2018

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

P. D. Chang

E. Kuoy

J. Grinband

B. D. Weinberg

M. Thompson

See next page for additional authors

HHS Public Access

Author manuscript

AJNR Am J Neuroradiol. Author manuscript; available in PMC 2019 March 01.

Published in final edited form as:

AJNR Am J Neuroradiol. 2018 September ; 39(9): 1609–1616. doi:10.3174/ajnr.A5742.

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

Peter D. Chang, MD[1,2], **Edward Kuoy, MD**[2], **Jack Grinband, PhD**[3], **Brent D. Weinberg, MD, PhD**[4], **Matthew Thompson, MD**[2], **Richelle Homo, BS**[2], **Jefferson Chen, MD**[5], **Hermelinda Abcede, MD**[6], **Mohammad Shafie, MD**[6], **Leo Sugrue, MD**[1], **Christopher G. Filippi, MD**[7], **Min-Ying Su, PhD**[2], **Wengui Yu, MD**[6], **Christopher Hess, MD**[1], and **Daniel Chow, MD**[2]

[1]Department of Radiology, University of California, San Francisco, CA

[2]Department of Radiology, University of California, Irvine, CA

[3]Department of Radiology, Columbia University, New York, NY

[4]Department of Radiology, Emory University School of Medicine, Atlanta, GA

[5]Department of Neurosurgery, University of California, Irvine, CA

[6]Department of Neurology, University of California, Irvine, CA

[7]Department of Radiology, North Shore University Hospital, Long Island, NY

Abstract

Background—Convolutional neural networks (CNN) are a powerful technology for image recognition. This study evaluates a CNN optimized for the detection and quantification of intraparenchymal (IPH), epidural/subdural (EDH/SDH) and subarachnoid (SAH) hemorrhages on non-contrast CT (NCCT).

Methods—This study was performed in two 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 R-CNN 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, AUC, sensitivity, specificity, PPV, and NPV was assessed for full and balanced datasets, and further stratified by hemorrhage type and size. Quantification was assessed by Dice score coefficient and Pearson correlation.

Results—A total of 10,159-exam training cohort (512,598 images; 901/8.1% hemorrhages) and 862-exam test cohort (23,668 images; 82/12% hemorrhags) were used in this study. Accuracy, area under the curve, sensitivity, specificity, PPV, and NPV for hemorrhage detection were 0.975, 0.983, 0.971, 0.975, 0.793, and 0.997 upon training cohort cross-validation, and 0.970, 0.981,

Corresponding author's contact: University of California, Irvine Medical Center, 101 The City Drive South, Douglas Hospital,, Route 140, Room 0115, Orange, CA 92868-3201, chowd3@uci.edu, Phone: 714-497-5735, Fax: 714-456-7864.
The name and street address of the institution from which the work originate: University of California, Irvine Medical Center, 101 The City Drive South, Douglas Hospital,, Route 140, Room 0115, Orange, CA 92868-3201

0.951, 0.973, 0.829, and 0.993 for the prospective test set. Dice scores for IPH, EDH/SDH, and SAH were 0.931, 0.863 and 0.772, respectively.

Conclusions—A customized deep learning tool is accurate in 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.

INTRODUCTION

Intracranial hemorrhages (ICHs) represent a significant 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 utilization coupled with distractions from noninterpretive tasks are known to cause delays in diagnosis[4] with turn-around-time (TAT) for non-contrast CT (NCCT) head examinations reported to be up to 1.5-4 hours in the emergency room setting[4]. These delays impact patient care as 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 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 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 overall sensitivity and specificity of 90% and 85% respectively. We extend this preliminary work by customizing a new mask region-of-interest-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 exams as part of an automated inference pipeline. By testing performance in a realistic environment of consecutive NCCT exams we hope to assess the feasibility of future implementation in clinical practice.

In summary, the three key objectives of this study include deep learning algorithm development and assessment of final trained CNN performance in: (1) detection of ICH including intraparenchymal, epidural/subdural (EDH/SDH), and subarachnoid (SAH)

hemorrhages; (2) quantification of ICH volume; (3) prospective, real-time inference on an independent test set as part of an automated pipeline.

METHODS

Patient Selection

After IRB approval, two 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 room between February 1, 2018 and February 28, 2018. For both cohorts, cases of positive 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 positive hemorrhage cases using a custom semi-automated 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 (Figure 1). In the first step, a preconfigured distribution of bounding boxes at various shapes and resolutions are 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 non-maximum suppression and used as input into a classifier to determine presence or absence of hemorrhage. In the case of positive hemorrhage detection, 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 (Figure 2). The backbone network used in this paper is a custom hybrid 3D/2D variant of the feature pyramid network (FPN)[14]. This custom backbone network was constructed using standard residual bottleneck blocks[15] without iterative tuning given the observation that R-CNN architectures, particular those based on FPNs, are robust to many design choices. In this implementation, a 3D input matrix of size $5 \times 512 \times 512$ is mapped to 2D output feature maps at various resolutions, with 3D inputs from the FPN bottom-up pathway added to the 2D feature maps of the top-down pathway using a projection operation to match matrix dimensions. In this way, the network can use contextual information from the five slices immediately surrounding the region-of-interest to predict the presence and location of hemorrhage.

Implementation

The approximate joint training method as described in the original Faster R-CNN implementation[16] was used for parallel optimization of the region-proposal network (RPN), 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 RPN are pruned using non-maximum suppression and used to generate detection boxes for classification. The RPN 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.

Software code for this study was written in Python 3.5 using the open-source TensorFlow r1.4 library (Apache 2.0 license)[19]. Experiments were performed on a GPU-optimized workstation with four NVIDIA GeForce GTX Titan X cards (12GB, Maxwell architecture). 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 +240 HU were clipped, and the entire volume was rescaled to a range of [−3, 3].

Statistical Analysis

The primary endpoint of this study was the detection of hemorrhage on a per-study basis. A given NCCT volume was considered to be positive for hemorrhage if any single region proposal prediction on any given slice was determined to contain hemorrhage. Based on this, algorithm performance including accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated. Furthermore, by varying the softmax score threshold for hemorrhage classification, an AUC value was calculated.

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

The secondary endpoint of this study was the ability of the algorithm to accurately estimate hemorrhage volume. This was assessed in two ways. First, predicted binary masks of hemorrhage were compared to gold-standard manual segmentations using a Dice score coefficient. Second, predicted volumes of hemorrhage were compared to gold-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 five-fold cross validation scheme was used for evaluation of the initial training cohort. In this experimental paradigm, 80% of the data is randomly assigned into the training cohort while the remaining 20% is used for validation. This process is then repeated five times until each study in the entire data set is used for validation once. Validation results below are reported for the cumulative statistics across the entire data set.

Independent Test Cohort Evaluation

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

RESULTS

Patient Selection

The initial training set cohort comprised a total of 10,159 NCCT examinations, 901 (8.9%) of which contained hemorrhage including IPH (*n*=358; 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 mL to 44.7 mL).

The independent test set cohort compromised a total of 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 a total of 23,668 images. The median hemorrhage size was 24.9 mL (interquartile range 8.3 mL to 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 data set as measured by accuracy, AUC, sensitivity, specificity, PPV, and NPV 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, 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%) of hemorrhages were missed in the cross-validation cohort compared to 4/81 (4.9%) of hemorrhages in the prospective test set (Figures 3 and 4).

Balanced data set results stratified by hemorrhage size show that in general algorithm accuracy for hemorrhages >5 mL (range 0.977 to 0.999) is higher than for hemorrhages <5

mL (range 0.872 to 0.965), with only four cases of missed hemorrhage >5 mL across both cohorts (all representing EDH/SDH). Detection accuracy of punctate hemorrhages <0.01 mL (range 0.872 to 0.883) is noticeably more challenging than small hemorrhages between 0.01 mL and 5 mL (range 0.906 to 0.965). When further stratifying results by hemorrhage type, the most challenging combinations to detect are punctate SAH or EDH/SDH with accuracy ranges of 0.830 to 0.881 across both cohorts. Complete stratification of balanced data set 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 to 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 to 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 number of GPU cards for training distribution, this process required on average 6 to 12 hours per fold. Once trained, the mask R-CNN network was able to determine 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 to 0.931) and Pearson correlations (0.953 to 0.999). Finally, while embedded for one month in an automated inference pipeline, the deep learning tool was able to accurately detect and quantify ICHs from prospective NCCT exams ordered from the emergency room.

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 thresholds[23], and decision tree analysis[24]. However, the significant 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.[25] reported a sensitivity of 0.60 and PPV 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] report a level set technique for hemorrhage quantification yielding a Dice score range between 0.858-0.917, however the algorithm is limited for hemorrhage detection as it is not designed to exclude hemorrhage on a negative exam.

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.[26] developed a multi-scale CNN for lung nodule detection with CT images while Wang et al.[27] devised a 12-layer CNN for predicting cardiovascular disease from mammograms as well as for detecting spine metastasis[28]. More recently, Phong et al.[29] described a deep learning approach for hemorrhage detection using several pre-trained networks on a small test set of 20 cases.

However, while these preliminary efforts are important, there remain several key limitations that need to be addressed prior to 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 generation of saliency maps[30] or class activation maps[31], 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 finding(s) 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 attempt 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 for use, namely every NCCT head examination performed in the emergency radiology department. The fact 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 database size for proper algorithm validation. While large data sets 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 most difficultly learning and generalizing to (e.g. punctate hemorrhages <0.01 mL represent approximately 56/10841 = 0.5% of all exams yet are also the most difficult to detect), and thus a large representative data set is required to assess performance on these critical rare entities. A large database also facilitates algorithm learning whereby the increased diversity of training examples helps the network choose more generalizable and predictive features. Finally, it is important to emphasize that cases without ICH are just as important as those with ICH, as the algorithm must also be able to correctly identify the absence of hemorrhage in the vast majority of cases despite any possible underlying pathology that may be present. To address these issues, this study takes advantages of a large training dataset comprising over 512,598 images from more than 10,000 patients, at least an order of magnitude higher than any previous study.

The most salient use-case of an accurate tool for hemorrhage detection is a triage system that alerts physicians of potentially positive exams for expedited interpretation, thus facilitating reduced TAT. The recent 2013 Imaging Performance Partnership survey of over 80 institutions rated the importance of reduced TAT as one of their highest priorities, scoring 5.7 out of a 6.0 rating[32], allowing for expedited triage of patients for therapeutic management. As an example, rapid identification of IPH patients would facilitate immediate control of blood pressure during the vulnerable first few 3-4.5 hours of symptom onset where acute deterioration is most likely[5–7]. This is supported further by the recent INTERACT-2 trial, which concluded that intensive treatment afforded by early diagnosis was associated with improved functional outcome[33].

In addition to hemorrhage detection, ICH volume metrics can be used to precisely and efficiently quantify initial burden of disease as well as serial changes, which in turn may have important clinical implications[34, 35]. For IPHs, this is most relevant within the first 2-3 hours of onset where 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 Kwak et al's ABC/2 formula[10, 36], where 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% as compared to manual segmentation[10]. While the gold standard remains manual delineation, this approach can be both time-consuming and technically challenging in the emergency room setting. By comparison, the ability of the trained CNN to rapidly and accurately quantify IPH volume with over 0.999 correlation to human experts offers a clinically feasible, improved alternative to 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 data sets obtained at our hospital center, further work is necessary to evaluate performance on a variety of vendors and scanning protocols at other institutions. Acknowledging this, CT examinations are inherently normalized by Hounsfield Units and show less image variability than plain radiographs or magnetic resonance imaging. Second, deep learning algorithms are known to be susceptible to the phenomenon of adversarial noise[37], where small but highly patterned perturbations in images may result in unexpected predictions. However, this is rare and was not encountered in the current data set, and to some extent can be mitigated by using network ensembles and denoising autoencoders[38]. Finally, while the current data set 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 data sets become available.

In conclusion, this study demonstrates 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 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 exams for interpretation and/or as a method for rapid quantification of ICH volume, overall expediting triage of patient care and offering more accurate, detailed information to guide clinical decision making.

Acknowledgments

Funding: PC: The work of PC was in part supported by grant T32EB001631.

DSC: The work of DC was in part supported by Canon Medical Systems.

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; 9:167–176. [PubMed: 20056489]
2. Goldstein JN, Gilson AJ. Critical care management of acute intracerebral hemorrhage. Curr Treat Options Neurol. 2011; 13:204–216. [PubMed: 21222062]
3. Heit JJ, Iv M, Wintermark M. Imaging of Intracranial Hemorrhage. J Stroke. 2017; 19:11–27. [PubMed: 28030895]
4. Glover MtAlmeida RR, Schaefer PW. , et al. Quantifying the Impact of Noninterpretive Tasks on Radiology Report Turn-Around Times. J Am Coll Radiol. 2017
5. Davis SM, Broderick J, Hennerici M, et al. Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage. Neurology. 2006; 66:1175–1181. [PubMed: 16636233]
6. Kazui S, Naritomi H, Yamamoto H, et al. Enlargement of spontaneous intracerebral hemorrhage. Incidence and time course. Stroke. 1996; 27:1783–1787. [PubMed: 8841330]
7. Qureshi A, Palesch Y, Investigators AI. Expansion of recruitment time window in antihypertensive treatment of acute cerebral hemorrhage (ATACH) II trial. J Vasc Interv Neurol. 2012; 5:6–9.
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; 24:987–993. [PubMed: 8322400]
9. Butcher K, Laidlaw J. Current intracerebral haemorrhage management. J Clin Neurosci. 2003; 10:158–167. [PubMed: 12637041]
10. Scherer M, Cordes J, Younsi A, et al. Development and Validation of an Automatic Segmentation Algorithm for Quantification of Intracerebral Hemorrhage. Stroke. 2016; 47:2776–2782. [PubMed: 27703089]
11. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016.
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; 162664
13. He K, Gkioxari G, Dollár P, et al. Mask R-CNN. IEEE International Conference on Computer Vision (ICCV). 2017
14. Lin T-Y, Dollár P, Girshick R. , et al. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. Honolulu, HI: IEEE; 2016. Feature Pyramid Networks for Object Detection.
15. He K, Zhang X, Ren S. , et al. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE Computer Society; 2016. Deep residual learning for image recognition.
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; 39:1137–1149. [PubMed: 27295650]

17. He K, Zhang X, Ren S. , et al. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society; 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification; 1026–1034.
18. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014 abs/1412.6980.
19. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. CoRR. 2015 abs/1603.04467.
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; 25:1163–1172. [PubMed: 18986221]
21. Ćosić D, Lončarić S. Rule-based labeling of CT head image. In: Keravnou E, Garbay C, Baud R. , et al., editorsArtificial Intelligence in Medicine: 6th Conference on Artificial Intelligence in Medicine Europe, AIME'97 Grenoble, France, March 23–26, 1997 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg; 1997. 453–456.
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; 7:507–516. [PubMed: 22081264]
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; 7:785–798. [PubMed: 22293946]
24. Gong T, Liu R, Tan CL. , et al. Classification of CT Brain Images of Head Trauma. In: Rajapakse JC, Schmidt B, Volkert G, editorsPattern Recognition in Bioinformatics: Second IAPR International Workshop, PRIB 2007, Singapore, October 1-2, 2007 Proceedings. Vol. 2007. Berlin, Heidelberg: Springer Berlin Heidelberg; 401–408.
25. Gong T, Liu R, Tan CL, et al. Classification of CT Brain Images of Head Trauma. PRIB. 2007
26. Shen W, Zhou M, Yang F, et al. Multi-scale Convolutional Neural Networks for Lung Nodule Classification. Inf Process Med Imaging. 2015; 24:588–599. [PubMed: 26221705]
27. Wang J, Ding H, Bidgoli FA, et al. Detecting Cardiovascular Disease from Mammograms With Deep Learning. IEEE Trans Med Imaging. 2017; 36:1172–1181. [PubMed: 28113340]
28. 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; 84:137–146. [PubMed: 28364643]
29. Phong TD, Duong HN, Nguyen HT. , et al. Proceedings of the 2017 International Conference on Machine Learning and Soft Computing. Ho Chi Minh City, Vietnam: ACM; 2017. Brain Hemorrhage Diagnosis by Using Deep Learning; 34–39.
30. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR. 2013 abs/1312.6034.
31. Selvaraju RR, Das A, Vedantam R, et al. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. CoRR. 2016 abs/1610.02391.
32. Nataraj S. 2013 Imaging Turnaround Times Survey Results. Advisory.com. 2014
33. Anderson CS, Heeley E, Huang Y, et al. Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage. N Engl J Med. 2013; 368:2355–2365. [PubMed: 23713578]
34. Jung SW, Lee CY, Yim MB. The relationship between subarachnoid hemorrhage volume and development of cerebral vasospasm. J Cerebrovasc Endovasc Neurosurg. 2012; 14:186–191. [PubMed: 23210046]
35. Bullock MR, Chesnut R, Ghajar J, et al. Surgical management of acute epidural hematomas. Neurosurgery. 2006; 58:S7–15. discussion Si-iv. [PubMed: 16710967]
36. Kwak R, Kadoya S, Suzuki T. Factors affecting the prognosis in thalamic hemorrhage. Stroke. 1983; 14:493–500. [PubMed: 6606870]
37. Goodfellow IJ, Shlens J, Szegedy C. International Conference on Learning Representations. San Diego, CA: 2015. Explaining and Harnessing Adversarial Examples.
38. Gu S, Rigazio L. Towards Deep Neural Network Architectures Robust to Adversarial Examples. CoRR. 2014 abs/1412.5068.

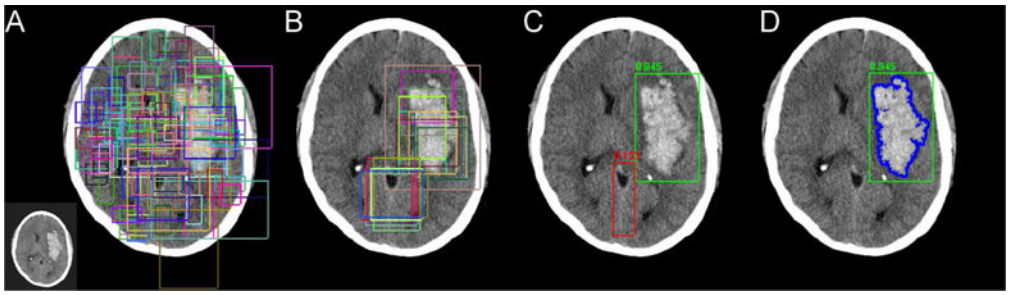

Figure 1. Overview of Mask R-CNN Approach

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 non-maximum suppression and used as input into a classifier to determine presence or absence of hemorrhage. (D) Segmentation masks are generated for positive cases of hemorrhage.

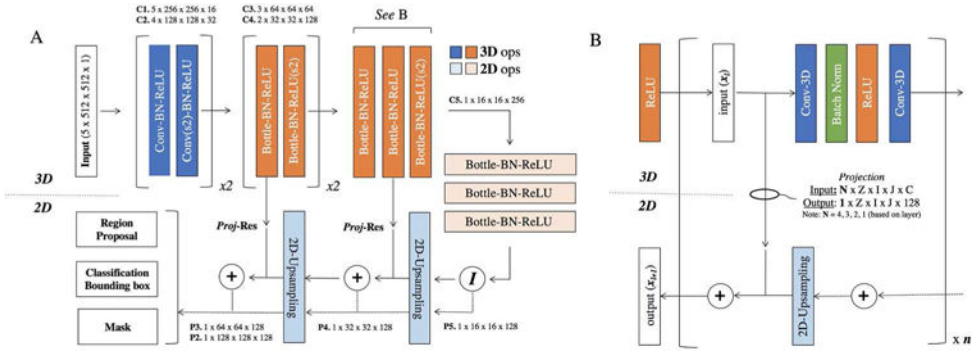

Figure 2. Convolutional Neural Network Architecture

(a) Hybrid 3D-contracting (bottom-up) and 2D-expanding (top-down) fully convolutional feature-pyramid network architecture used for 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-direction is implemented via 1 × 2 × 2 strided convolutions (marked by **s2**). Subsampling in the *z*-direction is mediated by a 2 × 1 × 1 convolutional kernel with valid padding. The expanding arm is composed entirely of 2D operations. (b) Connections between the contracting and expanding arm 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 (1) *z*-dimension shape in addition to input (**C**) and output (128) feature map size.

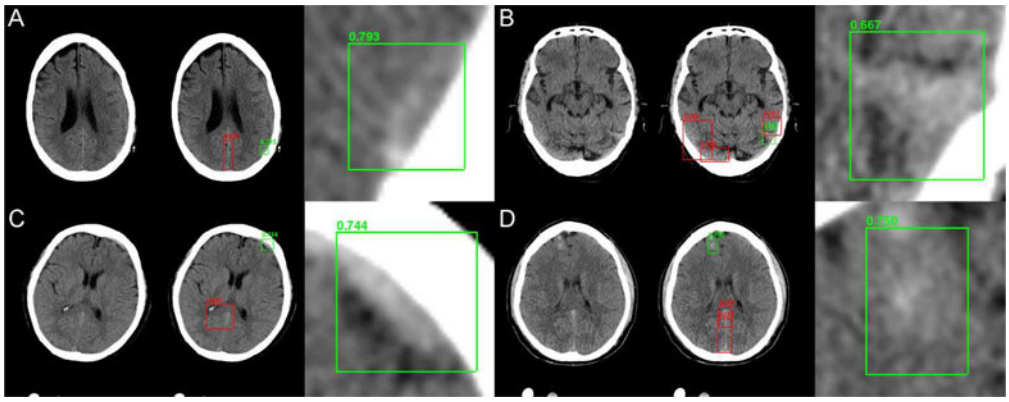

Figure 3. Example Network Predictions: True Positives

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 result. Correctly identified areas of hemorrhage (green) include subtle abnormalities representing subarachnoid (A), subdural (B and C) as well as 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, D) and beam-hardening along the periphery (B).

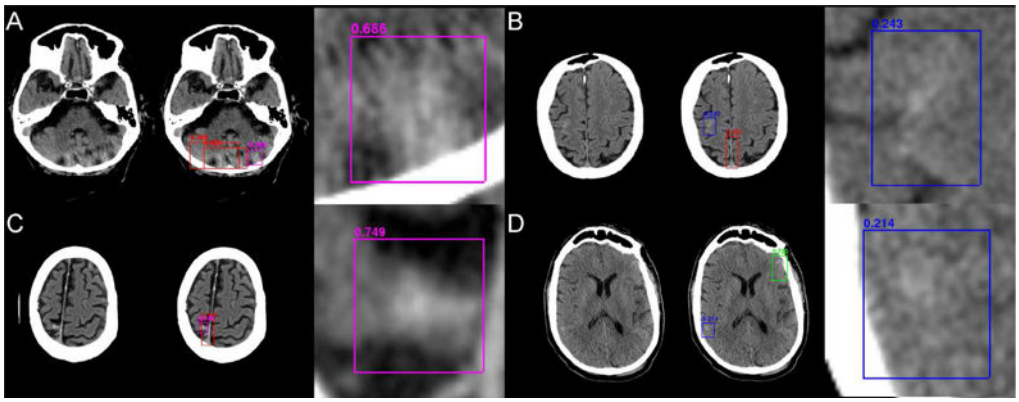

Figure 4. Example Network Predictions: False Positives and False Negatives

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 result. 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).

Table 1

Distribution of Hemorrhages by Type and Size

	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

Abbreviations: IPH, intraparenchymal hemorrhage, EDH/SDH, epidural hemorrhage/subdural hemorrhage; SAH, subarachnoid hemorrhage. Large, medium, small, and punctate hemorrhages were defined as > 25 mL, 5 to 25 mL, 0.01 to 5.0 mL, and < 0.01 mL, respectively.

Table 2

Balanced Dataset Performance Statistics Stratified By Hemorrhage Type and Size

	Accuracy		AUC		Sensitivity		Specificity		PPV		NPV	
Size	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**	**Valid**	**Test**
All ICH	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	N/A	0.857	N/A	0.667	N/A	0.975	0.973	0.963	N/A	0.745	N/A
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

Abbreviations: ICH, intracranial hemorrhage; IPH, intraparenchymal hemorrhage, EDH/SDH, epidural hemorrhage/subdural hemorrhage; SAH, subarachnoid hemorrhage. Large, medium, small, and punctate hemorrhages were defined as > 25 mL, 5 to 25 mL, 0.01 to 5.0 mL, and < 0.01 mL, respectively.