optimization was implemented using the Adam method, an algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower order moments.[18] An initial learning rate of 2×10^{-4} was used and annealed whenever a plateau in training loss was observed.

The software code for this study was written in Python 3.5 using the open-source TensorFlow r1.4 library (Apache 2.0 license; https://github.com/tensorflow/tensorflow/blob/master/LICENSE).[19] Experiments were performed on a graphics processing unit (GPU)-optimized workstation with 4 GeForce GTX Titan X cards (12GB, Maxwell architecture; NVIDIA, Santa Clara, California). Inference benchmarks for speed were determined using a single-GPU configuration.

Image Preprocessing

For each volume, the axial soft-tissue reconstruction series was automatically identified by a custom CNN-based algorithm. If necessary, this volume was resized to an in-plane resolution matrix of 512 × 512. Furthermore, all matrix values less than −240 HU or greater than +240 HU were clipped, and the entire volume was rescaled to a range of [−3, 3].

Statistical Analysis

The primary end point of this study was the detection of hemorrhage on a per-study basis. A given NCCT volume was considered positive for hemorrhage if any single region-proposal prediction on any given slice was determined to contain hemorrhage. Thus, algorithm performance including accuracy, sensitivity, specificity, positive predictive value, and negative predictive value was calculated. Furthermore, by varying the softmax score threshold for hemorrhage classification, we calculated an area under the curve.

In addition to complete dataset evaluation, performance statistics on a balanced dataset (an equal number of positive and negative cases) were also calculated. By means of a balanced distribution, accuracy could also be further stratified by hemorrhage type (IPH, EDH/SDH, and SAH) and size (punctate, small, medium, and large, defined as <0.01, 0.01–5.0, 5.0–25, and >25 mL).

The secondary end point of this study was the ability of the algorithm to accurately estimate hemorrhage volume. This was assessed in 2 ways. First, predicted binary masks of hemorrhage

Table 1: Distribution of hemorrhages by type and size[a]

	IPH		EDH/SDH		SAH	
Size	Valid	Test	Valid	Test	Valid	Test
Large	192	13	188	19	85	9
Medium	88	8	79	15	53	3
Small	63	1	49	4	52	6
Punctate	15	1	3	0	34	3
Total	358	23	319	38	224	21

[a] Large, medium, small, and punctate hemorrhages were defined as >25, 5–25, 0.01–5.0, and <0.01 mL, respectively.

were compared with criterion standard manual segmentations using a Dice score coefficient. Second, predicted volumes of hemorrhage were compared with criterion standard annotated volumes using a Pearson correlation coefficient (r). As a comparison, estimates of IPH volume were also calculated using the simplified ABC/2 formula.

Training Cohort Evaluation

A 5-fold cross-validation scheme was used for evaluation of the initial training cohort. In this experimental paradigm, 80% of the data are randomly assigned into the training cohort, while the remaining 20% are used for validation. This process is then repeated 5 times until each study in the entire dataset is used for validation once. Validation results below are reported for the cumulative statistics across the entire dataset.

Independent Test Cohort Evaluation

After fine-tuning the algorithm design and parameters, we applied the final trained network to a new, prospective cohort of all consecutive NCCT examinations ordered from the emergency department for 1 month. The entire pipeline for inference was fully automated, including real-time transfer of newly acquired examinations to a custom GPU server from the PACS, identification of the correct input series, and trained network inference. In addition to initial validation statistics, results from this independent test dataset are also reported.

RESULTS

Patient Selection

The initial training set cohort comprised 10,159 NCCT examinations, 901 (8.9%) of which contained hemorrhage including IPH (n = 358/10,159, 3.5%), EDH/SDH (n = 319, 3.1%), and SAH (n = 224, 2.2%), yielding a total of 512,598 images. The median hemorrhage size was 28.2 mL (interquartile range, 9.4–44.7 mL).

The independent test set cohort compromised 682 prospective NCCT examinations, 82 (12.0%) of which contained hemorrhage including IPH (n = 23, 3.4%), EDH/SDH (n = 38, 5.6%), and SAH (n = 21, 3.1%), yielding 23,668 images. The median hemorrhage size was 24.9 mL (interquartile range, 8.3–35.6 mL). Further baseline stratification of both training and test set cohorts by hemorrhage type and size can be found in Table 1.

ICH Detection

Overall algorithm performance on the full dataset as measured by accuracy, area under the curve, sensitivity, specificity, positive predictive value, and negative predictive value was 0.975, 0.983, 0.971, 0.975, 0.793, and 0.997 for the cross-validation cohort and 0.970, 0.981, 0.951, 0.973, 0.829, and 0.993 for the prospective test set. When stratified by ICH type, the sensitivity for IPH, EDH/SDH, and SAH detection was 98.6% (353/358), 97.4% (311/319), and 94.2% (211/224) for the cross-validation cohort and 100% (23/23), 94.7% (36/38), and 90.5% (19/21) for the prospective test set. In total, 26/901 (2.9%) hemorrhages were missed in the cross-validation cohort compared with 4/81 (4.9%) hemorrhages in the prospective test set (Figs 3 and 4).

Balanced dataset results stratified by hemorrhage size show that in general, algorithm accuracy for hemorrhages of >5 mL (range, 0.977–0.999 mL) is higher than for hemorrhages of <5 mL (range, 0.872–0.965 mL) with only 4 cases of missed hemorrhage of >5 mL across both cohorts (all representing EDH/SDH). Detection accuracy of punctate hemorrhages of <0.01 mL (range, 0.872–0.883 mL) is noticeably more challenging than that of small hemorrhages between 0.01 and 5 mL (range, 0.906–0.965 mL). When we further stratify results by hemorrhage type, the most challenging combinations to detect are punctate SAH or EDH/SDH with accuracy ranges of 0.830–0.881 across both cohorts. Complete stratification of balanced dataset results by hemorrhage and size can be found in Table 2.

ICH Quantification

Estimates of IPH, EDH/SDH, and SAH segmentation masks by the CNN demonstrated Dice score coefficients of 0.931, 0.863, and 0.772, respectively, compared with manual segmentations. Estimates of IPH, EDH/SDH, and SAH volume by the CNN demonstrated Pearson correlation coefficients of 0.999, 0.987, and 0.953 compared with volumes derived from manual segmentations. By comparison, estimates of IPH volume derived from the simplified ABC/2 formula demonstrated a Pearson correlation of 0.954. On average, the ABC/2-derived hemorrhage volumes overestimated ground truth by an average of 20.2%, while the CNN-derived hemorrhage volumes underestimated ground truth by an average of just 2.1%.

Network Statistics

Each network for a corresponding validation fold trained for approximately 100,000 iterations before convergence. Depending on the number of GPU cards for training distribution, this process required, on average, 6–12 hours per fold. Once trained, the mask R-CNN network was able to determine the presence of hemorrhage in a new test case within an average of 0.121 seconds, including all preprocessing steps on a single GPU workstation.

DISCUSSION

In this study, we demonstrate that a deep learning solution is highly accurate in the detection of ICHs, including IPHs, EDHs/SDHs, and SAHs. In addition, this study demonstrates that a CNN can quantify ICH volume with high accuracy as reflected by Dice score coefficients (0.772–0.931) and Pearson correlations (0.953–0.999). Finally, while embedded for 1 month in an automated inference pipeline, the deep learning tool was able to accurately detect and quantify ICHs from prospective NCCT examinations ordered from the emergency department.

There are several previously described approaches to ICH detection with traditional machine-learning techniques such as fuzzy clustering,[20,21] Bayesian classification,[22] level-set thresh-

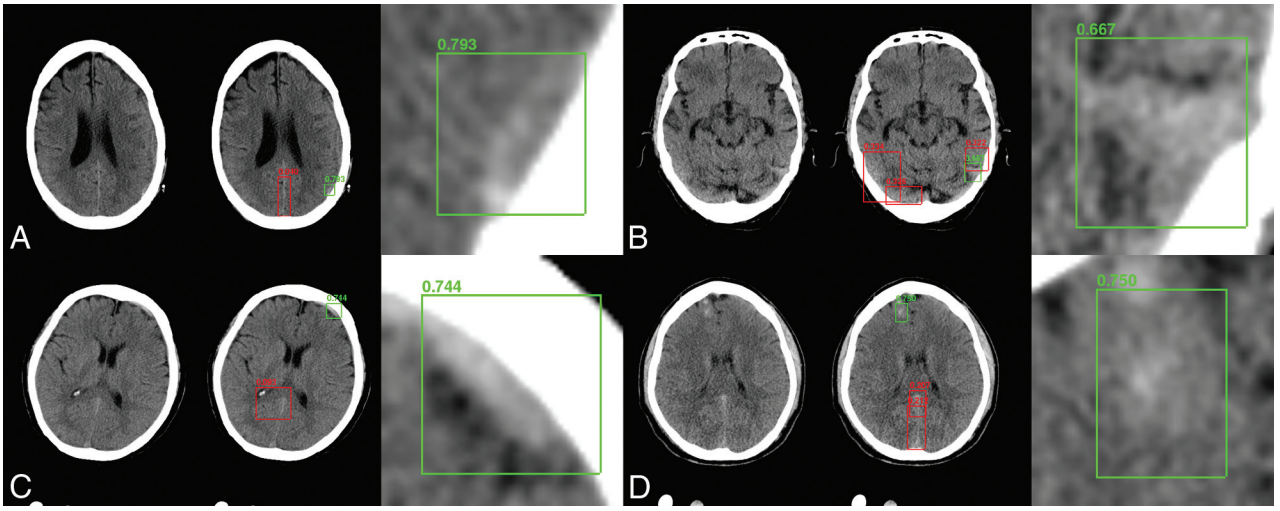

FIG 3. Sample network predictions: true-positives. Network predictions by the algorithm include bounding-box region proposals for potential areas of abnormality (to focus algorithm attention) and final network predictions, including confidence of results. Correctly identified areas of hemorrhage (green) include subtle abnormalities representing subarachnoid (*A*), subdural (*B* and *C*), and intraparenchymal (*D*) hemorrhage. Correctly identified areas of excluded hemorrhage often include common mimics for blood on NCCT, including thickening/high density along the falx (*A*, *C*, and *D*) and beam-hardening along the periphery (*B*).

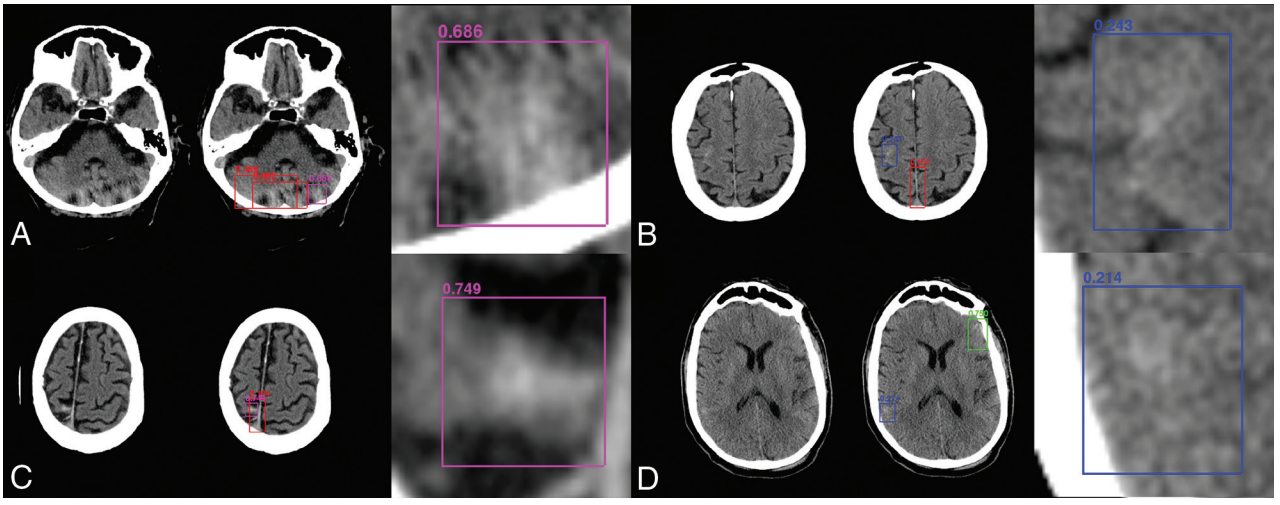

FIG 4. Sample network predictions: false-positives and false-negatives. Network predictions by the algorithm include bounding-box region proposals for potential areas of abnormality (to focus algorithm attention) and final network predictions including confidence of results. False-positive predictions for hemorrhage (purple) often include areas of motion artifacts and/or posterior fossa beam-hardening (*A*) or high-density mimics such as cortical calcification (*C*). False-negative predictions for excluded hemorrhage often include small volume abnormalities with relatively lower density, resulting in decreased conspicuity. Examples include subtle subarachnoid hemorrhage along the posterior right frontal lobe (*B*) and right inferior parietal lobe (*D*).

olds,[23] and decision tree analysis.[24] However, the image diversity present on any given NCCT head examination ultimately limits the accuracy of algorithms that are derived from a priori rules and hard-coded assumptions. For example, Gong et al[24] reported a sensitivity of 0.60 and a positive predictive value of 0.447 for IPH detection using decision tree analysis. Furthermore, hard-coded logic tends to produce narrow algorithms optimized for just a single task. For example, Prakash et al[23] reported a level-set technique for hemorrhage quantification yielding a Dice score range between 0.858 and 0.917; however, the algorithm is limited for hemorrhage detection because it is not designed to exclude hemorrhage on an examination with negative findings.

Given the increasing awareness of deep learning potential in medical imaging, there has been a gradual paradigm shift increasingly favoring convolutional neural networks over other approaches. For example, Shen et al[25] developed a multiscale CNN for lung nodule detection with CT images, while Wang et al[26] devised a 12-layer CNN for predicting cardiovascular disease from mammograms as well as for detecting spine metastasis.[27] More recently, Phong et al[28] described a deep learning approach for hemorrhage detection using several pretrained networks on a small test set of 20 cases.

However, while this preliminary effort is important, there are several key limitations to be addressed before clinical deployment of deep learning tools. First, in addition to high algorithm performance, a clinically viable tool must address the traditional "black box" critique of being unable to rationalize a given interpretation. While there are some techniques to ameliorate this through gen-

Table 2: Balanced dataset performance statistics stratified by hemorrhage type and size[a]

	Accuracy		AUC		Sensitivity		Specificity		PPV		NPV	
Size	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test	Valid	Test
All ICHs	0.984	0.972	0.991	0.989	0.971	0.951	0.975	0.973	0.975	0.972	0.971	0.952
Large	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Medium	0.992	0.977	0.995	0.982	0.986	0.962	0.975	0.973	0.975	0.972	0.986	0.962
Small	0.965	0.906	0.972	0.987	0.933	0.818	0.975	0.973	0.974	0.968	0.936	0.843
Punctate	0.883	0.872	0.895	0.903	0.769	0.750	0.975	0.973	0.968	0.965	0.809	0.796
IPH	0.992	0.997	0.996	0.999	0.986	1.000	0.975	0.973	0.975	0.973	0.986	1.000
Large	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Medium	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Small	0.983	0.997	0.999	0.999	0.968	1.000	0.975	0.973	0.974	0.973	0.968	1.000
Punctate	0.899	0.997	0.921	0.999	0.800	1.000	0.975	0.973	0.969	0.973	0.830	1.000
EDH/SDH	0.986	0.970	0.989	0.974	0.975	0.947	0.975	0.973	0.975	0.972	0.975	0.949
Large	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Medium	0.980	0.963	0.983	0.971	0.962	0.933	0.975	0.973	0.974	0.971	0.963	0.936
Small	0.958	0.872	0.968	0.882	0.918	0.750	0.975	0.973	0.973	0.965	0.923	0.796
Punctate	0.832	NA	0.857	NA	0.667	NA	0.975	0.973	0.963	NA	0.745	NA
SAH	0.970	0.949	0.972	0.953	0.942	0.905	0.975	0.973	0.974	0.971	0.944	0.911
Large	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Medium	0.999	0.997	0.999	0.999	1.000	1.000	0.975	0.973	0.975	0.973	1.000	1.000
Small	0.950	0.913	0.960	0.928	0.904	0.833	0.975	0.973	0.973	0.968	0.910	0.854
Punctate	0.881	0.830	0.891	0.833	0.765	0.667	0.975	0.973	0.968	0.961	0.806	0.745

Note:—AUC indicates area under the curve; NA, not applicable; PPV, positive predictive value; NPV, negative predictive value.

[a] Large, medium, small, and punctate hemorrhages were defined as >25, 5–25, 0.01–5.0, and <0.01 mL, respectively.

eration of saliency maps[29] or class-activation maps,[30] this is a known limitation of conventional global CNN-based classification of an image (or volume). By contrast, the proposed custom mask R-CNN architecture, through combining an attention-based object-detection network with more traditional classification and segmentation components, allows the algorithm to explicitly localize suspicious CT findings and provide visual feedback regarding which findings are likely to represent ICH or a mimic.

Second, a clinically viable tool needs to be tested on unfiltered data in a setting that reflects the expected context for deployment. In this study, we attempted to simulate this by deploying the trained network in a fully automated inference pipeline that can perform all the requisite steps to support algorithm prediction, ranging from PACS image transfer to series identification to GPU-enabled inference, all without human supervision. Furthermore, the prospectively acquired, independent test set used in this context is a reflective sample of the target population used, namely every NCCT head examination performed in the emergency radiology department. That algorithm performance in this setting remains favorable suggests that the deep learning tool has promising potential for clinical utility in the near future.

An additional point should also be made of the requisite data base size for proper algorithm validation. While large datasets are rare in medical imaging, a representative sample of pathology is critical for validating algorithm accuracy. As evidenced in this study, it is often the uncommon findings for which a neural network has the most difficultly learning and generalizing to (eg, punctate hemorrhages of <0.01 mL represent approximately 56/10,841 = 0.5% of all examinations yet are also the most difficult to detect); thus, a large representative dataset is required to assess performance on these critical rare entities. A large data base also facilitates algorithm learning, whereby the increased diversity of training examples helps the network choose more generalizable and predictive features. Finally, cases without ICH are just as important as those with ICH because the algorithm must also be able to correctly identify the absence of hemorrhage in most cases despite any possible underlying pathology that may be present. To address these issues, this study takes advantage of a large training dataset comprising over 512,598 images from >10,000 patients, at least an order of magnitude higher than that in any previous study.

The most salient use case of an accurate tool for hemorrhage detection is a triage system that alerts physicians of examinations potentially positive for hemorrhage for expedited interpretation, thus facilitating reduced turn-around time. The recent 2013 Imaging Performance Partnership survey of >80 institutions rated the importance of reduced turn-around time as one of their highest priorities, scoring 5.7 of a 6.0 rating,[31] allowing an expedited triage of patients for therapeutic management. As an example, rapid identification of patients with IPH would facilitate immediate control of blood pressure during the vulnerable first few 3–4.5 hours of symptom onset when acute deterioration is most likely.[5-7] The importance of rapid diagnosis is supported further by the recent Intensive Blood Pressure Reduction in Acute Cerebral Hemorrhage Trial-2, which concluded that intensive treatment afforded by early diagnosis was associated with improved functional outcome.[32]

In addition to hemorrhage detection, ICH volume metrics can be used to precisely and efficiently quantify the initial burden of disease as well as serial changes, which, in turn, may have important clinical implications.[33,34] For IPHs, this is most relevant within the first 2–3 hours of onset when the hemorrhagic volume can shift dramatically.[5-7] Furthermore, the volume of hemorrhage is a known predictor of 30-day mortality and morbidity.[8,9] Presently, the clinical standard for estimation of IPH volume is by the ABC/2 formula of Kwak et al,[10,35] in which A and B represent maximum single-dimensional perpendicular measurements on the largest axial region of hemorrhage and C represents a graded estimate of the craniocaudal extent. While easy to use, this limited

approach assumes an ellipsoid shape for all IPHs. In this study, we show that this assumption results in overestimation of hemorrhage by 20.2%, a statistic that has been previously reported with discrepancies up to 30% compared with manual segmentation.[10] While the criterion standard remains manual delineation, this approach can be both time-consuming and technically challenging in the emergency department setting. By comparison, the ability of the trained CNN to rapidly and accurately quantify IPH volume with >0.999 correlations to human experts offers a clinically feasible, improved alternative to the current standards of practice.

Several limitations should be addressed when considering our results. First, examinations in this study were performed at a single academic institution. Therefore, while we have demonstrated that our results generalize well to independent datasets obtained at our hospital center, further work is necessary to evaluate performance on a variety of vendors and scanning protocols at other institutions. While we acknowledge this drawback, CT examinations are inherently normalized by Hounsfield Units and show less image variability than plain radiographs or MR imaging. Second, deep learning algorithms are known to be susceptible to the phenomenon of adversarial noise,[36] where small but highly patterned perturbations in images may result in unexpected predictions. However, this is rare and was not encountered in the current dataset and, to some extent, can be mitigated using network ensembles and denoising autoencoders.[37] Finally, while the current dataset is quite large, there are, nonetheless, rare findings and contexts that occur at a prevalence of less than our 1/10,000 cases, and it is foreseeable that such studies may be incorrectly interpreted. To this end, we plan to incorporate continued iterative algorithm updates as new, increasingly larger datasets become available.

CONCLUSIONS

This study demonstrates the high performance of a fully automated, deep learning algorithm for detection and quantification of IPH, EDH/SDH, and SAH on NCCT examinations of the head. Furthermore, confirmation of high algorithm performance on a prospectively acquired, independent test set while embedded in an automated inference environment suggests the clinical viability of this deep learning tool in the near future. Such a tool may be implemented either as a triage system to assist radiologists in identifying high-priority examinations for interpretation and/or as a method for rapid quantification of ICH volume, overall expediting the triage of patient care and offering more accurate, detailed information to guide clinical decision-making.

Disclosures: Peter Chang—*RELATED*: *Grant*: National Institutes of Health (National Institute of Biomedical Imaging and Bioengineering) T32 Training Grant T32EB001631.* Christopher G. Filippi—*UNRELATED*: *Consultancy*: Syntactx, *Comments*: I read MR imaging brain scans for their clinical trials; *Grants/Grants Pending*: The Foundation of the American Society of Neuroradiology (FASNR), *Comments*: Alzheimer research study grant*; *Payment for Lectures Including Service on Speakers Bureaus*: Grand Rounds Lectures: Montefiore, Albert Einstein College of Medicine, and Iowa University. Daniel Chow—*RELATED*: *Grant*: Canon Medical. *Money paid to the institution.

REFERENCES

1. van Asch CJ, Luitse MJ, Rinkel GJ, et al. **Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis.** *Lancet Neurol* 2010;9:167–76 CrossRef Medline
2. Goldstein JN, Gilson AJ. **Critical care management of acute intracerebral hemorrhage.** *Curr Treat Options Neurol* 2011;13: 204–16 CrossRef Medline
3. Heit JJ, Iv M, Wintermark M. **Imaging of intracranial hemorrhage.** *J Stroke* 2017;19:11–27 CrossRef Medline
4. Glover M 4th, Almeida RR, Schaefer PW, et al. **Quantifying the impact of noninterpretive tasks on radiology report turn-around times.** *J Am Coll Radiol* 2017;14:1498–1503 CrossRef Medline
5. Davis SM, Broderick J, Hennerici M, et al. Recombinant Activated Factor VII Intracerebral Hemorrhage Trial Investigators. **Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage.** *Neurology* 2006;66:1175–81 CrossRef Medline
6. Kazui S, Naritomi H, Yamamoto H, et al. **Enlargement of spontaneous intracerebral hemorrhage: incidence and time course.** *Stroke* 1996;27:1783–87 CrossRef Medline
7. Qureshi A, Palesch Y, ATACH II Investigators. **Expansion of recruitment time window in antihypertensive treatment of acute cerebral hemorrhage (ATACH) II trial.** *J Vasc Interv Neurol* 2012; 5:6–9 Medline
8. Broderick JP, Brott TG, Duldner JE, et al. **Volume of intracerebral hemorrhage: a powerful and easy-to-use predictor of 30-day mortality.** *Stroke* 1993;24:987–93 CrossRef Medline
9. Butcher K, Laidlaw J. **Current intracerebral haemorrhage management.** *J Clin Neurosci* 2003;10:158–67 CrossRef Medline
10. Scherer M, Cordes J, Younsi A, et al. **Development and validation of an automatic segmentation algorithm for quantification of intracerebral hemorrhage.** *Stroke* 2016;47:2776–82 CrossRef Medline
11. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge: MIT Press; November 2016. ISBN: 9780262035613
12. Prevedello LM, Erdal BS, Ryu JL, et al. **Automated critical test findings identification and online notification system using artificial intelligence in imaging.** *Radiology* 2017;285:923–31 CrossRef Medline
13. He K, Gkioxari G, Dollár P, et al. **Mask R-CNN.** arXiv:1703.06870. 2017. In: *Proceedings of the IEEE International Conference on Computer Vision,* Venice, Italy. October 22–29, 2017
14. Lin TY, Dollár P, Girshick R, et al. **Feature pyramid networks for object detection.** arXiv:1612.03144. 2017. In: *Proceedings of the IEEE* Conference on *Computer Vision and Pattern Recognition,* Honolulu, Hawaii. July 21–27, 2017
15. He K, Zhang X, Ren S, et al. **Deep residual learning for image recognition.** arXiv:1512.03385. 2016. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vagas, Nevada. June 27–30, 2016
16. Ren S, He K, Girshick R, et al. **Faster R-CNN: towards real-time object detection with region proposal networks.** *IEEE Trans Pattern Anal Mach Intell* 2017;39:1137–49 CrossRef Medline
17. He K, Zhang X, Ren S, et al. **Delving deep into rectifiers: surpassing human-level performance on imagenet classification.** arXiv:1502. 01852. 2015. In: *Proceedings of the IEEE International Conference on Computer Vision,* Santiago, Chile. December 7–13, 2015:1026–34
18. Kingma DP, Ba J Adam. **A method for stochastic optimization.** arXiv:1412.6980. 2015. In: *Proceedings of the International Conference for Learning Representations*, San Diego, California. May 7–9, 2015
19. Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. http://download.tensorflow.org/paper/whitepaper2015.pdf. Accessed March 25, 2018
20. Yuh EL, Gean AD, Manley GT, et al. **Computer-aided assessment of head computed tomography (CT) studies in patients with suspected traumatic brain injury.** *J Neurotrauma* 2008;25:1163–72 CrossRef Medline
21. Cósić D, Lončarić S. **Rule-based labeling of CT head image.** In: Keravnou E, Garbay C, Baud R, et al, eds. *Artificial Intelligence in Medicine: 6th Conference on Artificial Intelligence in Medicine Europe, AIME'97 Grenoble, France, March 23–26, 1997 Proceedings.* Berlin: Springer-Verlag;1997:453–56

22. Li YH, Zhang L, Hu QM, et al. **Automatic subarachnoid space segmentation and hemorrhage detection in clinical head CT scans.** *Int J Comput Assist Radiol Surg* 2012;7:507–16 CrossRef Medline
23. Prakash KN, Zhou S, Morgan TC, et al. **Segmentation and quantification of intra-ventricular/cerebral hemorrhage in CT scans by modified distance regularized level set evolution technique.** *Int J Comput Assist Radiol Surg* 2012;7:785–98 CrossRef Medline
24. Gong T, Liu R, Tan CL, et al. **Classification of CT brain images of head trauma.** In: Rajapakse JC, Schmidt B, Volkert G, eds. *Pattern Recognition in Bioinformatics: Second IAPR International Workshop, PRIB 2007, Singapore, October 1–2, 2007 Proceedings.* Berlin: Springer-Verlag; 2007:401–08
25. Shen W, Zhou M, Yang F, et al. **Multi-scale convolutional neural networks for lung nodule classification.** *Inf Process Med Imaging* 2015;24:588–99 Medline
26. Wang J, Ding H, Bidgoli FA, et al. **Detecting cardiovascular disease from mammograms with deep learning.** *IEEE Trans Med Imaging* 2017;36:1172–81 CrossRef Medline
27. Wang J, Fang Z, Lang N, et al. **A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks.** *Comput Biol Med* 2017;84:137–46 CrossRef Medline
28. Phong TD, Duong HN, Nguyen HT, et al. **Brain hemorrhage diagnosis by using deep learning.** In: *Proceedings of the International Conference on Machine Learning and Soft Computing,* Ho Chi Minh City, Vietnam. January 13–16, 2017:34–39
29. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. https://arxiv.org/abs/1312.6034. Accessed March 25, 2018
30. Selvaraju RR, Das A, Vedantam R, et al. Grad-CAM: why did you say that? visual explanations from deep networks via gradient-based localization. https://www.researchgate.net/publication/308964930_GradCAM_Why_did_you_say_that_Visual_Explanations_from_Deep_Networks_via_Gradient-based_Localization. Accessed March 25, 2018
31. Nataraj S. 2013 Imaging Turnaround Times Survey Results. 2014. https://www.advisory.com/research/imaging-performance-partnership/expert-insights/2014/2013-turnaround-times-survey-results. Accessed March 25, 2018
32. Anderson CS, Heeley E, Huang Y, et al; INTERACT2 Investigators. **Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage.** *N Engl J Med* 2013;368:2355–65 CrossRef Medline
33. Jung SW, Lee CY, Yim MB. **The relationship between subarachnoid hemorrhage volume and development of cerebral vasospasm.** *J Cerebrovasc Endovasc Neurosurg* 2012;14:186–91 CrossRef Medline
34. Bullock MR, Chesnut R, Ghajar J, et al; Surgical Management of Traumatic Brain Injury Author Group. **Surgical management of acute epidural hematomas.** *Neurosurgery* 2006;58:S7–S15; discussion Si-iv Medline
35. Kwak R, Kadoya S, Suzuki T. **Factors affecting the prognosis in thalamic hemorrhage.** *Stroke* 1983;14:493–500 CrossRef Medline
36. Goodfellow IJ, Shlens J, Szegedy C. **Explaining and harnessing adversarial examples.** In: *Proceedings of the International Conference for Learning Representations,* San Diego, California. May 7–9, 2015
37. Gu S, Rigazio L. **Towards deep neural network architectures robust to adversarial examples.** https://arxiv.org/abs/1412.5068. Accessed March 25, 2018