The Utility of Deep Learning: Evaluation of a Convolutional Neural Network for Detection of Intracranial Bleeds on Non-Contrast Head Computed Tomography Studies

Ojeda P, Zawaideh M, Mossa-Basha M, Haynor D
Department of Diagnostic Radiology, University of Washington Medical Center, Seattle WA.

ABSTRACT

While rapid detection of intracranial hemorrhage (ICH) on computed tomography (CT) is a critical step in assessing patients with acute neurological symptoms in the emergency setting, prioritizing scans for radiologic interpretation by the acuity of imaging findings remains a challenge and can lead to delays in diagnosis at centers with heavy imaging volumes and limited staff resources. Deep learning has shown promise as a technique in aiding physicians in performing this task accurately and expeditiously and may be especially useful in a resource-constrained context. Our group evaluated the performance of a convolutional neural network (CNN) model developed by Aidoc (Tel Aviv, Israel). This model is one of the first artificial intelligence devices to receive FDA clearance for enabling radiologists to triage patients after scan acquisition. The algorithm was tested on 7112 non-contrast head CTs acquired during 2016–2017 from a two, large urban academic and trauma centers. Ground truth labels were assigned to the test data per PACS query and prior reports by expert neuroradiologists. No scans from these two hospitals had been used during the algorithm training process and Aidoc staff were at all times blinded to the ground truth labels. Model output was reviewed by three radiologists and manual error analysis performed on discordant findings. Specificity was 99%, sensitivity was 95%, and overall accuracy was 98%.

In summary, we report promising results of a scalable and clinically pragmatic deep learning model tested on a large set of real-world data from high-volume medical centers. This model holds promise for assisting clinicians in the identification and prioritization of exams suspicious for ICH, facilitating both the diagnosis and treatment of an emergent and life-threatening condition.

Keywords: Intracranial hemorrhage, Aidoc, convolutional neural network, non-contrast head CT, FDA

1. INTRODUCTION

Non-contrast computed tomography (CT) scans of the head are the first-line diagnostic modality in the emergent setting in patients with head trauma or symptoms suggestive of a stroke or rise in intracranial pressure. The last several decades have seen a progressive increase in head CT utilization within the emergency context, and it is a prerequisite study for determining the need for neurosurgical intervention in the context of intracranial hemorrhage (ICH). While rapid detection of ICH on CT is a critical step in assessing patients with neurological symptoms in the emergency setting, prioritizing scans for interpretation by the acuity of imaging findings remains a challenge and can lead to delays in diagnosis at centers with heavy imaging volumes and limited staff resources. An automated head CT scan screening and triage system would be valuable for case prioritization in a busy trauma care setting or to facilitate decision-making in remote locations without immediate radiologist availability. Deep learning has shown promise as a technique in aiding physicians in performing this task accurately and expeditiously and may be especially useful in a resource-constrained context. Recently, there have been several advances in the application of deep learning for radiologic tasks in both the classification and segmentation domains, utilizing a variety of imaging modalities including radiography, CT, and MRI. Developing robust, generalizable algorithms usable in clinical settings requires creation of a large number of accurately labeled scans from a patient distribution that reflects the population of interest. In this study, we evaluated the performance of a fully automated and cloud based diagnostic server running a convolutional neural network developed by Aidoc (Tel Aviv, Israel). This model is one of the first artificial intelligence (AI) devices to receive Federal Drug Administration (FDA) clearance for enabling radiologists to triage patients after acquisition. We report promising results of a scalable and clinically pragmatic tool tested on a large set of real-world data from high-volume medical centers.

Medical Imaging 2019: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE
Vol. 10949, 109493J · © 2019 SPIE · CCC code: 1605-7422/19/$18 · doi: 10.1117/12.2513167

2. METHODS

2.1 Aidoc Model Development

A proprietary convolutional neural network architecture was used.
The training dataset included approximately 50,000 CT studies collected from 9 different sites. In total, data was derived from 17 different scanner models. CT data slice thickness (z-axis) ranged from 0.5 to 5 mm. Data from all anatomic planes was used, when available (axial, sagittal, coronal). Only soft-tissue kernel images were used. Ground truth labeling structure varied depending on hemorrhage type and size, and included both weak and strong labeling schema. Label types included study-level classification for diffuse SAH, slice-level bounding boxes around indistinct extra and intra-axial hemorrhage foci, and pixel-level semantic segmentation of well-defined intraparenchymal hemorrhage.

2.2 Validation Dataset

After approval by the international review board of the University of Washington, we retrospectively collected a validation dataset composed of 7112 non-contrast head CTs acquired from two large urban academic and trauma centers. No data from these hospitals had been used during the algorithm training process and Aidoc staff were at all times blinded to the ground truth. The data was pulled from local PACS servers and anonymized in compliance with institutional HIPAA guidelines. Studies with ICH were labeled as ‘positive’ and those without evidence of hemorrhage were labeled as ‘negative.’ For an exam to be included in either category, the patient had to be 18 years or older at the time of image acquisition and the non-contrast head CT studies had to be performed between January 1, 2016 and December 31, 2017. Of these studies, 30.7% were performed at the academic center with a General Electric Healthcare Revolution CT scanner. 69.3% of the studies were performed at the trauma center, using either a Siemens SOMATOM Force or the Siemens Definition AS Plus CT scanners.

We queried the PACS for *presumed ICH present* studies and *presumed ICH absent* studies. We defined *presumed ICH present* studies as those in which an initial study was followed-up with 1 or more additional studies spaced one to six hours of one another. This criterion was based on our institutional protocol of reimaging patients with intracranial hemorrhage within 4 hours of the initial non-contrast CT head which demonstrated the abnormality. We defined *presumed ICH absent* studies as those in which only a single study was performed, without a close follow-up. These studies served as a crude, or preliminary, ground truth for algorithm evaluation.

2.3 Model Evaluation

Cases which met selection criteria were submitted to Aidoc for testing; in patients who received multiple CT scans within a six-hour period, only the first exam was submitted. The primary Aidoc output is binary and reported as either ICH present or absent. If ICH is detected, representative images are marked and incorporated as an additional series in the initial DICOM study. Results for each study were then analyzed by our team and compared to the ground truth labels. Sensitivity, specificity, and accuracy at a single algorithm output operating point was used for evaluation.

2.4 Error Analysis

Cases in which the Aidoc output and the preliminary ground truth were discrepant were reviewed manually by one of four radiologists – two senior neuroradiologists and two diagnostic radiology residents. Discrepant CT scans were reviewed by the radiologists on a custom web-based viewer built upon the Aidoc framework. Intracranial presence of blood products due to any etiology such as hemorrhagic contusion, acute vessel thrombus, or tumor/infarct with hemorrhagic component was included in the definition of true positive. Repetitive imaging patterns which emerged as common causes of false positive or negative interpretations were binned into discrete categories.
In cases of discordant findings between the model and ground truth, a combination of efforts were made to determine whether the crude ground truth label was in fact correct. This involved review of the associated clinical report originally dictated by a subspecialty neuroradiologist, review of the source images, and review of any available subsequent

imaging. If the crude label did not match the reported radiologic findings, the studies were reassigned the appropriate ground truth label.

Cases in which the crude ground truth and the Aidoc output were concordant were assumed to be true positives or negatives. However, to assess the error rate of the concordant cases between Aidoc output and the crude labels, 100 concordant cases were cross-checked with the associated radiology reports.

3. RESULTS

Of the 7112 cases, 1661 were positive for hemorrhage (after review of discordant cases and possible label reassignment), of which the model detected 1579 (Sensitivity = 95%). The cases of hemorrhage missed by Aidoc are labeled as false negatives, of which there were a total of 82 cases. The most common false negatives (59 cases) involved small hemorrhages, all of which were managed conservatively. The Aidoc algorithm missed 11 cases of a hyperdense middle cerebral artery (MCA) sign. Of these, 8 required emergent thrombectomies. In total, 12 of the 82 false negative cases required some form of medical/surgical intervention. The details of these false negative cases are presented in Tables 2 and 3.

Table 1. Error Matrix

Aidoc Categorization	Total Case Validation	Number of Cases (out of 7112)	Percentage of Total
Aidoc ICH Negative (n = 5462)	True Negatives	5380	75.6%
	False Negatives	82	1.2%
Aidoc ICH Positive (n = 1650)	True Positives	1579	22.2%
	False Positives	71	1.0%

Table 2. Aidoc false negative studies categorized by reason for erroneous classification.

Reasons for False Negative	Number	Percent of False Negatives	Percent of True Positives + False Negatives
MCA Sign	11	13.4%	0.7%
Large Bleed	6	7.3%	0.4%
Small Bleed	59	72.0%	3.6%
Hemorrhagic Metastasis	1	1.2%	0.1%
Other	5	6.1%	0.3%

Table 3. Management outcome of Aidoc false negative studies

Reasons for False Negative	Number	Conservative Management	Intervention Needed
MCA Sign	11	3	8
Large Bleed	6	4	2
Small Bleed	59	59	0
Hemorrhagic Metastasis	1	1	0
Other	5	3	2

5451 cases were negative for hemorrhage, of which the model correctly assigned 5380 cases (Specificity = 99%). There were 71 total false positive cases. The most common error (16 cases) involved incorrect identification of a thick falx as a hyperdense hemorrhage (Table 4). The ‘Other’ category includes miscellaneous imaging findings which were undefined prior to manual validation analysis. These cases are further categorized in Table 5.

Table 4. Aidoc false positive studies categorized by reason for erroneous classification.

Reasons for False Positives	Number	Percent of False Positives
Bone volume averaging	4	5.6%
Soft tissue averaging	5	7.0%
Dense calcifications	3	4.2%
Bridging cerebral veins	7	9.9%
Thick Falx	16	22.5%
Other	36	50.7%

Table 5. The Aidoc false positive studies categorized as 'Other' further classified by reason for erroneous classification.

Reasons for 'Other' False Positives	Number	Percent of 'Other'
Noise in Image	1	2.9%
Choroid Plexus	2	5.7%
Fahr disease	1	2.9%
Calcified structure	5	14.3%
Dural thickening	5	14.3%
Dural sinus	4	11.4%
Hyperdense mass	2	5.7%
Dolichoectasia	1	2.9%
Deep cerebral vein	1	2.9%
Pseudosubarachnoid in hypoxic ischemic encephalopathy	1	2.9%
Beam hardening	3	8.6%
Thrombosed aneurysm	1	2.9%
Basilar artery	2	5.7%
Cranioplasty	1	2.9%
Calvarial metastasis	1	2.9%
Artifactual focal parenchymal hyperattenuation	1	2.9%
Streak artifact	3	8.6%
Post-surgical	2	5.7%

Overall accuracy was 98%. Table 6 demonstrates model performance metrics.

Of the 100 concordant cases which were cross-checked with the associated radiology reports, there was a 2% error rate. Both cases were labeled as *ICH present* by Aidoc output and labeled *presumed ICH present* by the crude, preliminary ground truth schema - although when compared to the radiology reports, these cases did not have hemorrhage.

Table 6. Model performance metrics.

Sensitivity	0.95
Specificity	0.99
Positive Predictive Value/Precision	0.96
Negative Predictive Value	0.98
Accuracy	0.98

4. DISCUSSION

4.1 Model Performance

We have shown that a clinically pragmatic deep neural network trained on a large-sized clinical imaging dataset (~50,000 cases) can detect critical radiological conditions such as ICH with high accuracy (98%) when tested on real-world data obtained at high-volume academic and trauma centers. Importantly, the proposed algorithm detected a variety of ICH cases from a heterogeneous dataset without a priori information about the hemorrhage location and without controlling for factors such as scanner type, patient comorbidities or image acquisition parameters.

Deep learning is increasingly demonstrating great utility in medical diagnosis. Convolutional neural networks (CNNs) are a specific class of multi-layer, feed-forward deep learning algorithms which excel in object detection ("is ICH present in these images?"). CNNs have recently demonstrated success in retinal photography evaluation, tuberculosis detection on chest X-rays, and mammography. The models are intended to act as sophisticated digital evaluation tools, assisting with prioritization of studies highly suspicious for abnormality, and providing the radiologist with extra data to supplement their final diagnostic decisions [12].

The CNN developed by Aidoc to identify intracranial hemorrhage on non-contrast head CTs has recently received FDA clearance as an AI tool intended to enable radiologists to triage acquired studies. The algorithm provides a means to reduce the time between image acquisition and identification of an acute intracranial bleed by the radiologist. The expedited diagnosis results in timely treatment which is crucial in the setting of brain hemorrhage. There is extensive evidence supporting the need for intensive control of blood pressure within 6 hours of symptom onset for most patients with ICH, which results in improved functional outcomes. [345]

This model demonstrates high sensitivity and specificity based on our current evaluation, 95% and 98%, respectively. In addition, the fact that performance was excellent on a dataset not used for training suggests that the system is robust. The analysis has been promising to indicate the potential impact this model may have on the daily work-flow and efficacy of the radiologist.

False negative model output carries the most clinically relevant consequence, as a missed finding may delay appropriate time-sensitive management. Identification of a hyperdense middle cerebral artery (MCA) sign indicates an acute thrombus which may require emergent thrombectomy - as was the case for 8 of the 11 missed MCA signs. Although the Aidoc algorithm most often missed small intraparenchymal hemorrhages (59 cases), none of these required emergent management. It should be noted, however, that an estimated 15% of intracranial hemorrhages are associated with concurrent anticoagulation therapy [6], which inhibits physiologic clotting. ICH in this patient population is associated with subsequent hemorrhage growth [7] and worse outcomes, particularly for those on warfarin [8]. Consequently, identification of even small hemorrhages in anticoagulated patients is of clinical relevance, as immediate reversal of the anticoagulation therapies may reduce morbidity [9].

With respect to clinical impact, false positive cases may incorrectly alert the interpreting physician that an abnormality is present. Which this may lead to mild workflow inefficiencies, such errors are less harmful than false negatives. Nevertheless, these cases demonstrate potential areas for algorithmic improvement.

Clinically, a meaningful assessment of model performance should include comparison to reported human-level performance. In a study performed at the same trauma institution in 2009, Miyakoshi reported a radiology resident error rate of 1.1% for intracranial hemorrhage detection[10], closely approximating Aidoc model error rate of 1.2%.

4.2 Data Collection

Deep learning algorithms require large data sets in order to train a successful, generalizable model. This necessity is, in principle, easily addressed due to the ease of accessibility to large amounts of digital data. However, clinical data typically needs extensive curation and standardization before it can be used for algorithm training.

Our group was confronted with this issue upon initial PACS query for non-contrast head CTs demonstrating either an ICH or no bleed. While a large number of studies (nearly 40,000, including a large number of repeat examinations, met the age and date ranges) were potentially available, automatically determining which studies fell fall into the categories of *present* or *absent* ICH was more challenging. Reviewing every study's associated radiology report for determination of ground truth would have required an inordinate amount of time with such a large data set. Text mining the associated radiology reports for specific words such as 'hemorrhage' would also be difficult because of the "negative problem": head CTs without bleeds are most often reported as "No evidence of *hemorrhage*" or "No *hemorrhage* is identified", etc. As natural language processing continues to improve, however, this may become less of an issue.

In the setting of ICH, a neurosurgeon is consulted for further work-up and treatment. At our medical centers, a pre-existing protocol is followed by the Department of Neurosurgery in which patients with ICH receive follow-up imaging with non-contrast enhanced head CTs within 4 hours of their initial scan. This is done for surveillance of the bleed as rapid progression of these hemorrhages is not rare and may lead to increased disability or death if treatment is delayed. We therefore determined that patients with repeat non-contrast head CT exams performed within a few hours of one another would likely be representative of patients with hemorrhage, while patients who received only one non-contrast head CT were unlikely to have an ICH.

We applied these parameters to our selection criteria for determining which studies represented ICH *presumed present* versus ICH *presumed absent.* This search algorithm allowed us to reduce our data set to 7112 studies. The system was not flawless since it was not based on the ground truth information present in the associated radiology reports, but it allowed for substantial narrowing of our data set to include only studies highly likely to have ICH truly present or highly likely to have no ICH. We labeled this data as *presumed ICH present* or *presumed ICH absent* and used the labels to compare to the CNN model output. All discrepant imaging studies were then reviewed by comparison to the associated radiology report. In the cases where labeling error were identified, these were easily re-labeled and these final labels were used in the calculation of sensitivity and specificity. Overall, this data collection schema greatly reduced (by over 90%) the effort required to validate the Aidoc output rigorously.

Review of a sample of concordant cases showed that when the Aidoc evaluation and the crude ground truth label were in agreement, the crude ground truth label was almost always correct with a 2% error rate. The errors identified were false positives, labeled as *ICH present* by Aidoc and *ICH presumed present* by the crude labeling technique, although no hemorrhage was identified on manual review of the original reports/source images. As these errors were limited to false positives, this type of discrepancy has little impact on clinical management of these patients. As described previously, false negative cases may affect time to treatment in medically urgent cases whereas false positives would be limited to causing work flow delays.

4.3 Limitations

Several study limitations should be considered. Occasionally, the radiologist who provided the initial CT head report would equivocate whether a finding represented hemorrhage or artifact. Determining the ground truth in these instances proved more challenging. To address this, our group re-reviewed the source images and subsequent imaging studies to assign a final label.

Secondly, to gather a substantial number of validation studies, a crude labeling schema was utilized. However, this schema resulted in approximately a 2% error rate, which we felt was acceptable given that discordant findings were manually reviewed, and no better alternative query schema was available for the degree of throughput. Other institutions with this same query limitation could consider this approach given the low error-rate, and expedite the data acquisition process.

Lastly, our validation dataset represented two centers, a single academic site and a single trauma center. Future studies at additional centers would further assess the generalizability of the model's performance characteristics. Deep learning models utilized in medical settings have been described to generalize poorly in tested on a dataset from a different distribution[11–13]. Despite this, the model performed well at our two institutions, which holds promise for utility at other sites.

5. CONCLUSION

In summary, we report promising results of a scalable and clinically pragmatic deep learning algorithm tested on a large set of real-world data from high-volume medical centers that requires no human intervention to accurately characterize the presence or absence of ICH. This model holds promise for assisting clinicians in the identification and prioritization of exams suspicious for ICH, facilitating both the diagnosis and treatment of an emergent and life-threatening condition.

We also show that when extracting ground truth data from a single institution, the use of knowledge about institutional protocols can shorten the labeling. This would expedite the training data acquisition process in cases when manual labeling of every data point for definitive ground truth determination would be excessively laborious.

This work has not been previously presented or submitted for publication

REFERENCES

1. Thrall JH, Li X, Li Q, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *J Am Coll Radiol*. 2018;15:504-508. doi:10.1016/j.jacr.2017.12.026
2. Savadjiev P, Chong J, Dohan A, et al. Demystification of AI-driven medical image interpretation: past, present and future. *Eur Radiol*. August 2018:1-9. doi:10.1007/s00330-018-5674-x
3. Sprigg N, Flaherty K, Appleton JP, et al. Tranexamic acid for hyperacute primary IntraCerebral Haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. *Lancet (London, England)*. 2018;391(10135):2107-2115. doi:10.1016/S0140-6736(18)31033-X
4. Song L, Sandset EC, Arima H, et al. Early blood pressure lowering in patients with intracerebral haemorrhage and prior use of antithrombotic agents: pooled analysis of the INTERACT studies. *J Neurol Neurosurg Psychiatry*. 2016;87(12):1330-1335. doi:10.1136/jnnp-2016-313246
5. Cordonnier C, Demchuk A, Ziai W, Anderson CS. Intracerebral haemorrhage: current approaches to acute management. *Lancet*. 2018;392(10154):1257-1268. doi:10.1016/S0140-6736(18)31878-6
6. Bejot Y, Cordonnier C, Durier J, Aboa-Eboule C, Rouaud O, Giroud M. Intracerebral haemorrhage profiles are changing: results from the Dijon population-based study. *Brain*. 2013;136(2):658-664. doi:10.1093/brain/aws349
7. Cucchiara B, Messe S, Sansing L, Kasner S, Lyden P. Hematoma Growth in Oral Anticoagulant Related Intracerebral Hemorrhage. *Stroke*. 2008;39(11):2993-2996. doi:10.1161/STROKEAHA.108.520668
8. Dequatre-Ponchelle N, Hénon H, Pasquini M, et al. Vitamin K Antagonists–Associated Cerebral Hemorrhages. *Stroke*. 2013;44(2):350-355. doi:10.1161/STROKEAHA.112.672303

9. Huttner HB, Kuramatsu JB. Aktuelle Therapieziele bei intrazerebralen Blutungen. *Medizinische Klin - Intensivmed und Notfallmedizin*. 2017;112(8):695-702. doi:10.1007/s00063-017-0361-2

10. Miyakoshi A, Nguyen QT, Cohen WA, Talner LB, Anzai Y. Accuracy of preliminary interpretation of neurologic CT examinations by on-call radiology residents and assessment of patient outcomes at a level I trauma center. *J Am Coll Radiol*. 2009;6(12):864-870. doi:10.1016/j.jacr.2009.07.021

11. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys*. 2018;45(3):1150-1158. doi:10.1002/mp.12752

12. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. Sheikh A, ed. *PLOS Med*. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683

13. Therrien R, Doyle S. Role of training data variability on classifier performance and generalizability. In: *Proc. SPIE Medical Imaging 2018: Digital Pathology*. ; 2018. doi:10.1117/12.2293919