Neuroradiology (2020) 62:335–340
https://doi.org/10.1007/s00234-019-02330-w

DIAGNOSTIC NEURORADIOLOGY

Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage

Daniel T. Ginat[1]

Received: 9 September 2019 / Accepted: 20 November 2019 / Published online: 11 December 2019

Abstract

Purpose To analyze the implementation of deep learning software for the detection and worklist prioritization of acute intracranial hemorrhage on non-contrast head CT (NCCT) in various clinical settings at an academic medical center.

Methods Urgent NCCT scans were reviewed by the Aidoc (Tel Aviv, Israel) neural network software. All cases flagged by the software as positive for acute intracranial hemorrhage on the neuroradiology worklist were prospectively included in this assessment. The scans were classified regarding presence and type of hemorrhage, whether these were initial or follow-up scans, and patient visit location, including trauma/emergency, inpatient, and outpatient departments.

Results During the 2 months of enrollment, 373 NCCT scans were flagged by the Aidoc software for possible intracranial hemorrhage out of 2011 scans analyzed (18.5%). Among the flagged cases, 275 (72.4%) were positive; 290 (77.7%) were inpatient cases, 75 (20.1%) were trauma/emergency cases, and eight (2.1%) were outpatient cases, and 229 of 373 (62.5%) were follow-up cases, of which 219 (95.6%) inpatient cases. Among the 144 new cases flagged for hemorrhage, 66 (44.4%) were positive, of which 39 (58.2%) were trauma/emergency cases. The overall sensitivity, specificity, positive predictive value, negative predictive value, and accuracy were 88.7%, 94.2% and 73.7%, 97.7%, and 93.4%, respectively. The accuracy of the intracranial hemorrhage detection was significantly higher for emergency cases than for inpatient cases (96.5% versus 89.4%).

Conclusion This study reveals that the performance of the deep learning software for acute intracranial hemorrhage detection varies depending upon the patient visit location. Furthermore, a substantial portion of flagged cases were follow-up exams, the majority of which were inpatient exams. These findings can help optimize the artificial intelligence-driven clincical workflow.

Keywords Deep learning · Neural networks · CT · Intracranial hemorrhage

Abbreviations

ED	Emergency department
ICH	Intracranial hemorrhage
NCCT	Non-contrast head CT

Introduction

Intracranial hemorrhage (ICH) is the second leading cause of stroke worldwide and has various etiologies, including trauma, infarction, aneurysm rupture, and anticoagulant therapy [1, 2]. It is estimated that 37,000 to 50,000 cases of ICH occur in the USA annually [3]. Despite a slight historical decrease in the incidence of ICH, the case fatality has not decreased significantly, with the 30-day mortality rate still as high as 47% [4]. Half of the resulting mortality occurs in the first 24 hours, and early treatment has shown to improve outcomes [5, 6]. Hematoma expansion can lead to worsening of neurologic deficits, and irreversible damage can occur as early as the first few hours after onset of ICH, making accurate diagnosis crucial for appropriate management of these patients [3]. Therefore, a timely diagnosis of intracranial hemorrhage on CT is important. Neuroimaging is essential for diagnosing and characterizing ICH, in terms of the particular type, location, and size, thereby guiding patient management [2].

The increasing study size and patient volume has an added burden on the practicing radiologist [7]. On average, a radiologist currently interprets one image every 3 to 4 seconds and scans are prone to wait in a queue for their interpretation [7]. Artificial intelligence software that can screen CT images for

✉ Daniel T. Ginat
dtg1@uchicago.edu

[1] Department of Radiology, Section of Neuroradiology, University of Chicago, 5841 S Maryland Avenue, Chicago, IL 60637, USA

Fig. 1 Schematic diagram of the neural network integration and workflow on PACS

acute findings has the potential to assist radiologists with their burgeoning case-loads [8]. One approach is to use those screening tools for prioritization of acute ICH cases in a worklist. Thus, it is important to ascertain the potential impact of such a system in terms of not only accuracy of ICH detection, but also the type of hemorrhage, the patient location, and the acuity of the hemorrhage in order to best allocate resources.

The Aidoc software accuracy for the detection of acute ICH on CT was reported in a prior study, with a sensitivity of 95%, a specificity of 99%, and overall accuracy of 98% [9]. However, deep learning models utilized in medical settings have been found to generalize poorly to datasets from different distributions [10]. Thus, it is of interest to test the Aidoc software at additional centers in order to assess the generalizability of the neural network performance characteristics. Accordingly, the purpose of this study is to further evaluate the neural network performance for the detection of ICH on NCCT in different clinical settings, in terms of patient visit location, types of hemorrhage, and the influence of follow up exams, since such factors may be relevant for optimal triaging.

Materials and methods

AI system

An FDA approved convolutional neural network algorithm was developed by Aidoc (Tel Aviv, Israel) for ICH detection on NCCT. The algorithm was trained and tested on CT scans from 9 medical centers and 17 different scanners [9].CT data slice thickness (z-axis) ranged from 0.5 to 5 mm. Data from all anatomic planes was used, when available. 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 subarachnoid hemorrhage, slice-level bounding boxes around indistinct extra- and intra-axial hemorrhage, and pixel-level semantic segmentation of well-defined intraparenchymal hemorrhage [9].

Integration and deployment

All urgent NCCT scans performed at a single academic medical center were automatically and immediately forwarded for analysis by the software in real-time. The NCCT scans were performed on nine different scanners, each using 120 kVp, but variable tube currents ranging from 185 to 350 mA, and axial sections with 5 mm slice thickness, but coronal and sagittal sections with 3 mm slice thickness. The scans were sent to analysis with no additional data or clinical context of the study. Once a positive ICH case is detected, the software sends a notification through a standalone application and would also flag and elevate the case in the worklist. In addition, the Aidoc software shows a preview of the abnormal key images and those are added to the existing studies. A schematic of the integration and workflow is depicted in Fig. 1.

Validation dataset

All cases that were flagged by the software as potentially positive for acute intracranial hemorrhage, during January and February 2019, were prospectively included in this assessment. For evaluation purposes, all the flagged cases were analyzed by a single board-certified radiologist with a certificate of added qualification in neuroradiology. The corresponding NCCT reports were also reviewed for agreement. This assessment constituted the ground truth for the study.

Table 1 Error matrix

Aidoc classification	Case validation	Number of cases and percentage among positive or negative cases	Percentage among total cases ($N = 2011$)
Positive for ICH ($N = 373$)	True- positive	275 (73.5%)	13.6%
	False- positive	98 (26.5%)	4.9%
Negative for ICH ($N = 1638$)	True- negative	1603 (97.9%)	79.7%
	False- negative	35 (2.1%)	1.7%

Table 2 Types of ICH in this study

Type of ICH	Number of cases with the finding among the positive flagged cases and proportion ($N = 431$)	Number of cases with the finding among the non-flagged cases and proportion ($N = 39$)
Intraparenchymal hemorrhage	152 (35.3%)	11 (28.2%)
Subarachnoid hemorrhage	96 (22.3%)	7 (17.9%)
Subdural hemorrhage	82 (19.0%)	16 (41.0%)
Intraventricular hemorrhage	80 (18.6%)	1 (2.6%)
Epidural hematoma	6 (1.4%)	2 (5.1%)
Extra axial hematoma not otherwise specified	10 (2.3%)	1 (2.6%)
Indeterminate	5 (1.2%)	1 (2.6%)

Some cases had more than one type of hemorrhage

Analysis

Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of the software for the detection of intracranial hemorrhage on NCCT were computed. The results were further analyzed based on the different patient visit locations and whether the algorithm flagged new findings, which may require more urgent prioritization, versus follow up cases. This study was approved by the Institutional Review Board and no patient identifiers were recorded. Since this study was determined to involve quality improvement research, the need for patient consent was waived.

Results

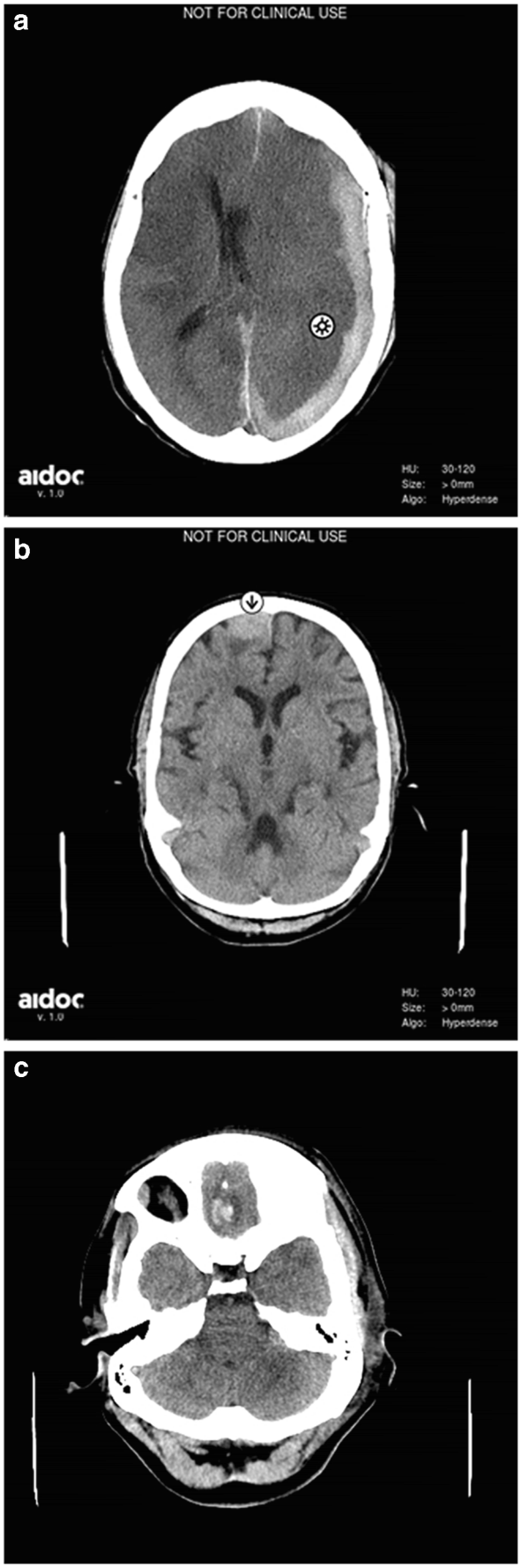
During the enrollment period, 2011 urgent NCCT scans of the head were performed and analyzed by the software. A total of 373 (18.5%) of these exams were flagged by the software for possible acute intracranial hemorrhage, while the prevalence of acute ICH in this series was 20.3%. Among the flagged cases, 275 (72.4%) were positive for ICH, while 98 (26.3%) were negative for hemorrhage. Among the 1639 cases not flagged by the software, 1603 (97.9%) were negative for hemorrhage and 35 (2.1%) were positive for hemorrhage.

The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy for all cases were 88.7%, 94.2% and 73.7%, 97.7%, and 93.4% (95% confidence interval: 92.2% to 94.4%), respectively, with additional details listed in Table 1. The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy for inpatient cases were 89.4%, 89.4% and 78.6%, 95.1%, and 89.4% (95% confidence interval: 87.1% to 91.4%), respectively. The sensitivity, specificity, positive predictive value, negative predictive value, and accuracy for emergency cases were 86.3%, 97.0% and 58.7%, 99.3%, and 96.5% (95% confidence interval: 95.2% to 97.5%), respectively. Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy for outpatient cases were 75.0%, 95.0% and 37.5%, 99.0%, and 94.2% (95% confidence interval: 87.9% to 97.9%), respectively.

Table 3 Types of false- positive ICH findings

Type of false-positive finding	Number of cases that showed the finding and percentage out of positive flagged cases ($N = 373$)
Intraparenchymal tumor[a]	15 (4.0%)
Thrombus[a]	3 (0.8%)
Ischemic infarct with cortical laminar necrosis or retained contrast[a]	3 (0.8%)
Vessels outlined by cytotoxic edema	10 (2.7%)
Craniotomy/craniectomy meningogaleal complex	14 (3.8%)
Choroid plexus calcification	19 (5.1%)
Bone	1 (0.3%)
Normal brain and vessels	16 (4.3%)
Falx cerebri	15 (4%)
Beam hardening artifacts	21(5.7%)

[a] Pathological findings



◀ **Fig. 2** Examples of a true-positive case with subdural hemorrhage (**a**), a false-positive case with a meningioma (**b**), and a false-negative case with hemorrhagic contusion (**c**)

False-positive flagged cases were attributed to various causes, such as artifacts, thick dura, intra-arterial clot, calcifications, and tumors. Interestingly, the software flagged 295 (79.0%) cases with any type of pathological finding. The types and rates of different pathological and non-pathological true-positive findings and false-positive findings are included in Tables 2 and 3 and examples of true-positive, false-positive, and false-negative cases flagged by the software are shown in Fig. 2.

Among all cases flagged by the deep learning system, 144 (38.6%) were initial scans, while 229 (61.3%) cases were follow-up scans on patients who had recently been found to have ICH. Among the 275 true positive cases that were flagged by the software, 67 (24.4%) were initial scans, and 208 (75.6%) were follow-up cases. The visit location of flagged scans comprised 75 (20.1%) trauma/emergency cases, 290 (77.7%) inpatient cases, and eight (2.1%) outpatient cases. Among the 229 follow-up cases flagged by the software, 219 (95.6%) were inpatient cases. Among the 144 new cases flagged for ICH, 66 (44.4%) were positive, of which 39 (58.2%) were trauma/emergency cases. Among the 1638 cases not flagged by the software, 994 (60.7%), 548 (33.5%), and 96 (5.9%) were trauma/emergency, inpatient, and outpatient cases, respectively. Among the 35 positive cases for ICH not flagged by the software, 10 (28.6%) were initial scans, while 25 (71.4%) were follow up cases.

Discussion

While general machine learning techniques have been used for detecting ICH on CT with success [10, 11], there is limited information regarding the impact of deep learning methods in actual clinical workflow pertaining to acute ICH on CT in an academic setting. This study indicates that the Aidoc deep leaning software has variable performance for the detection of ICH depending on the clinical setting, in which the accuracy is significantly higher for trauma/emergency cases than for inpatient cases. This disparity may be attributable to the presence of more confounding features on the inpatient scans, such as postoperative findings and also due to differences in scan quality, although the main scan parameters were consistent across the different scanners in this study. There was also variation in the software detection performance based on the type of ICH detected. Intraparenchymal hemorrhages were the most common type of ICH correctly detected by Aidoc, while the software disproportionately missed subdural and epidural hemorrhage, although these tended to be relatively small in size.

The deep learning software flagged several non-hemorrhagic pathologies due to their hyperattenuating appearce. Although such instances were counted as false positives, these findings may also be of clinical significance. Otherwise, there were a few cases that should not have been flagged, such as those with artifacts. Furthermore, a substantial portion of flagged exams were follow-up stability scans, the vast majority of which were inpatient cases. These flagged follow up cases for ICH may not be as noteworthy as newly identified cases of ICH and could also potentially prioritize these at the expense of more significant non-hemorrhagic conditions on the worklist that are otherwise not flagged by the deep learning system. The fraction of inpatient studies can vary significantly between medical institutions, but inpatient studies are generally considered less time sensitive than trauma/emergency cases. Nevertheless, newly diagnosed ICH of clinical significance does occur in inpatients and as such should be flagged as a priority.

The overall accuracy of the detection of ICH on CT was substantially lower in this study than in a prior study that also assessed the Aidoc software performance (93% versus 98%) [9], and was based on a smaller subset (373 flagged cases compared to 1660). In particular, the positive predictive value was most discrepant between this study (73%) and the prior study (96%). This disparity may be attributable to the differences in patient visit locations encompassed, in which the prior study only comprised trauma centers. The trauma center scans likely had fewer confounding features on the CT that could be prone to misinterpretation as false positive by the software. Otherwise, the negative predictive value was 98% in both studies. This relatively high negative predictive value is beneficial, since a missed hemorrhage may be more problematic from a clinical perspective than a false- positive or a false- negative.

This study has several limitations. For example, the true-positive and false-negative rates may be inflated due to over-representation of cases with actual ICH due to serial follow up as opposed to cases without hemorrhage. In addition, the number of flagged cases is too small to produce a decent statistic, and the Aidoc software also does not currently assess changes in size of the intracranial hemorrhage, which may be relevant for follow up cases, whereby those that have enlarged may warrant an alert. Furthermore, the clinical impact of the software, in terms of the significance of flagged cases with pathology not related to ICH, reduction of the turnaround time, a survey of radiologists regarding their personal perspectives regarding the software implementation, and whether there was improved patient outcome were not a part of this study, but can be addressed in future studies. Nevertheless, this study identified potential deficiencies in the current software version, such as not accounting for patient visit location and whether there are prior head CTs. Such information could provide important clinical context to improve the overall algorithm accuracy, thereby flagging cases in a more useful manner.

Conclusion

The analysis in this study reveals that the performance of the neural network software for acute intracranial hemorrhage detection on NCCT varies depending upon the patient visit location. Notably, a substantial portion of flagged cases were follow-up exams, the majority of which were inpatient exams. While the Aidoc software holds promise for prioritizing exams and assisting in the detection of ICH, further optimization of the software implementation for different clinical practice settings is warranted.

Funding Information None

Compliance with ethical standards

Conflict of interest The author declares no conflict of interest.

Ethical approval The study was approved by the Office of Clinical Effectiveness of the University of Chicago, Chicago, USA.

Informed consent For this type of study formal consent is not required.

References

1. van Asch CJJ, Luitse MJA, Rinkel GJE, van der Tweel I, Algra A, Klijn CJM (2010) 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 9:167–176
2. Heit JJ, Iv M, Wintermark M (2017) Imaging of intracranial hemorrhage. J Stroke 19:11–27
3. Carter JA, Curry W (2017) Intracerebral hemorrhage: pathophysiology and management for generalists. Hosp Med Clin 6:95–111
4. Zahuranec DB, Lisabeth LD, Sánchez BN et al (2014) Intracerebral hemorrhage mortality is not changing despite declining incidence. Neurology 82:2180
5. Elliott J, Smith M (2010) The acute management of intracerebral hemorrhage: a clinical review. Anesth Analg 110:1419–1427
6. Fujitsu K, Muramoto M, Ikeda Y, Inada Y, Kim I, Kuwabara T (1990) Indications for surgical treatment of putaminal hemorrhage. 73:518
7. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, Erickson BJ, Kallmes DF (2015) The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 22:1191–1198
8. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 392:2388–2396
9. Ojeda P, Zawaideh M, Mossa-Basha M, Haynor D et al The utility of deep learning: evaluation of a convolutional neural network for

detection of intracranial bleeds on non-contrast head computed tomography studies. SPIE Medical Imaging, 2019, Proceedings Volume 10949, Medical Imaging 2019: Image Processing; 109493J

10. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ et al (2018) Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. npj Digit Med 1:9
11. Al-Ayyoub M, Alawad D, Al-Darabsah K et al (2013) Automatic detection and classification of brain hemorrhages. WSEAS Trans Comput 10:395–405