Neuroradiology (2022) 64:981–990
https://doi.org/10.1007/s00234-021-02874-w

FUNCTIONAL NEURORADIOLOGY

FDA-approved deep learning software application versus radiologists with different levels of expertise: detection of intracranial hemorrhage in a retrospective single-center study

Thomas Kau[1,2] · Mindaugas Ziurlys[1] · Manuel Taschwer[1] · Anita Kloss-Brandstätter[3] · Günther Grabner[4] · Hannes Deutschmann[5]

Received: 26 August 2021 / Accepted: 1 December 2021 / Published online: 6 January 2022

Abstract

Purpose To assess an FDA-approved and CE-certified deep learning (DL) software application compared to the performance of human radiologists in detecting intracranial hemorrhages (ICH).

Methods Within a 20-week trial from January to May 2020, 2210 adult non-contrast head CT scans were performed in a single center and automatically analyzed by an artificial intelligence (AI) solution with workflow integration. After excluding 22 scans due to severe motion artifacts, images were retrospectively assessed for the presence of ICHs by a second-year resident and a certified radiologist under simulated time pressure. Disagreements were resolved by a subspecialized neuroradiologist serving as the reference standard. We calculated interrater agreement and diagnostic performance parameters, including the Breslow–Day and Cochran–Mantel–Haenszel tests.

Results An ICH was present in 214 out of 2188 scans. The interrater agreement between the resident and the certified radiologist was very high ($\kappa = 0.89$) and even higher ($\kappa = 0.93$) between the resident and the reference standard. The software has delivered 64 false-positive and 68 false-negative results giving an overall sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of 68.2%, 96.8%, 69.5%, 96.6%, and 94.0%, respectively. Corresponding values for the resident were 94.9%, 99.2%, 93.1%, 99.4%, and 98.8%. The accuracy of the DL application was inferior ($p < 0.001$) to that of both the resident and the certified neuroradiologist.

Conclusion A resident under time pressure outperformed an FDA-approved DL program in detecting ICH in CT scans. Our results underline the importance of thoughtful workflow integration and post-approval validation of AI applications in various clinical environments.

Keywords Artificial intelligence · Deep learning · Intracranial hemorrhage · Computed tomography · Diagnostic accuracy

✉ Thomas Kau
thomas.kau@kabeg.at

1 Department of Radiology, Landeskrankenhaus Villach, Nikolaigasse 43, 9500 Villach, Austria

2 Division of Pediatric Radiology, Department of Radiology, Medical University of Graz, Auenbruggerplatz 9, 8036 Graz, Austria

3 Carinthia University of Applied Sciences, Europastrasse 4, 9500 Villach, Austria

4 Department of Medical Engineering, Carinthia University of Applied Sciences, Primoschgasse 8, 9020 Klagenfurt, Austria

5 Division of Neuroradiology, Vascular and Interventional Radiology, Department of Radiology, Medical University of Graz, Auenbruggerplatz 9, 8036 Graz, Austria

Introduction

Deep learning (DL), a type of machine learning, is inspired by human brain structures and is increasingly used for medical image analysis [1]. It has evolved into a state-of-the-art machine learning approach where convolutional neural networks (CNN) are most commonly used [2, 3]. These CNNs are trained with "ground truth" data, which is called supervised learning. During the learning process, image features are automatically extracted using multiple levels of abstraction. After being trained, a CNN can be used to classify new data. Radiological DL solutions might have just passed the peak of exaggerated expectations [4, 5]. However, since peer review is not mandatory for Food and Drug Administration (FDA) approval, we see

a need for external evidence of actual performance to fill the translational gap with implementing AI solutions in real-world clinical scenarios [6–8].

On the other hand, there is evidence of rapidly increasing workloads for radiologists during on-call hours and shortages of radiologists in many countries [9]. The latter is reflected in the upgrowth in teleradiological services. It is mainly CT examinations that contribute to this challenge for hospital staff to maintain quality and safety standards in diagnostic imaging. The models for 24-h coverage of radiological services vary in different hospitals. While the 24/7 presence of a consultant is a practiced standard in some hospitals and countries, overnight resident coverage may be beneficial to the quality of training during the day in case of limited human resources [10]. This is supported by previous studies showing sufficiently low resident overnight error rates [11, 12]. Still, we can assume that error rates increase with case burden, fatigue, and circadian effects [13, 14].

Neuroimaging significantly adds to the total CT workload. With various etiologies including trauma, infarction, and aneurysm rupture, intracranial hemorrhage (ICH) is the second leading cause of stroke worldwide [15]. Because of high early case fatality and few effective interventions, ICH cases require optimal patient management starting with a timely and accurate diagnosis [16]. In most hospitals, CT is the mainstay of emergency neuroimaging. The role of the radiologist in ICH cases is to make the diagnosis, characterize, quantify, and localize the bleeding. Basic information, including the binary decision of whether a hemorrhage is present or not, is derived from a non-contrast CT, the indispensable part of every protocol. Acute ICHs are hyperdense compared to brain tissue and cerebrospinal fluid (CSF). However, making a confident diagnosis is not always straightforward. Especially for tiny or subacute bleedings, radiological performance may be challenged by the limited spatial resolution, pre-existing structural changes, subtle calcifications, beam hardening artifacts, and locations near the skull base.

Individual AI algorithms have been developed to support clinicians in the identification and prioritization of cases suspected to be ICHs [17–19]. So far, only a few vendors have received FDA approval for their solutions. The Aidoc software has been reported to perform with accuracy levels of up to 98%, notably with even higher specificity than sensitivity [20]. Nonetheless, the generalization of different datasets and clinical translation are known challenges restraining convolutional neural networks (CNN) [18, 21, 22].

The objective of our study was to assess the Aidoc software in a diverse clinical setting compared to the performance of human radiologists under simulated time pressure. Specifically, we investigated whether the diagnostic accuracy of this workflow-integrated AI solution was equivalent to that of a second-year resident.

Materials and methods

This was a single-center study designed in accordance with the updated Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement [23]. It was approved by the responsible Federal Ethics Committee, thus waiving the need for patient consent.

Patients

All adult inpatients and outpatients who have had a cranial CT including non-contrast scans (NCCT) from January to May 2020 were enrolled in the study. They were selected from a wide variety of indications in a regional stroke and trauma center in a general teaching hospital. We chose a 20-week period based on statistical power calculations to determine the minimal required total sample. Retrospective inclusion was irrespective of urgency or indication of the examinations and irrespective of whether they were initial or follow-up scans. In all cases, imaging data were automatically analyzed by the AI software program. Exclusion criteria were an incomplete NCCT scan or severe motion-related image artifacts. For the study-related radiological assessment, eligible NCCT scans were anonymized, separated from their reports, and sorted chronologically.

Image acquisition and reconstruction

Depending on resource capacity and urgency, non-contrast cranial CT (NCCT) imaging was either performed using a 2 × 192-detector row dual-source CT scanner (Somatom Force, Siemens Healthineers) or a 128-detector row CT scanner (Somatom Edge Plus, Siemens Healthineers). The following acquisition parameters were used for the dual-energy scanner: collimation, 0.6 mm; spatial resolution, 0.24 mm; rotation time, 0.25 s; table speed, 13.4 cm/s; tube voltage, 80 kV; adaptive tube current; windowing width/level, 35/80; convolution kernel, Hr40s\3. On the single-source scanner, the acquisition parameters were collimation, 0.6 mm; spatial resolution, 0.30 mm; rotation time, 0.28 s; table speed, 21.1 cm/s; tube voltage, 120 kV; adaptive tube current; windowing width/level, 40/90; convolution kernel, Hr40s\3. Only axial images with a reconstructed section thickness of 5 mm (January–March) or 3 mm (March–May) were selected for further assessment. Radiologists were allowed to manually change window settings during the assessment.

The AI software application

The Aidoc software application, FDA class II approved and CE-certified class I, was implemented in the radiological workflow for a clinical trial [20]. This included the integration into our picture archiving and communication system (PACS). The commercial solution is offered as a triage and notification software program that flags and communicates suspected ICHs on NCCTs (www.aidoc.com) [23]. Referring to the company's literature, the convolutional neural network had been trained and tested on CT scans from 9 medical centers and 17 different CT machines [18, 20]. The ground truth labeling structure varied with hemorrhage type and size. It included both weak and strong labeling schemes with study-level classification for diffuse subarachnoid hemorrhage, slice-level bounding boxes around indistinct extraaxial and intraaxial hemorrhages, and pixel-level semantic segmentation of well-defined intraparenchymal hemorrhages [20]. According to a non-peer-reviewed publication, the software application had also been tested on 7112 NNCTs from two large urban academic and trauma centers with a dataset different from that used for the algorithm training process [20]. The authors reported sensitivity, specificity, and accuracy levels of 95%, 99%, and 98%, respectively, for purely AI-based detection of ICH [20, 23]. The software application does not provide any notification of inadequate data quality.

For the purpose of our study, the source images of all non-contrast CT examinations showing the entire neurocranium were automatically and immediately forwarded by the application for analysis [22]. They were sent without any context data of the study. In the case of a positive result, the application sent a notification to a widget on the radiologist's desktop and added a preview of the abnormal key images to the existing study in the PACS system. The workflow and integration have been described elsewhere in the literature [18]. All cases meeting the above-mentioned inclusion criteria were retrospectively included in this assessment.

Radiological assessment/reference standard for proof of diagnosis

Generally, the gold standard for the detection of intracranial hemorrhage is an expert neuroradiological diagnosis achieved by the visual inspection of CT images with the support of a software tool for density measurements. In order to best simulate an acute setting, only native axial CT scans were presented in close succession to the human readers who were asked to document their diagnosis within 30 s for each scan. A resident who had completed his first year of training with 1100 supervised cranial CT reports and a radiologist who has passed board-examination with 5 years of experience in neuroradiology was asked to make a binary decision (hemorrhage/no hemorrhage) separately for each examination. Disagreements were resolved by a radiologist with 15 years of experience and additional qualifications in neuroradiology, who also provided further analysis of discordant results. A CT scan was considered positive for intracranial hemorrhage if there were visually detectable hyperdensities (> 60 Hounsfield units) of any size or number with no better explanation (i.e., calcification, beam hardening artifacts, intact vascular structures, colloid cyst) [24]. All three radiologists were aware of the patients' age and sex but were not provided any data relating to the history, the indication of examination, further clinical information, and the AI algorithm output. The PACS file folder created for the fast retrospective analysis did not allow any conclusions to be drawn from any contextual factors.

Statistical analysis

Statistical analyses were performed using IBM SPSS version 27 and the pROC package version 1.17.0.1 from the free software environment R (https://www.r-project.org/). All data, irrespective of whether flagged by the software application as potentially positive for acute intracranial hemorrhage or not, were forwarded on a per-examination basis. We calculated the interrater agreement between human reviewers by using Cohen's κ. It was interpreted as follows: less than or equal to 0.20, poor; 0.21–0.40, slight; 0.41–0.60, moderate; 0.61–0.80, good; 0.81–0.90, very good; and 0.91–1.0, excellent. Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values, and accuracy in the detection of ICH on NCCT were computed for the software application as well as for the resident. To test for homogeneity of the odds ratio between the contingency tables of the software and the resident, a Breslow–Day test was applied. To test for conditional independence between the software and the resident contingency tables, a Cochran–Mantel–Haenszel test was applied.

Results

During the 20-week review period, 212 out of 2210 NCCT scans analyzed were flagged by the Aidoc application for the presence of ICHs (Fig. 1). After excluding 22 examinations – two of which had been flagged by the application – due to severe motion artifacts, 2188 scans in 1782 patients (921 females, 861 males) were eligible for further analysis. Consequently, according to the software widget, the rate of positive cases (210/2188) in our series was 9.6%. The resident found 218 scans (10.0%) to be positive for ICH; the certified radiologist found 211 scans (9.6%) with a very good interrater agreement of $\kappa = 0.89$. The

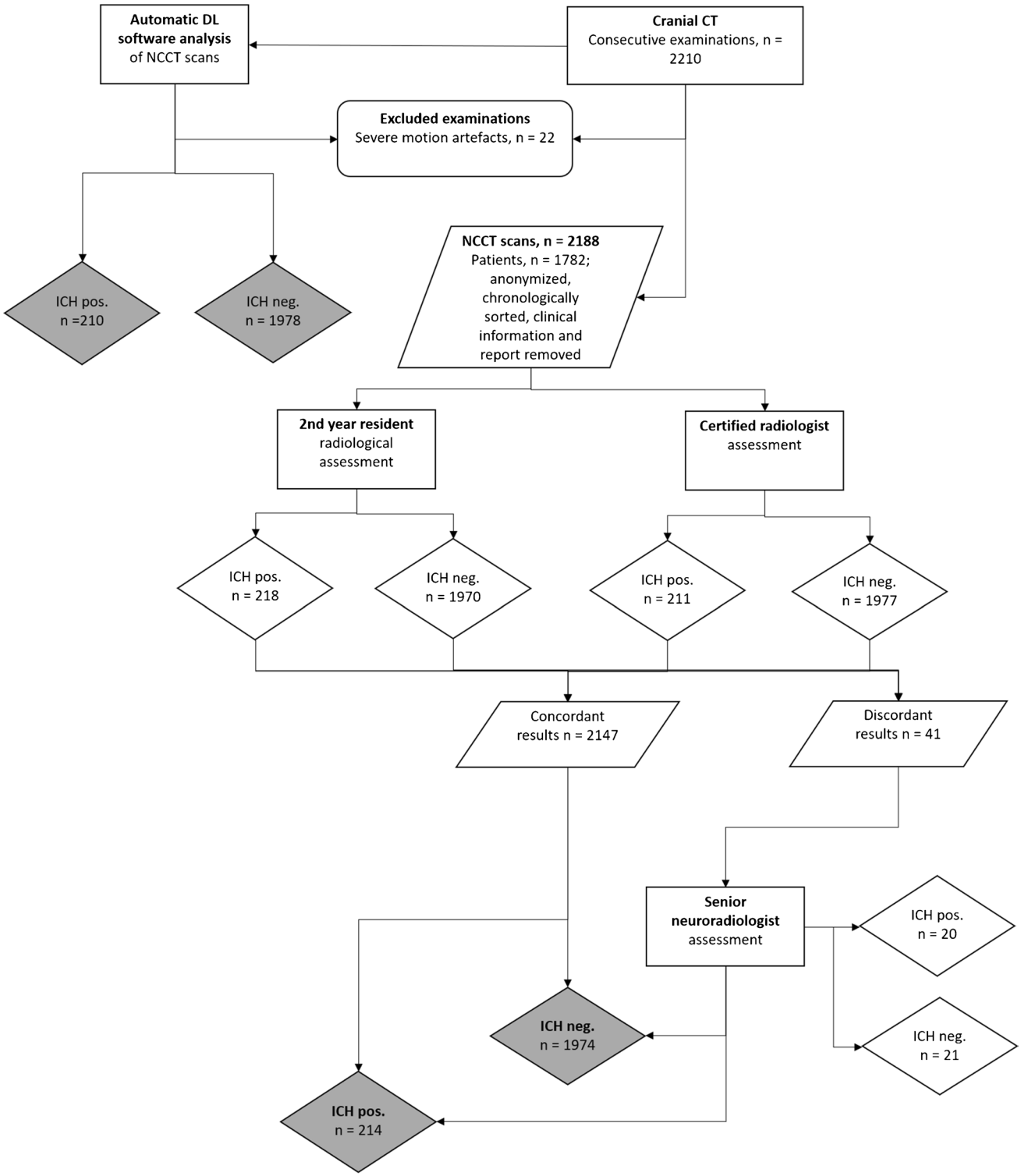

Fig. 1 Overview of study workflow and main results

agreement was excellent between the resident and the neuroradiologist ($\kappa = 0.93$) as well as between the certified radiologist and the neuroradiologist ($\kappa = 0.96$). The neuroradiologist resolved 41 discordant results out of 2188 cases. In 2 of 64 cases with false-positive AI results, a qualified radiologist was overruled by the subspecialized neuroradiologist. In 3 cases, the resident was overruled. In 5 cases (0.2%), the software application indicated subtle

ICHs that were overlooked or interpreted differently by the resident.

In conclusion, the radiological assessment detected ICHs in a total of 214 examinations (9.8%). Measured against the reference standard, the initial automatic DL software analysis delivered 64 false-positive and 68 false-negative results. This results in overall sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of 68.2%, 96.8%, 69.5%, 96.6%, and 94.0%, respectively, for the software-based detection of ICH. In contrast, the resident achieved respective accuracy values of 94.9%, 99.2%, 93.1%, 99.4%, and 98.8%; the certified radiologist achieved values of 95.8%, 99.7%, 97.2%, 99.5%, and 99.3%, respectively (Table 1).

Based on the comparison of their contingency tables and measured against the reference standard, the diagnostic accuracies of the resident and the software differed highly significantly ($p < 0.001$). The flow diagram provides an overview of the study workflow and the main results (Fig. 1).

Beam hardening artifacts and relatively hyperdense brain structures, partly in the vicinity of a chronic ischemic infarct, were the main cause of false positives, followed by calcifications (falx cerebri, choroid plexus, or unspecific cerebral), dural thickening long after osteoplastic craniotomy, tumor, dense venous sinus or tentorium, "dual-energy " artifacts, hyperdense vessels in a subacute ischemic infarct, and chronic subdural hematoma (SDH) (Table 2; Fig. 2). In eight false-positive cases (12.5%) affecting six patients, the software application produced false-positive results in at least one (1–3) previous scan.

Radiologically confirmed ICHs affecting males in 30 cases and females in 38 cases at a mean age of 72 years were not flagged by the Aidoc application. They represented SDH in 35 cases, primarily subarachnoid bleeding in nine cases, and intracerebral hemorrhage in 24 cases (Fig. 3). Of those, six cases were due to a hemorrhagic transformation of an ischemic infarct, and three were interpreted as tumor bleeding. The mean short diameter of unmarked hematomas was 0.9 cm (0.1–5.7 cm; median, 0.5 cm). First-time imaging produced a false-negative software result in 31 cases, while a follow-up examination was affected in 37 cases. In eight cases, the software application provided discrepant results in consecutive examinations of a specific patient (Fig. 4). In two examinations of a single patient, the software-flagged calcifications of the falx while failing to mark an SDH.

The software application detected five ICHs not spotted by the resident. Discrepancies between the resident and the neuroradiologist were caused in nine cases by beam hardening artifacts; in four cases due to discordant assessment of subdural bleeding; in three cases related to subtle subarachnoid hemorrhage (SAH); in three cases due to dural thickening after osteoplastic craniotomy; and relating to the following conditions in one case each: intracerebral hematoma, vascular malformation, brain tumor, dense venous sinus, and ischemic infarct with possible hemorrhagic transformation. Discrepancies between the certified radiologist and the neuroradiologist occurred in six cases of SDH; in three cases of subtle tumor hemorrhage or calcification, respectively; in two cases of dense sinus or tentorium; and in each one case related to beam hardening, tiny cortical bleeding, and calcifications of the falx or brain parenchyma.

Table 1 Contingency tables for the software, the certified radiologist, and the resident, each with reference to the neuroradiologist

		Reference (neuroradiologist)	
		Hemorrhage	No hemorrhage
Software	Hemorrhage	146	64
	No hemorrhage	68	1910
Resident	Hemorrhage	203	15
	No hemorrhage	11	1959
Radiologist	Hemorrhage	205	6
	No hemorrhage	9	1968

Table 2 Corresponding image content and frequency of false-positive AI software results

Correlate	*n*
Beam hardening artifacts	16
Falx calcification	13
Chronic ischemic brain infarct	10
Post craniotomy dural thickening	9
Meningeoma	5
Calcified choroid plexus	4
Cerebral tumor or metastasis	4
Dense venous sinus	3
Dense tentorium	3
Dual-energy artifacts	2
Normal brain tissue	2
Hyperdense vessels in a subacute ischemic infarct	1
Chronic subdural hematoma	1
Unspecific cerebral calcification	1

Discussion

In our study of over 2000 cases, the diagnostic accuracy of an FDA-approved and CE-certified deep learning software application for the detection of ICHs in routine clinical practice was significantly inferior to that of a certified radiologist and a second-year resident under simulated stress conditions. Two competing aspects must be taken into account after a thorough analysis of the results. On the one hand, false positives of the software tool were almost

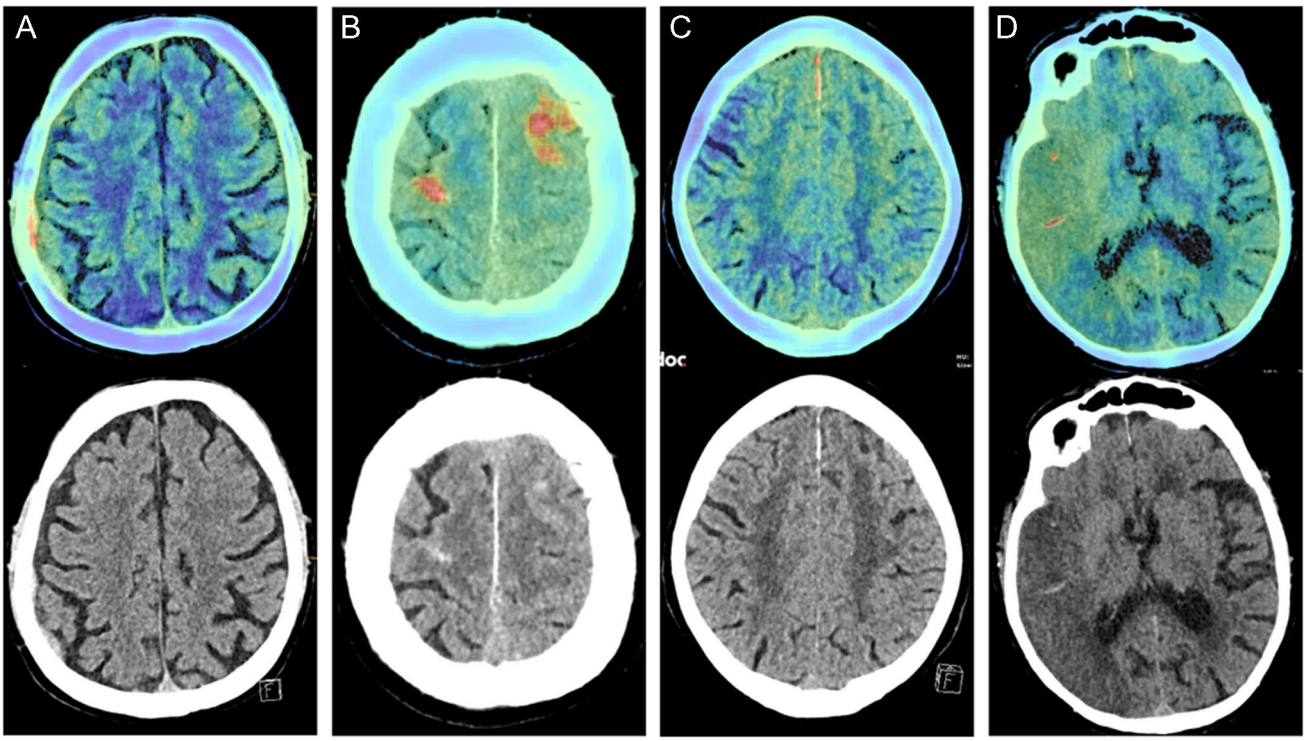

Fig. 2 Marked non-contrast CT images versus untagged equivalents. By tagging suspicious image content (*red overlay*), the deep learning software correctly detected intracranial hemorrhage in some cases (**A**, **B**) and produced false-positive results in others (**C**, **D**). (**A**) Post-traumatic epidural hematoma. (**B**) Bilateral subarachnoid hemorrhage. (**C**) Calcification of the falx instead of subdural bleeding. (**D**) Hyperdense peripheral segments of the middle cerebral artery in demarcated subacute ischemic stroke, not compatible with hemorrhagic transformation

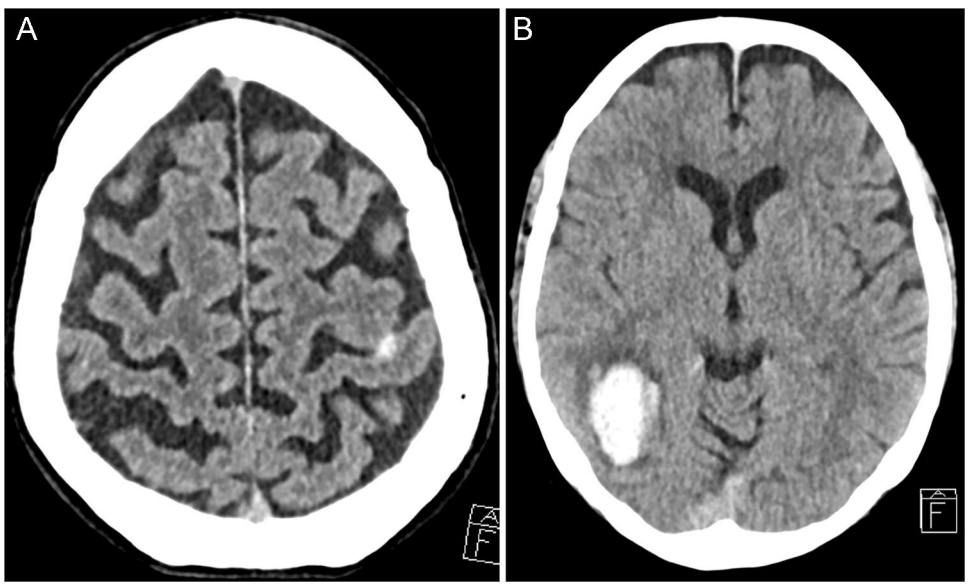

Fig. 3 Imaging findings that did not trigger an alert by the DL software. (**A**) Subtle subarachnoid bleeding in the central sulcus. (**B**) Subacute intracerebral hematoma in the right occipito-temporal region. The results generated by the software were rated as false negative in both cases

exclusively findings that an experienced radiologist should be able to reject with a high degree of certainty. On the other hand, a remarkable rate of false negatives was noted, partly related to scans showing large hematomas. Interestingly, in some cases, the software application marked a hemorrhage while failing to detect it during the follow-up exam, or vice versa.

False-negative software results are the most critical issue for clinical implementation. In the hypothetical scenario of an AI-supported case preselection, the limited NPV would

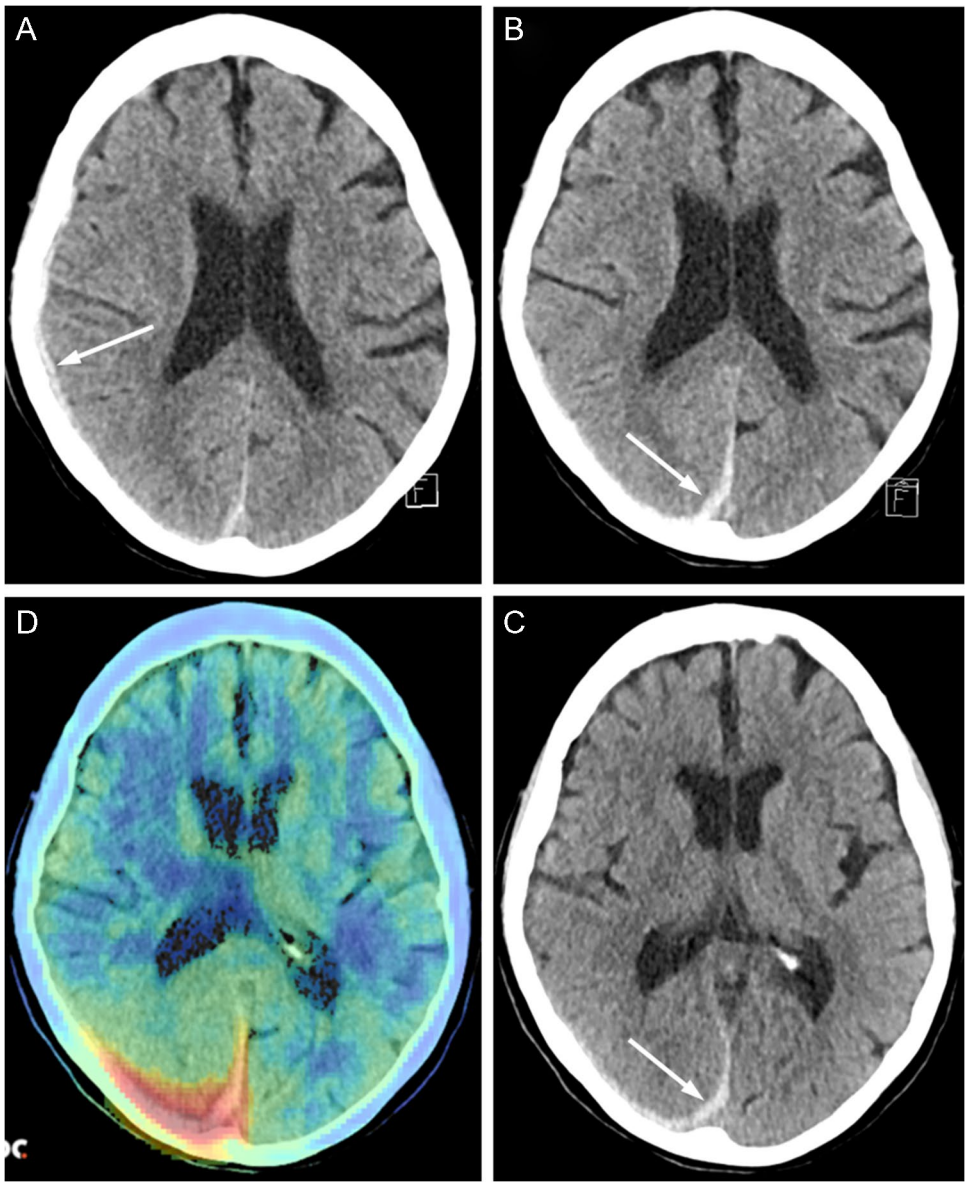

Fig. 4 Serial imaging in a single patient with a thin subdural hematoma (*arrows*). False-negative software results in two consecutive CT scans (**A**, **B**), followed by a true positive result (**C**, **D**, *marked in red*)

mean a critical reduction of reliability. The data were verifiably transferred to the software application in all relevant cases. The retrospective data tracking of false-negative results, however, could not differentiate with certainty whether the imaging data of a specific case were 'stuck' in the orchestrator or whether the AI-based analysis did not recognize an ICH as such. This is in accordance with a very recent study using the same decision support software [22]. It must therefore remain open to what extent the DL algorithm per se was responsible for a failure to notify. Ultimately, it is more of developmental importance than of clinical importance to make a distinction between lack of recognition of an imaging data set and incorrect assessment of the image content. With the clinical value in mind, we see a special need for further analysis of the restricted sensitivity in this AI application.

Beam hardening artifacts were the most common source of false-positive results. This was true for the DL tool as well as for the resident. Interestingly, not only calcifications of the cerebral falx and choroid plexus but also cortical structures in close vicinity to chronically infarcted tissue were quite frequently flagged by the software application. Dural thickening late after osteoplastic craniotomy provided another challenge for both the machine and the human reader. Apart from different proportions, our findings are in concordance with the etiology of false positives as published by Voter et al. [22]. They reported decreased diagnostic performance in association with prior neurosurgery and type and number of hemorrhages but not with image quality.

It goes without saying that the limitations of an AI algorithm only coincide to a limited extent with the challenges that a human radiologist must overcome. For example,

choroid plexus calcifications are rarely misinterpreted as hemorrhages by human readers, and those false-negative software results related to large cerebral hematomas will unlikely be missed by a radiological resident. On the other hand, the AI tool is impervious to fatigue or loss of concentration and may be particularly helpful in alerting the radiologist to a narrow, high SDH. Especially in the subacute phase, the latter can easily be overseen on axial images due to partial volume effects.

Ultimately, the combination of man and machine will most likely achieve the highest diagnostic accuracy. Our results support the expectation that well-integrated algorithms should be further improved to assist radiologists, especially in high-output situations and during on-call hours [9]. O'Neill et al. recently reported that AI-assisted reprioritization of the reading worklist was beneficial in terms of turnaround time, especially for examinations ordered as routine [19]. It remains to be seen which frequent tasks will be integrated into neuroradiological solutions over time, and to what extent rare differential diagnoses may move into the focus of development.

The studied software application has been one of the first AI programs to receive FDA approval [20]. It is meant to aid radiologists in the identification and prioritization of CT scans suspicious for ICH in an urgent context [19]. According to the developers, the algorithm had been tested on CT scans from two trauma centers which were different from those used for ground truth labeling [20]. Initially published performance characteristics with sensitivity and specificity values of 95% and 99%, respectively, were promising [20]. In their prospective single-reader assessment with a focus on software-flagged cases, Ginat et al. reported an overall sensitivity, specificity, PPV, NPV, and accuracy of 88.7%, 94.2%, 73.7%, 97.7%, and 93.4%, respectively, with significantly higher accuracy for emergency scans than for inpatient scans [18]. With corresponding performance metrics of 68.2%, 96.8%, 69.5%, 96.6%, and 94.0%, respectively, our results slightly lag behind the data from this previous study. Based exclusively on emergent cases with a similar ICH rate, Voter et al. also found a lower PPV (81.3%) than previously reported raising concerns about the generalizability of this DL algorithm [22]. Their radiomic feature analysis failed to reveal significant differences between concordant and discordant cases.

Despite simulated time-constraint, the resident in our own study significantly outperformed the FDA-approved DL solution in all parameters. While the frequency of its erroneous warnings appears to be acceptable, the PPV achieved by the Aidoc application calls for close control of unflagged scans. We believe that, prior to market approval, AI solutions like the one under investigation need more and clinically diverse training data derived from different scanners in multiple institutions. Site-specific training of diagnostic decision support systems, currently not permitted by the FDA, may be another movement toward improving CNN performance [25].

There were limitations to our study. Firstly, stress conditions in clinical practice may not only be due to time pressure but also due to fatigue, technical overload, lack of concentration, and other factors [14]. Nevertheless, we decided to create a reader study taking into account a realistic scenario with limitations of both time and neuroradiological experience. Secondly, we needed to consider a certain selection bias since ICH cases were over-represented due to serial follow-up. This, in turn, corresponded to clinical needs, therefore reflecting routine practice. Thirdly, for the purpose of comparability with the software, equivocal decisions were not an option for human readers. Fourthly, the readers were blind to previous imaging and clinical information, which may be only partially in contrast to routine conditions. Fifthly, as with many CT studies, patients under 18 years of age were excluded. Sixthly, despite standardized training content and therefore representative radiological knowledge of the single resident, our study results must be interpreted in the context of his individual circumstances. Finally, with respect to the integrated software widget, lacking notifications were interpreted as false-negative results without further interpretability, which reflects the inherent black box problem of AI systems [22].

Nagendran et al. recently pointed out that for the sake of patient safety and appropriateness of health expenses, research efforts in the field of AI should focus more on real-world clinical scenarios [8]. On the other hand, developers and companies affirm the challenge of translating AI into clinical practice [26]. Our large-scale study adds to the evidence on the actual performance of an FDA-approved DL tool. It underlines the usefulness of collaboration between healthcare providers and the industry for validation after regulatory approval. Moreover, radiologists will need to learn from their experience to recognize AI alerts, critically evaluate them, check their plausibility, appraise their relevance, and integrate them into the final decision-making process [27]. Even if not yet fully defined, AI will, with high probability, play a major role in future diagnostic support systems.

Conclusion

Our study adds to the body of evidence required for the implementation of AI solutions in real-world scenarios. We assessed an FDA-approved DL software application for ICH detection in a routine setting compared to the performance of human radiologists under time-constrained study conditions. A second-year resident outperformed the AI tool in terms of both sensitivity and specificity. Since most

erroneous alerts can be resolved by experienced radiologists, the software holds promise for prioritizing cranial CT exams. However, due to a notable rate of unflagged ICH scans, we doubt generalizability and recommend this AI solution be improved. Our results underline the need for external post-approval validation in various clinical environments. They warrant further research with a focus on the combination of human and artificial intelligence for an accurate and timely diagnosis of ICH.

Acknowledgements The authors thankfully acknowledge the IT support given by Gerhard Orlitsch, Hannes Mössler, and Markus Kurej. No external funding or study support by a company has to be disclosed.

Author contribution Thomas Kau: made substantial contributions to the conception and design of the work as well as to the interpretation of data; drafted the work; approved the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Mindaugas Ziurlys: made substantial contributions to the acquisition of data; revised the work critically for important intellectual content; approved the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Manuel Taschwer: made substantial contributions to the acquisition of data; revised the work critically for important intellectual content; approved the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Anita Kloss-Brandstätter: made substantial contributions to the analysis and interpretation of data: revised the work critically for important intellectual content: approved the version to be published: agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Günther Grabner: made substantial contributions to the design of the work and interpretation of data; revised the work critically for important intellectual content; approved the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Hannes Deutschmann: made substantial contributions to the design of the work and interpretation of data; revised the work critically for important intellectual content; approved the version to be published; agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval The study was performed with approval from the local ethics committee (S2020-32). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

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

References

1. Zaharchuk G, Gong E, Wintermark M et al (2018) Am J Neuroradiol 39:1776–1784
2. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
3. Chan HP, Samala RK, Hadjiiski LM et al (2020) Deep learning in medical image analysis. Adv Exp Med Biol 1213:3–21
4. Banja J (2020) AI hype and radiology: a plea for realism and accuracy. Radiol Artif Intell 2:e190223
5. Desai AN (2020) Artificial intelligence: promise, pitfalls, and perspective. JAMA 323:2448–2449
6. Rockall A (2020) From hype to hope to hard work: developing responsible AI for radiology. Clin Radiol 75:1–2
7. Bluemke DA, Moy L, Bredella MA et al (2020) Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers – from the Radiology editorial board. Radiology 294:487–489
8. Nagendran M, Chen Y, Lovejoy CA et al (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368:m689
9. Bruls RJM, Kwee RM (2020) Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging 11:121
10. Bruno MA, Duncan JR, Bierhals AJ, Tappouni R (2018) Overnight resident versus 24-hour attending radiologist coverage in academic medical centers. Radiology 289:809–813
11. Cooper VF, Goodhartz LA, Nemcek AA Jr, Ryu RK (2008) Radiology resident interpretations of on-call imaging studies: the incidence of major discrepancies. Acad Radiol 15:1198–1204
12. Mellnick V, Raptis C, McWilliams S, Picus D, Wahl R (2016) On-call radiology resident discrepancies: categorization by patient location and severity. J Am Coll Radiol 13:1233–1238
13. Terreblanche OD, Andronikou S, Hlabangana LT, Brown T, Boshoff PE (2012) Should registrars be reporting after-hours CT scans? A calculation of error rate and the influencing factors in South Africa. Acta Radiol 53:61–68
14. Ruutiainen AT, Durand DJ, Scanlon MH, Itri JN (2013) Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight. Acad Radiol 20:305–311
15. 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
16. Steiner T, Al-Shahi Salman R, Beer R, Christensen H et al (2014) European Stroke Organisation (ESO) guidelines for the management of spontaneous intracerebral hemorrhage. Int J Stroke 9:840–855
17. Chilamkurthy S, Ghosh R, Tanamala S et al (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392:2388–2396
18. Ginat DT (2020) Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. Neuroradiology 62:335–340
19. O'Neill TJ, Xi Y, Stehel E et al (2020) Active reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head CT with

intracranial hemorrhage. Radiol Artif Intell 3(2):e200024. https://doi.org/10.1148/ryai.2020200024
20. Ojeda P, Zawaideh M, Mossa-Basha M, Haynor D (2019) The utility of deep learning: evaluation of a convolutional neural network for detection of intracranial bleeds on non-contrast head computed tomography studies. Proc. SPIE 10949, Medical Imaging 2019: Image Processing, 109493J. https://doi.org/10.1117/12.2513167
21. 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
22. Voter AF, Meram E, Garrett JW, Yu JJ (2021) Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. J Am Coll Radiol S1546–1440(21):00227–00231
23. Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology 277:826–832
24. Parizel PM, Makkat S, Van Miert E, Van Goethem JW, van den Hauwe L, De Schepper AM (2021) Intracranial hemorrhage: principles of CT and MRI interpretation. Eur Radiol 11:1770–1783
25. US FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. Available at: https://www.fda.gov/media/122535/download. Published April 2, 2019. Accessed 15 June 2021
26. Tariq A, Purkayastha S, Padmanaban GP et al (2020) Current clinical applications of artificial intelligence in radiology and their best supporting evidence. J Am Coll Radiol 17:1371–1381
27. Simpson SA, Cook TS (2020) Artificial intelligence and the trainee experience in radiology. J Am Coll Radiol 17:1388–1393