Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263984

NEURORADIOLOGY

ORIGINAL ARTICLE

Real-world multimetric comparison of four commercial artificial intelligence solutions for intracranial hemorrhage detection

Jimin Kim[1]
Ha Young Lee[1]
Jin Eun[2]
Sanghyuk Im[2]
Eun Jeong Min[3]
Se Won Oh[1]

[1]Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Department of Radiology, Seoul, Republic of Korea

[2]Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Department of Neurosurgery, Seoul, Republic of Korea

[3]The Catholic University of Korea, College of Medicine, Department of Medical Life Sciences, Seoul, Republic of Korea

PURPOSE

To evaluate the real-world multimetric performance of four commercially available computed tomography (CT)-based artificial intelligence (AI) solutions for acute intracranial hemorrhage (AIH).

METHODS

Patients who underwent non-contrast brain CT for suspected AIH in our emergency room between February and March 2024 were screened. After applying the inclusion and exclusion criteria, 436 CT scans were included in the final analysis. Three neuroradiologists established the ground truth for AIH and hemorrhage volume. For detection performance, the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and Brier score were calculated based on the available probability score, whereas sensitivity, specificity, precision, and F1 score were calculated based on binary classification. Bland–Altman analysis was performed to assess volumetric agreement for AIH between each algorithm's calculations and the neuroradiologists' measurements.

RESULTS

A total of 436 patients (mean age, 62 years ± 20; male patients, 209) were enrolled. The AUROC (0.96 to 0.99) and sensitivity (0.85 to 0.92) were high across all solutions, with no statistically significant differences in pairwise comparisons ($P > 0.05$). However, solution B demonstrated the highest AUPRC [0.98, 95% confidence interval (CI): 0.94, 1.00] and the lowest Brier score [0.02 (95% CI: 0.02, 0.03)]. In binary performance, both solutions B and D exhibited significantly higher specificity (1.00 and 0.99), precision (0.90 to 0.98), and F1 score (0.87 to 0.94) than the other solutions ($P < 0.05$). For volumetric agreement of AIH, solution D showed the lowest mean difference [−0.87 mm^3 (95% CI: −1.47, −0.27)] and the narrowest limits of agreement (−13.4 to 11.6) relative to the neuroradiologists' measurements.

CONCLUSION

In a real-world emergency setting, all four commercially available CT-based AI solutions for AIH demonstrated uniformly excellent performance; however, meaningful differences emerged in confirmatory performance and volumetric agreement. These distinct, algorithm-specific trade-offs provide practical guidance for selecting and integrating appropriate AI solutions to improve AIH diagnosis and management workflows.

CLINICAL SIGNIFICANCE

The algorithm-specific performance trade-offs identified in this study suggest that no single AI solution is universally optimal; solutions with superior confirmatory performance may reduce unnecessary notifications in high-volume emergency settings, whereas those with more consistent volumetric agreement may better support treatment planning and longitudinal monitoring. A structured, multimetric evaluation aligned with institutional priorities is essential for evidence-based AI procurement in acute stroke imaging.

KEYWORDS

Acute intracranial hemorrhage, computed tomography, deep learning, artificial intelligence, performance

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Se Won Oh

E-mail: oasis1979@gmail.com

Received 18 March 2026; revision requested 26 April 2026; accepted 29 June 2026.

Epub: 17.07.2026

Publication date:

DOI: 10.4274/dir.2026.263984

You may cite this article as: Kim J, Lee HY, Eun J, Im S, Min EJ, Oh SW. Real-world multimetric comparison of four commercial artificial intelligence solutions for intracranial hemorrhage detection. *Diagn Interv Radiol*. 17 July 2026 DOI: 10.4274/dir.2026.263984 [Epub Ahead of Print].

The rapid expansion of commercially available artificial intelligence (AI) applications in radiology has led to the widespread clinical availability of stroke-focused algorithms, particularly for the detection of acute intracranial hemorrhage (AIH) on non-contrast computed tomography (CT).[1-3] As an increasing number of AI solutions receive regulatory approval and reimbursement pathways expand,[4] hospitals operating under constrained resources are faced with a growing range of options and must evaluate and select the most appropriate tools for integration into acute stroke workflows.[5-8]

AIH represents a time-sensitive neurologic emergency.[9] Rapid detection is essential for initiating appropriate management strategies, including blood pressure control, anticoagulation reversal, neurosurgical consultation, and intensive monitoring.[10-12] Beyond binary detection, hemorrhage volume is a well-established prognostic marker that plays a central role in risk stratification, treatment decision-making, and longitudinal assessment. Accordingly, AI tools in this domain are positioned not only as triage systems but also as quantitative decision-support instruments.[4-6,13,14]

Despite the proliferation of commercial AI solutions, independent head-to-head comparisons under real-world clinical conditions remain critically lacking.[15] Procurement decisions are frequently informed by vendor-reported performance metrics or single-metric validation studies, which do not fully reflect clinical priorities such as discriminative ability, confirmatory performance, calibration quality, and quantitative agreement. Moreover, different algorithms may be optimized for distinct tasks, raising the possibility that they are not functionally interchangeable.[16-19]

Therefore, we performed an external, real-world, multimetric evaluation of four CT-based AI solutions for AIH that were approved and designated as innovative medical devices by the Korean Ministry of Food and Drug Safety. By assessing both discrimination performance and volumetric agreement, including subtype-specific analyses, this study aims to provide practical, evidence-based guidance for the selection and clinical integration of AI tools in acute stroke imaging.

Main points

- Commercial artificial intelligence (AI) tools for detecting acute intracranial hemorrhage (AIH) on brain computed tomography (CT) are becoming increasingly available, yet real-world comparisons among algorithms remain limited.
- A multimetric comparison of four commercially available CT-based AI solutions was conducted to evaluate detection performance and volumetric agreement using real-world data.
- The solutions demonstrated distinct, task-specific performance, with algorithm-specific trade-offs among discriminative performance, confirmatory performance, and volumetric agreement.
- This study provides practical guidance for selecting and integrating the most appropriate AI solution according to institutional priorities to improve AIH diagnosis and management workflows.

Methods

This retrospective study was performed in accordance with the principles of the Declaration of Helsinki and approved by Eunpyeong St. Mary's Hospital's Institutional Review Board (IRB; protocol number: PC25RISI0154, approval date: 14 August 2025). The requirement for informed consent was waived by the IRB because of the retrospective nature of the study.

Sample eligibility

A total of 615 non-contrast brain CT scans of adult patients with suspected AIH, including intraparenchymal hemorrhage (IPH), subarachnoid hemorrhage (SAH), subdural hemorrhage (SDH), epidural hemorrhage, and intraventricular hemorrhage (IVH), obtained between February and March 2024 in Eunpyeong St. Mary's Hospital emergency room were eligible for inclusion. During this period, eligible scans were included using the following criteria: (1) the first CT scan was performed during the patient's clinical course, (2) the cranial structure was preserved, and (3) the CT image quality was acceptable for radiologist interpretation, which was defined as the absence of significant motion or metal artifacts that would preclude diagnostic evaluation, consistent with prior validation studies that excluded non-diagnostic-quality images to ensure a fair and standardized comparison across the evaluated AI solutions.[8,20-22] The exclusion criteria were as follows: (1) mixed-stage intracranial hemorrhage, considering the intended use of the enrolled AI solutions; (2) patients with intracranial surgical materials (e.g., clips, coils, or drainage catheters); and (3) intracranial hemorrhages caused by other co-pathologic factors (e.g., tumor bleeding).[8,20-22] All eligible CT scans were reviewed by a board-certified neuroradiologist (J.K., with 12 years of relevant experience) according to the inclusion and exclusion criteria. After review, 63 follow-up scans, 15 scans of patients who had undergone craniectomy, and 21 scans with significant motion artifacts were excluded. Among the potentially eligible 516 scans, 7 scans presenting mixed-stage intracranial hemorrhage, 71 scans with significant metal artifacts, and 2 scans presenting tumor hemorrhage were excluded. Finally, 436 non-contrast brain CT scans were included in this study.

Computed tomography scanning protocol

CT scans were performed using one of two CT machines at the study institution. Machine A was a 128-slice single-source CT scanner (SOMATOM Edge, Siemens Healthineers, Erlangen, Germany) with a tube potential range of 70–140 kVp and a tube current range of 20–800 mA, and machine B was a dual-source CT scanner (SOMATOM Force, Siemens Healthineers) with a tube potential range of 70–150 kVp and a tube current range of 20–1,300 mA. The acquisition parameters were as follows: slice thickness, 4 mm without a gap; rotation time, 1.0 s; pitch, 1; automatic tube voltage modulation (CARE kV, Siemens Healthineers) using a reference kV of 120; automatic tube current selection (CAREDose 4D, Siemens Healthineers) using a reference mAs of 250; and collimation, 128 × 0.6 for machine A and 192 × 0.6 for machine B.

Commercially available artificial intelligence software

In this study, four commercially available CT-based AI solutions for detecting AIH (solution A, JLK-ICH, version 4.1.0.2, JLK Inc.; solution B, HyperInsight–ICH, version 2.1.3, Purple AI Inc.; solution C, Heuron-ICH, version 1.0.5.13, Heuron Inc.; and solution D, AVIEW NeuroCAD, version 1.1, Coreline Soft Inc.) were used based on predefined criteria to reflect both regulatory status and real-world clinical applicability. Specifically, the included solutions (1) were designated as innovative medical devices by the Korean Ministry of Food and Drug Safety or had equivalent regulatory approval, (2) were practically accessible and deployable within our institutional environment at the time of the study, and (3) had supporting evidence from clinical validation studies.[6-8] All four solutions were designated as innovative medical devices by the Korean Ministry of

Food and Drug Safety, and solutions A, B, and C were additionally approved by the U.S. Food and Drug Administration.[22-24]

All AI solutions were deployed on-premises within Eunpyeong St. Mary's Hospital's secure internal network environment. CT images were processed locally without transfer to external servers or vendor platforms. The AI analyses were performed retrospectively using anonymized datasets, and each solution was applied under standardized conditions to ensure a fair comparison.

Solutions A, B, and C provided binary classifications with probability scores and calculated volumes for AIH. However, solution D provided only binary classification without probability scores. Detailed information on the algorithms used in each AI solution is provided in the Supplementary Material. None of the CT images included in this study were used for the training, validation, or internal testing of any of the evaluated AI solutions.

Ground truth and volume measurement for acute intracranial hemorrhage

To establish the ground truth for AIH, three board-certified neuroradiologists (J.M., S.W.O., and H.Y.L., with 12, 18, and 20 years of experience in brain imaging, respectively) independently reviewed the identical set of 436 non-contrast brain CT scans. The three neuroradiologists diagnosed AIH based solely on CT findings and were blinded to the patients' clinical information and the results of the AI solutions. In cases of disagreement, the ground truth was determined by consensus with reference to other available imaging studies. AIH volume in each case was also independently measured by the three neuroradiologists using manual free drawing in ITK-SNAP (version 4.0.1, http://www.itksnap.org). For objective evaluation of AIH, the Hounsfield unit (HU) cut-off for AIH was > 30 HU.[25,26] Therefore, a segmentation map for AIH with a threshold of 30–150 HU was provided to each neuroradiologist before initiating manual drawing. Each neuroradiologist edited the provided AIH segmentation map using the free-draw function. The reference standard for AIH volume was defined as the mean of the three independent measurements.

Statistical analysis

Considering the exploratory and retrospective head-to-head comparison of four AI algorithms in this study, the sample size was determined by the number of eligible non-contrast head CT examinations available during the study period, which was substantially greater than the preliminary sample size calculation based on a significance level of 0.05, a statistical power of 0.8, a specificity of 0.90 from a previous meta-analysis, a specificity of 0.98 from previous validation studies, and a dropout rate of 10%, yielding a minimum required sample size of 202 cases.[5-8,27]

To evaluate detection performance, a two-step analysis was conducted. First, an analysis based on probability scores was performed among solutions A, B, and C because solution D did not provide probability scores. Second, an analysis based on binary classification was performed among solutions A, B, C, and D based on their reported binary results. For the first analysis, the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were calculated. In addition, the Brier score was calculated to compare the overall quality of probability predictions across algorithms under identical conditions. For the second analysis, binary classification metrics, including sensitivity, specificity, precision, negative predictive value, false-positive ratio, false-negative ratio, and F1 score, were calculated. These performance comparisons were also conducted between the traumatic and non-traumatic AIH cohorts in a sub-analysis. In addition, a sub-analysis of detection-failure cases by solution was conducted. AIH volumes were compared between true-positive and false-negative cases. In addition, considering existing studies,[28-30] the false-negative rate was compared between the small (< 10 mm^3) and large (> 10 mm^3) AIH groups.

To evaluate the solutions' AIH volume calculation, Bland–Altman analysis was performed to assess volumetric agreement between each solution's calculation and the neuroradiologists' measurement. Inter-reader agreement of volume measurement among the three neuroradiologists was evaluated using the intraclass correlation coefficient (ICC). The ICCs were classified as follows: < 0.2, very poor; ≥ 0.2 and < 0.4, poor; ≥ 0.4 and < 0.6, moderate; ≥ 0.6 and < 0.8, good; and ≥ 0.8, excellent.

The baseline characteristics, including age, sex, AIH incidence, Glasgow Coma Scale scores, and modified Rankin Scale scores, were compared between patients with and without AIH, and AIH volume was compared between true-positive and false-negative cases using independent t-tests or Mann–Whitney U tests after the Shapiro–Wilk normality test. Diagnostic performance and volumetric agreement metrics among solutions were compared using patient-level paired bootstrap resampling (1,000 iterations). For each bootstrap sample, performance metrics were computed for each algorithm, and pairwise differences were obtained. Two-sided *P* values were calculated from the empirical distribution of bootstrap differences, and 95% confidence intervals (CIs) were derived using the percentile method. This bootstrap-based comparison was applied to both probability-based metrics and binary classification metrics. Stratified assessments by AIH subtype were additionally performed as sub-analyses.

All statistical analyses were conducted using MedCalc (version 23.2.1, MedCalc Software Ltd) and Python (version 3.14, Python Software Foundation) using scikit-learn. Multiple comparisons were controlled using the Benjamini–Hochberg procedure. Statistical significance was set at $P < 0.05$.

Results

Patient characteristics

A total of 436 patients, each with one initial CT scan, were included in the final analysis. A flow chart of patient enrollment is presented in Figure 1. The mean age of the total patient cohort was 62 ± 20 years [standard deviation (SD)], the proportion of male participants was 48%, and the proportion of AIH cases was 12% (n = 52; traumatic, 27; non-traumatic, 25). There were significant differences in median Glasgow Coma Scale scores and modified Rankin Scale scores between the AIH-positive and AIH-negative groups ($P < 0.001$). The mean volume of AIH cases was 33.15 mm^3 ± 48.24 (SD). The results are summarized in Table 1. The detailed numbers of AIH subtypes and volume information are presented in Supplementary Table 1 and Supplementary Figure 1.

Detection performance based on probability score

Solution B showed the highest AUROC [0.99 (95% CI: 0.97, 1.00)]; however, there were no significant differences in pairwise comparisons among solutions A, B, and C ($P > 0.05$). In contrast, the highest AUPRC [0.97 (95% CI: 0.94, 1.00)] and the lowest Brier score [0.02 (95% CI: 0.01, 0.03)] of solution B were significantly different from those of the other solutions ($P < 0.01$).

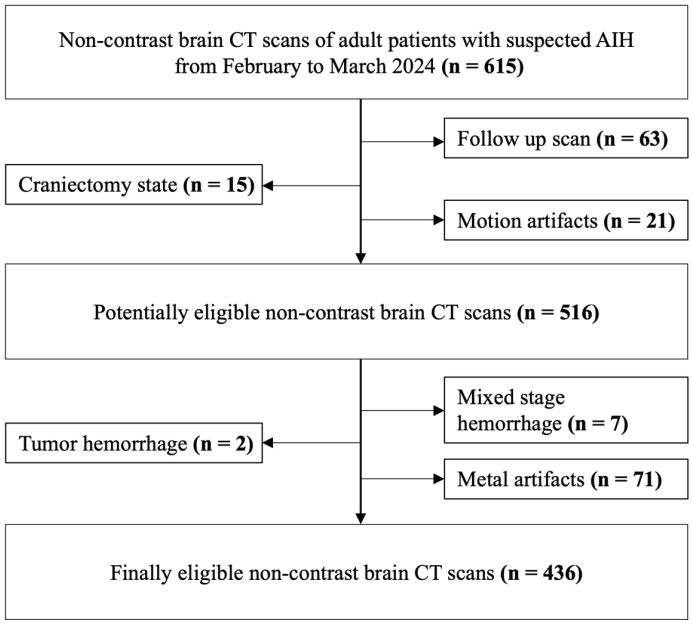

Figure 1. Flow chart of patient enrollment. AIH, acute intracranial hemorrhage; CT, computed tomography.

Table 1. Patients' characteristics

	AIH positive	AIH negative	*P* value
Age (years)†	71 ± 15	61 ± 20	< 0.001*
Number of patients (%)	52 (12)	384 (88)	< 0.001*
Number of male patients (%)	28 (54)	181 (47)	0.576
Glasgow Coma Scale score†	13.42 ± 3.77	14.93 ± 0.87	< 0.001*
Modified Rankin Scale score†	1.92 ± 2.26	0.15 ± 0.67	< 0.001*
Number of traumatic AIH	27		
Number of non-traumatic AIH	25		
Number of AIH subtypes‡			
IPH	30		
SAH	28		
SDH	27		
IVH	17		
EDH	3		
AIH volume (mm^3)	33.15 ± 48.24		

**P* value < 0.05, statistical significance.
†Independent t-test was used, and the values were described as the mean ± SD.
‡Counts are not mutually exclusive because a single patient may present with multiple hemorrhage subtypes (e.g., concurrent IPH and SAH).
SD, standard deviation; AIH, acute intracranial hemorrhage; EDH, extradural hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; SAH, subarachnoid hemorrhage; and SDH, subdural hemorrhage.

In sub-analyses according to AIH type, all solutions achieved AUROCs greater than 0.95 across all AIH subtypes. However, solution B consistently showed the highest AUPRCs (0.95 to 1.00) and the lowest Brier scores (0.01 to 0.02). Solution A showed the lowest AUPRCs (0.56 to 0.63), whereas solution C showed the highest Brier scores (0.15 to 0.16) across AIH types. These results are summarized in Table 2, Figure 2, and Supplementary Figure 2. In sub-analyses by etiology, there were no significant differences between traumatic and non-traumatic AIH across solutions ($P > 0.05$) (Supplementary Table 2).

Detection performance based on binary classification

For sensitivity, solution C showed the highest value [0.92 (95% CI: 0.84, 0.98)]; however, there was no significant difference in pairwise comparisons among the solutions ($P > 0.05$). Solution B showed the highest specificity [1.00 (95% CI: 0.99, 1.00)], precision [0.98 (95% CI: 0.93, 1.00)], and F1 score [0.94 (95% CI: 0.89, 0.98)]; however, these metrics were not significantly different from those of solution D. In sub-analyses according to AIH type, solutions A and B showed the highest sensitivity (0.97) for IPH. In addition, perfect sensitivity (1.00) was achieved by solution C for SAH, by solutions B and C for SDH, and by all four solutions for IVH. In contrast, solutions B and D consistently demonstrated the highest specificity and precision, resulting in the highest F1 scores among the four solutions across AIH types. These results are presented in Table 3, Supplementary Table 3, and Supplementary Figure 3. In sub-analyses by etiology, solutions C and D showed the highest sensitivity in traumatic AIH [0.93 (95% CI: 0.81, 1.00)] and non-traumatic AIH [0.96 (95% CI: 0.87, 1.00)], respectively. For all other metrics, solution B showed the highest specificity (1.00), precision (0.96), and F1 scores (0.92–0.94) in both traumatic and non-traumatic AIH. However, there were no significant differences between traumatic and non-traumatic AIH across all solutions ($P > 0.05$) (Supplementary Table 2).

False-negative analysis according to acute intracranial hemorrhage volume

Across all four solutions, the median AIH volumes of false-negative cases (range, 0.30 to 0.73 mm^3) were significantly smaller than those of true-positive cases (range, 15.50 to 16.74 mm^3) ($P < 0.01$). Volume-stratified analysis demonstrated that the false-negative rate was significantly higher for small AIHs (0.17 to 0.35) than for large AIHs (0.00) across all solutions ($P < 0.01$). These findings are summarized in Supplementary Tables 4 and 5.

Volumetric agreement between each solution and ground truth

The ICC of the measured AIH volume among the three neuroradiologists showed excellent inter-reader agreement [1.00 (95% CI: 0.99, 1.00)]. Among the four solutions, solution D showed the lowest mean difference [−0.87 (95% CI: −1.47, −0.27)] and the narrowest limits of agreement (−13.4 to 11.6) for overall AIH. In sub-analyses, solution D showed the lowest mean differences in IPH [−6.49 (95% CI: −12.89, −0.09)], SAH [−5.74 (95% CI: −8.83, −2.65)], SDH [−8.66 (95% CI: −17.24, −0.07)], and IVH [−6.62 (95% CI: −15.21, 1.96)] across all solutions. These results are presented in Table 4 and Figure 3. Examples of the AIH heatmaps generated by each solution are illustrated in Figure 4.

Table 2. Detection performance of artificial intelligence solutions for acute intracranial hemorrhage based on probability scores

Subtype	Solution	AUROC	AUPRC	Brier score
ICH (n = 52)	A	0.96 (0.92, 0.98)	0.70 (0.56, 0.85)	0.07 (0.05, 0.08)
	B	0.99 (0.97, 1.00)	0.97 (0.94, 1.00)	0.02 (0.01, 0.03)
	C	0.97 (0.93, 0.99)	0.91 (0.84, 0.96)	0.15 (0.15, 0.16)
IPH (n = 30)	A	0.97 (0.95, 0.99)	0.63 (0.48, 0.82)	0.06 (0.05, 0.08)
	B	1.00 (1.00, 1.00)	0.99 (0.97, 1.00)	0.02 (0.01, 0.02)
	C	0.95 (0.88, 0.99)	0.85 (0.73, 0.94)	0.16 (0.15, 0.16)
SAH (n = 28)	A	0.95 (0.90, 0.98)	0.56 (0.41, 0.77)	0.07 (0.05, 0.09)
	B	0.98 (0.95, 1.00)	0.95 (0.86, 1.00)	0.02 (0.01, 0.03)
	C	0.99 (0.98, 1.00)	0.92 (0.84, 0.98)	0.15 (0.15, 0.16)
SDH (n = 27)	A	0.96 (0.90, 0.99)	0.59 (0.41, 0.81)	0.07 (0.05, 0.08)
	B	1.00 (1.00, 1.00)	1.00 (0.98, 1.00)	0.01 (0.01, 0.02)
	C	1.00 (0.99, 1.00)	0.96 (0.91, 1.00)	0.15 (0.15, 0.16)
IVH (n = 17)	A	0.99 (0.97, 1.00)	0.59 (0.41, 0.87)	0.06 (0.05, 0.08)
	B	1.00 (1.00, 1.00)	1.00 (0.98, 1.00)	0.01 (0.01, 0.02)
	C	1.00 (0.99, 1.00)	0.94 (0.84, 0.99)	0.15 (0.15, 0.16)

Values are described with 95% CIs in parentheses. AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; ICH, intracranial hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; SAH, subarachnoid hemorrhage; and SDH, subdural hemorrhage; CI, confidence interval.

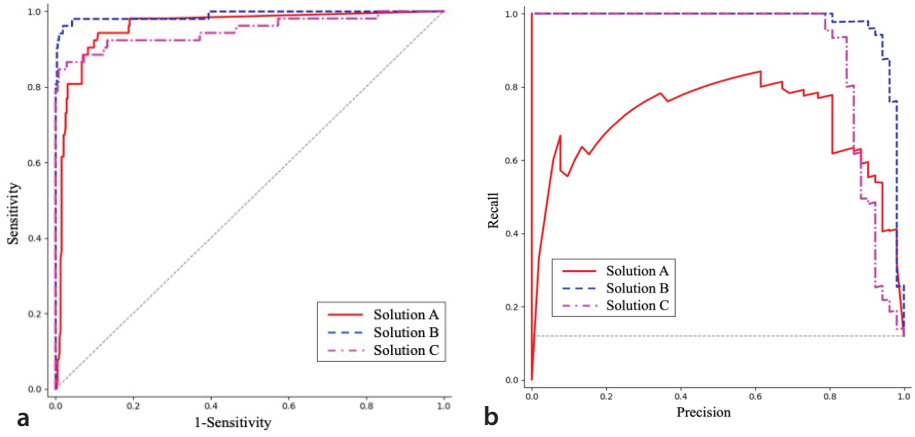

Figure 2. This figure presents the receiver operating characteristic curve **(a)** and precision-recall curve **(b)** for AIH detection. The curves depict solution A as a red solid line, solution B as a blue dashed line, and solution C as a pink dash-dotted line. The black dotted line represents the reference line. Solution B showed the highest AUROC [0.99 (95% CI: 0.97, 1.00)] and AUPRC [0.97 (95% CI: 0.94, 1.00)], followed by solutions A and C. AIH, acute intracranial hemorrhage; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval.

Discussion

In this real-world comparative study of four commercially available CT-based AI solutions for AIH, we observed distinct, task-specific performance profiles across algorithms. All four solutions demonstrated uniformly high discriminative performance; however, solutions B and D exhibited superior precision-oriented performance among all evaluated solutions. This pattern was consistent regardless of AIH subtype, and no significant performance differences were observed based on hemorrhage etiology. Regarding missed detections, false negatives occurred predominantly in small-volume hemorrhages, particularly SAH. In terms of volumetric agreement, solution D demonstrated the most stable and accurate performance across subtypes.

The clinical relevance of between-algorithm differences is well reflected in specific metrics. In this study, all solutions showed high sensitivity and AUROC, metrics emphasizing case discrimination. Among them, solutions B and D demonstrated consistently strong performance in false-positive control, as reflected by precision-oriented metrics. These results suggest consistent performance and a robust risk-scoring framework with a favorable operating point. Therefore, interpretation of the results requires caution because, in class-imbalanced settings such as this real-world cohort, precision-oriented confirmatory metrics may better reflect clinically meaningful differences.[15] In particular, differences in sensitivity or false-negative rate represent variations in the probability of a missed diagnosis at the point of care,[10-12] whereas differences in specificity translate directly into differential false-positive notification burdens in emergency workflows.[5] Therefore, differences in multimetric performance should be considered when evaluating clinical utility, and the optimal algorithm choice may depend on institutional priorities.[15,31] Notably, despite the distinct pathophysiological mechanisms, underlying comorbidities, and prognostic implications associated with traumatic vs. non-traumatic AIH,[32] no significant performance differences were observed between the two etiologic subgroups across all four solutions, suggesting that etiology did not influence algorithm performance in this cohort.

False-negative analysis according to AIH volume revealed that missed cases occurred predominantly in small-volume AIH. For small AIH, which is generally associated with more favorable clinical outcomes,[28-30] the false-negative rate was consistently higher across all solutions than for large AIH, for which perfect sensitivity was observed. In particular, SAH accounted for the majority of missed cases, whereas IVH was reliably detected across all solutions. These results highlight volume and subtype as key determinants of AI failure in AIH detection and suggest that future model development should specifically address and overcome weaknesses such as small-volume SAH.

The volumetric analysis demonstrated distinct performance characteristics across the evaluated solutions. Solutions A and B showed intermediate agreement, indicating reasonable accuracy with some variability. In contrast, solution C exhibited greater variability in volume estimation. Among the solutions, solution D exhibited minimal systematic bias and the most consistent volumetric estimation. Given the excellent inter-reader agreement of the reference standard, measurement error in the ground truth is unlikely to account for the observed

Table 3. Detection performance of artificial intelligence solutions for acute intracranial hemorrhage based on binary classification

Subtype	Solution	Sensitivity	Specificity	False-negative rate	Precision	F1 score
ICH (n = 52)	A	0.90 (0.81, 0.98)	0.91 (0.88, 0.93)	0.10 (0.02, 0.19)	0.57 (0.46, 0.68)	0.70 (0.61, 0.78)
	B	0.90 (0.82, 0.98)	1.00 (0.99, 1.00)	0.10 (0.02, 0.18)	0.98 (0.93, 1.00)	0.94 (0.89, 0.98)
	C	0.92 (0.84, 0.98)	0.93 (0.90, 0.95)	0.08 (0.02, 0.16)	0.64 (0.53, 0.75)	0.76 (0.67, 0.84)
	D	0.85 (0.74, 0.94)	0.99 (0.97, 1.00)	0.15 (0.06, 0.26)	0.90 (0.81, 0.98)	0.87 (0.79, 0.94)
IPH (n = 30)	A	0.97 (0.89, 1.00)	0.91 (0.88, 0.94)	0.03 (0.00, 0.11)	0.45 (0.33, 0.58)	0.62 (0.49, 0.73)
	B	0.97 (0.89, 1.00)	1.00 (0.99, 1.00)	0.03 (0.00, 0.11)	0.97 (0.89, 1.00)	0.97 (0.91, 1.00)
	C	0.87 (0.74, 0.97)	0.93 (0.90, 0.95)	0.13 (0.03, 0.26)	0.49 (0.36, 0.63)	0.63 (0.49, 0.74)
	D	0.93 (0.84, 1.00)	0.99 (0.98, 1.00)	0.07 (0.00, 0.16)	0.85 (0.72, 0.97)	0.89 (0.80, 0.96)
SAH (n = 28)	A	0.89 (0.76, 1.00)	0.91 (0.88, 0.94)	0.11 (0.00, 0.24)	0.42 (0.30, 0.55)	0.57 (0.44, 0.68)
	B	0.86 (0.70, 0.96)	1.00 (0.99, 1.00)	0.14 (0.04, 0.30)	0.96 (0.87, 1.00)	0.91 (0.81, 0.98)
	C	1.00 (1.00, 1.00)	0.93 (0.90, 0.95)	0.00 (0.00, 0.00)	0.51 (0.38, 0.64)	0.67 (0.55, 0.78)
	D	0.86 (0.71, 0.97)	0.99 (0.97, 1.00)	0.14 (0.03, 0.29)	0.83 (0.68, 0.96)	0.84 (0.73, 0.93)
SDH (n = 27)	A	0.93 (0.81, 1.00)	0.91 (0.88, 0.94)	0.07 (0.00, 0.19)	0.42 (0.28, 0.55)	0.57 (0.43, 0.69)
	B	1.00 (1.00, 1.00)	1.00 (0.99, 1.00)	0.00 (0.00, 0.00)	0.96 (0.88, 1.00)	0.98 (0.94, 1.00)
	C	1.00 (1.00, 1.00)	0.93 (0.90, 0.95)	0.00 (0.00, 0.00)	0.50 (0.37, 0.63)	0.67 (0.54, 0.77)
	D	0.89 (0.76, 1.00)	0.99 (0.97, 1.00)	0.11 (0.00, 0.24)	0.83 (0.68, 0.96)	0.86 (0.75, 0.95)
IVH (n = 17)	A	1.00 (1.00, 1.00)	0.91 (0.88, 0.94)	0.00 (0.00, 0.00)	0.33 (0.20, 0.46)	0.49 (0.34, 0.63)
	B	1.00 (1.00, 1.00)	1.00 (0.99, 1.00)	0.00 (0.00, 0.00)	0.94 (0.81, 1.00)	0.97 (0.90, 1.00)
	C	1.00 (1.00, 1.00)	0.93 (0.91, 0.95)	0.00 (0.00, 0.00)	0.39 (0.24, 0.53)	0.56 (0.38, 0.70)
	D	1.00 (1.00, 1.00)	0.99 (0.97, 1.00)	0.00 (0.00, 0.00)	0.77 (0.58, 0.93)	0.87 (0.73, 0.96)

Values are described with 95% CIs in parentheses. Additional metrics, including the false-positive ratio and negative predictive value, are presented in Supplementary Table 3.
ICH, intracranial hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; SAH, subarachnoid hemorrhage; and SDH, subdural hemorrhage.

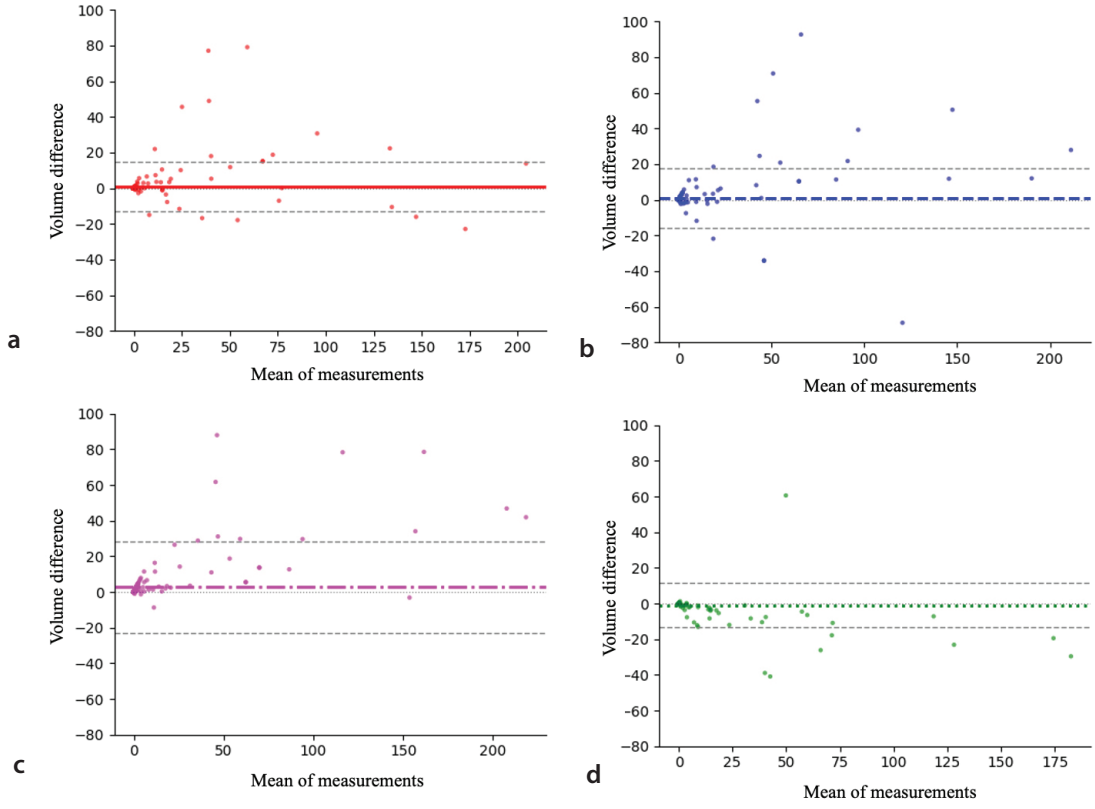

Figure 3. Bland–Altman plots illustrating the agreement between AIH volumes calculated by the solutions and the reference standard established by the neuroradiologists for solutions A **(a)**, B **(b)**, C **(c)**, and D **(d)**. The x-axis represents the mean of the measurements obtained from the AI solution and the reference standard, and the y-axis represents the volume difference (AI solution minus reference standard). The solid horizontal line indicates the mean difference, and the dashed gray lines indicate the 95% LoA. Among the four solutions, solution D showed the lowest mean difference [−0.87 (95% CI: −1.47, −0.27)] and the narrowest LoA (−13.4 to 11.6). AI, artificial intelligence; AIH, acute intracranial hemorrhage; LoA, limit of agreement; CI, confidence interval.

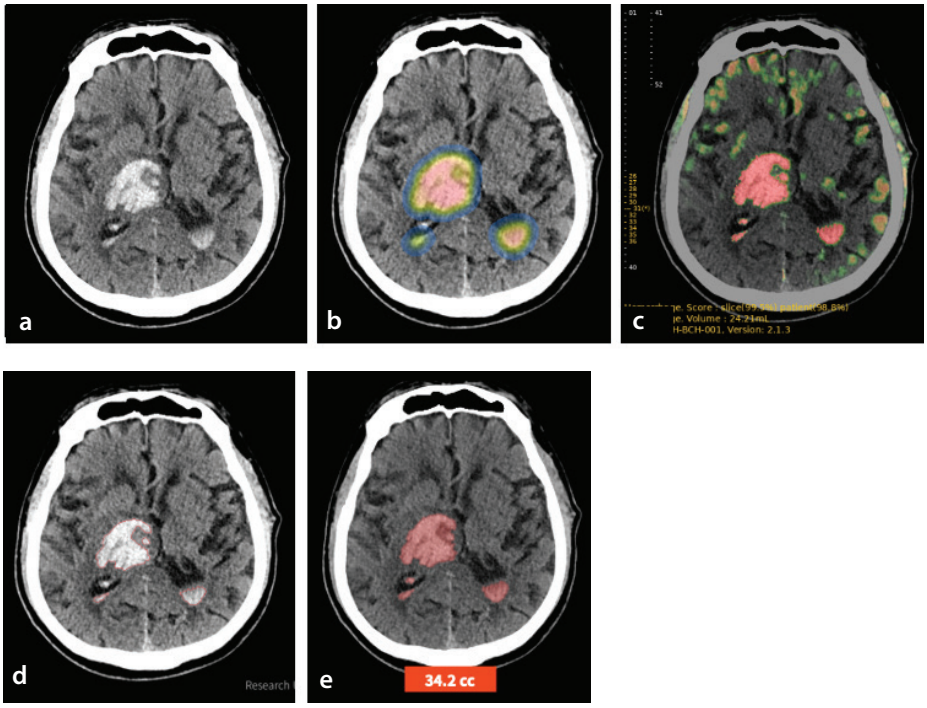

Figure 4. Representative axial non-contrast brain CT images demonstrating AI solution-generated heatmap outputs for AIH detection. The original CT image **(a)** shows IPH in the right thalamus with IVH in the dependent portions of both lateral ventricles. This image is overlaid with probability-based heatmaps from solutions A **(b)**, B **(c)**, C **(d)**, and D **(e)**, indicating areas contributing to hemorrhage detection. AI, artificial intelligence; AIH, acute intracranial hemorrhage; CT, computed tomography; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage.

Table 4. Volumetric agreement between each solution and ground truth according to acute intracranial hemorrhage

Subtype	Solution	Mean difference	Lower LoA	Upper LoA
ICH (n = 52)	A	0.84 (−0.17, 1.52)	−13.2 (−14.32, −12.02)	14.9 (13.71, 16.01)
	B	0.95 (0.14, 1.76)	−15.9 (−17.78, −14.49)	17.8 (16.39, 19.16)
	C	2.76 (1.52, 4.00)	−23.1 (−25.21, −20.97)	28.6 (26.49, 30.73)
	D	−0.87 (−1.47, −0.27)	−13.4 (−14.40, −12.35)	11.6 (10.61, 12.66)
IPH (n = 30)	A	6.29 (−0.28, 12.85)	−29.41 (−52.11, −28.18)	38.17 (16.04, 40.76)
	B	11.19 (1.60, 20.78)	−44.95 (−78.10, −39.15)	51.63 (44.95, 78.10)
	C	23.24 (10.20, 36.27)	−69.12 (−114.18, −45.18)	91.65 (69.12, 114.18)
	D	−6.49 (−12.89, −0.09)	−40.08 (−52.11, −29.41)	27.11 (16.04, 38.17)
SAH (n = 28)	A	7.06 (−0.05, 14.17)	−30.70 (−55.30, −28.88)	43.00 (30.70, 55.30)
	B	9.81 (2.21, 17.40)	−35.06 (−61.34, −28.59)	48.20 (35.06, 61.34)
	C	29.11 (14.81, 43.41)	−76.66 (−43.17, 101.40)	101.40 (76.66, 125.13)
	D	−5.74 (−8.83, −2.65)	−21.36 (−30.70, −15.21)	9.88 (4.54, 15.21)
SDH (n = 27)	A	3.43 (−4.79, 11.64)	−37.28 (−58.35, −29.91)	44.13 (29.91, 58.35)
	B	11.94 (−0.23, 24.10)	−51.17 (−93.28, −48.35)	72.22 (51.17, 93.28)
	C	28.41 (13.81, 43.01)	−75.48 (−126.01, −43.93)	100.75 (75.48, 126.01)
	D	−8.66 (−17.24, −0.07)	−48.73 (−72.01, −51.19)	33.87 (19.02, 48.73)
IVH (n = 17)	A	8.70 (−2.77, 20.18)	−35.05 (−72.47, −32.44)	52.46 (32.44, 72.47)
	B	14.39 (−2.30, 31.08)	−74.13 (−107.13, −49.45)	78.02 (48.91, 107.13)
	C	30.66 (9.65, 51.68)	−49.45 (−147.43, 74.13)	110.78 (74.13, 147.43)
	D	−6.62 (−15.21, 1.96)	−46.86 (−72.47, −32.44)	33.62 (15.21, 52.02)

Values are described with 95% CIs in parentheses. ICH, intracranial hemorrhage; IPH, intraparenchymal hemorrhage; IVH, intraventricular hemorrhage; LoA, limit of agreement; SAH, subarachnoid hemorrhage; and SDH, subdural hemorrhage.

differences in volumetric estimation across solutions, suggesting that these differences reflect algorithmic rather than annotator variability. From a clinical standpoint, volumetric accuracy represents an important surrogate for prognostic assessment and longitudinal monitoring in intracranial hemorrhage.[9,11,28,30,33] Therefore, accurate and reproducible volume estimation is important for evaluating interval changes on follow-up. Although solution D did not demonstrate the highest detection performance, its superior volumetric agreement suggests a distinct strength in quantitative assessment rather than case-level triage.[33-36] The consistent performance of solution D across multiple hemorrhage subtypes further supports its potential utility.

In this study, the differences in performance across solutions are likely attributable to a combination of training dataset characteristics and architectural design choices.[6-8] Although the evaluated solutions obtained regulatory approval based on multi-institutional, multi-racial validation datasets demonstrating robust generalizability, differences in the scale, composition, and case distribution of their respective training data may contribute to the algorithm-specific performance profiles observed in this evaluation.[5,7,8,22-24] Collectively, these findings suggest that commercially available AI tools for AIH are not interchangeable. Rather than being universally superior, the algorithms appear to emphasize different performance dimensions.[10-12,37,38]

Both accurate detection and reliable volume estimation are important; however, their relative priorities can vary across stages of care. In resource-limited settings with reduced staffing or expertise, AI may function as a second-opinion tool to reduce missed diagnoses, emphasizing detection performance. In high-volume emergency settings, solutions with high sensitivity and rapid notification may serve as reliable red-flag systems to prioritize urgent cases. In contrast, in well-resourced tertiary centers, where detection accuracy is less limiting, AI-driven volume quantification may reduce repetitive manual measurement tasks and support longitudinal monitoring.[9-11] This perspective suggests that no single solution is universally optimal and that algorithm selection should be guided by the clinical context and institutional workflow rather than by aggregate performance metrics alone. By characterizing algorithm-specific performance profiles across these dimensions, our multimetric

framework provides actionable guidance for matching AI solutions to the specific demands of each stage of care rather than relying on a single aggregate metric for procurement decisions.

To the best of our knowledge, this is the first external, head-to-head, multimetric comparison of four concurrently regulatory-approved commercial AI solutions under real-world emergency conditions. This study directly addresses the critical gap identified by recent systematic reviews regarding the scarcity of independent head-to-head comparisons in this field.[15,18,39] Unlike vendor-reported benchmarks or single-algorithm validations,[22-24] our comprehensive multimetric framework provides the granularity needed to move beyond the binary question of whether AI can detect hemorrhage toward the more clinically pressing question of which algorithm best serves a given institutional context. Although the observed performance differences likely reflect dataset characteristics such as case mix, subtype distribution, and prevalence rather than intrinsic algorithmic superiority,[22-24,40,41] these differences carry direct practical consequences. Therefore, amid the increasing number of commercially available AI solutions, we propose that AI procurement in acute stroke imaging should be guided not by a single performance metric but by a structured, multimetric evaluation aligned with institutional priorities—a framework this study is designed to support.

Our study had several limitations. First, this was a retrospective, single-center study utilizing CT scanners from a single vendor, which may limit generalizability to other CT vendors. However, considering the consistent cross-vendor performance of the solutions,[5-8,22-24] vendor-specific performance is unlikely to fully explain the algorithm-specific differences observed in this study. In addition, the evaluated AI solutions were limited to those available at Eunpyeong St. Mary's Hospital and may not represent the full spectrum of commercially available algorithms, potentially limiting generalizability. Second, the selection of AI solutions and the case enrollment criteria may limit generalizability. Although geographic-specific bias is substantially mitigated by the characteristics of prior training and validation datasets,[5-8,22-24] these factors remain potentially vulnerable to selection bias. Third, regarding subgroup analyses, individual cases could contain multiple hemorrhage subtypes with relatively small per-subgroup case counts, possibly resulting in non-independent analyses. Thus, even though previous validations have shown consistent results, subtype-level analyses should be interpreted as hypothesis-generating rather than confirmatory. Fourth, each solution's AIH volumetric agreement was evaluated; however, topographic information was not evaluated.[42] Finally, we did not evaluate the clinical impact of the AI solutions, which is the most important consideration when selecting a solution. For example, other important factors (e.g., processing speed, software usability, or workflow compatibility) may affect patient outcomes.[43,44] Therefore, further prospective, multicenter, and comparative clinical evaluations of these AI solutions should be performed in the future.

In conclusion, commercially available CT-based AI tools for AIH demonstrated distinct, task-specific performance profiles in a real-world emergency setting. Although all solutions exhibited excellent discriminative performance, some solutions demonstrated superior confirmatory performance, and certain solutions showed more consistent AIH volumetric agreement with expert measurements. Therefore, these results provide practical guidance for selecting and integrating appropriate AI algorithms to improve AIH diagnosis and management workflows.

Acknowledgement

The corresponding author wishes to acknowledge the financial support of the Catholic Medical Center Research Foundation made in the program year of 2021. The authors appreciate the respective companies for providing temporary demo versions of their software for research purposes. The companies had no role in study design, data collection, analysis, interpretation of the data, manuscript preparation, or the decision to submit the manuscript for publication.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Material: https://d2v96fxpocvxx.cloudfront.net/ecbfe058-16b4-48ee-945f-aeecbad1b86a/content-images/ed674ea0-b24d-497e-801e-8d08508e36d2.pdf

Supplementary Tables: https://d2v96fxpocvxx.cloudfront.net/ecbfe058-16b4-48ee-945f-aeecbad1b86a/content-images/781369fb-80a3-4234-a857-4bf2dd77ac9f.pdf

Supplementary Figures: https://d2v96fxpocvxx.cloudfront.net/580eb5e7-1480-44a6-9404-b8b7446acbcb/content-images/80c1897e-6d32-4c83-8cee-c084dd2f85f0.pdf

References

1. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol.* 2021;31(6):3797-3804. [Crossref]
2. Antonissen N, Tryfonos O, Houben IB, Jacobs C, de Rooij M, van Leeuwen KG. Artificial intelligence in radiology: 173 commercially available products and their scientific evidence. *Eur Radiol*. 2026;36(1):526-536. [Crossref]
3. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022. *Eur Radiol*. 2024;34(1):348-354. [Crossref]
4. Hwang EJ, Park JE, Song KD, et al. 2023 survey on user experience of artificial intelligence software in radiology by the Korean Society of Radiology. *Korean J Radiol*. 2024;25(7):613-622. [Crossref]
5. Kim J, Jang J, Oh SW, et al. Impact of a computed tomography-based artificial intelligence software on radiologists' workflow for detecting acute intracranial hemorrhage. *Diagn Interv Radiol*. 2025;31(5):518-525. [Crossref]
6. Kang DW, Kim M, Park GH, et al. Deep learning-assisted detection of intracranial hemorrhage: validation and impact on reader performance. *Neuroradiology*. 2025;67(6):1511-1519. [Crossref]
7. Yun TJ, Choi JW, Han M, et al. Deep learning based automatic detection algorithm for acute intracranial haemorrhage: a pivotal randomized clinical trial. *NPJ Digit Med.* 2023;6(1):61. [Crossref]
8. Kyung S, Shin K, Jeong H, et al. Improved performance and robustness of multi-task representation learning with consistency loss between pretexts for intracranial hemorrhage identification in head CT. *Med Image Anal*. 2022;81:102489. [Crossref]
9. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. 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(2):167-176. [Crossref]
10. Winder M, Owczarek AJ, Chudek J, Pilch-Kowalczyk J, Baron J. Are we overdoing it? Changes in diagnostic imaging workload during the years 2010-2020 including the impact of the SARS-CoV-2 pandemic. *Healthcare (Basel)*. 2021;9(11):1557. [Crossref]

11. Elliott J, Smith M. The acute management of intracerebral hemorrhage: a clinical review. *Anesth Analg*. 2010;110(5):1419-1427. [Crossref]
12. Kim J, Oh SW, Lee HY, et al. Assessment of deep learning-based triage application for acute ischemic stroke on brain MRI in the ER. *Acad Radiol*. 2024;31(11):4621-4628. [Crossref]
13. Kim J, Oh SW, Lee HY, et al. Efficacy of MRI-based deep learning algorithm for detecting acute ischemic stroke: evaluation among diverse readers. *Eur Radiol*. 2026;36(4):2674-2686. [Crossref]
14. Irgang L, Sestino A, Barth H, Holmén M. Healthcare workers' adoption of and satisfaction with artificial intelligence: the counterintuitive role of paradoxical tensions and paradoxical mindset. *Technol Forecast Soc Change*. 2025;212:123967. [Crossref]
15. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean J Radiol*. 2021;22(3):442-453. [Crossref]
16. Kuziemsky CE, Chrimes D, Minshall S, Mannerow M, Lau F. AI quality standards in health care: rapid umbrella review. *J Med Internet Res*. 2024;26:e54705. [Crossref]
17. Mincu D, Roy S. Developing robust benchmarks for driving forward AI innovation in healthcare. *Nat Mach Intell*. 2022;4:916-921. [Crossref]
18. Weissflog JS, Keller EJ, Neymeyer ML, Morotti A, Dowlatshahi D, Nawabi J. Systematic review of commercial artificial intelligence tools for the detection and volume quantification in intracerebral hemorrhage. *Eur Radiol*. 2026;36(1):367-395. [Crossref]
19. Campolo A. State-of-the-art: the temporal order of benchmarking culture. *Digit Soc*. 2025;4(2):35. [Crossref]
20. Heuron Co Ltd. Heuron ICH: neurovascular AI solution [Internet]. [Crossref]
21. Kang DW, Park GH, Ryu WS, et al. Strengthening deep-learning models for intracranial hemorrhage detection: strongly annotated computed tomography images and model ensembles. *Front Neurol*. 2023;14:1321964. [Crossref]
22. Taggart H. 510(k) premarket notification summary: Hyper Insight - ICH (K240353). Silver Spring (MD): U.S. Food and Drug Administration; 2024. [Crossref]
23. Kim D. 510(k) premarket notification summary: JLK-ICH (K243363). Silver Spring (MD): U.S. Food and Drug Administration; 2025. [Crossref]
24. Smith JJ. 510(k) premarket notification summary: Heuron ICH (K233247). Silver Spring (MD): U.S. Food and Drug Administration; 2024. [Crossref]
25. DenOtter TD, Schubert J. Hounsfield unit. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2019. [Crossref]
26. Ziya A. Determination of bleeding time by hounsfield unit values in computed tomography scans of patients diagnosed with intracranial hemorrhage: evaluation results of computed tomography scans of 666 patients. *Clin Neurol Neurosurg*. 2022;217:107258. [Crossref]
27. Agarwal S, Wood D, Grzeda M, et al. Systematic review of artificial intelligence for abnormality detection in high-volume neuroimaging and subgroup meta-analysis for intracranial hemorrhage detection. *Clin Neuroradiol*. 2023;33(4):943-956. [Crossref]
28. Loan JJ, Gane AB, Middleton L, et al. Association of baseline hematoma and edema volumes with one-year outcome and long-term survival after spontaneous intracerebral hemorrhage: a community-based inception cohort study. *Int J Stroke*. 2021;16(7):828-839. [Crossref]
29. Zhan C, Chen Q, Zhang M, et al. Radiomics for intracerebral hemorrhage: are all small hematomas benign? *Br J Radiol*. 2021;94(1119):20201047. [Crossref]
30. Dowlatshahi D, Smith EE, Flaherty ML, Ali M, Lyden P, Demchuk AM; VISTA Collaborators. Small intracerebral haemorrhages are associated with less haematoma expansion and better outcomes. *Int J Stroke*. 2011;6(3):201-206. [Crossref]
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. [Crossref]
32. Siddique MS, Gregson BA, Fernandes HM, et al. Comparative study of traumatic and spontaneous intracerebral hemorrhage. *J Neurosurg*. 2002;96(1):86-89. [Crossref]
33. Leary OP, Merck LH, Yeatts SD, et al. Computer-assisted measurement of traumatic brain hemorrhage volume is more predictive of functional outcome and mortality than standard ABC/2 method: an analysis of computed tomography imaging data from the progesterone for traumatic brain injury experimental clinical treatment phase-III trial. *J Neurotrauma*. 2021;38(5):604-615. [Crossref]
34. Hemphill JC 3rd, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2015;46(7):2032-2060. [Crossref]
35. Magid-Bernstein J, Girard R, Polster S, et al. Cerebral hemorrhage: pathophysiology, treatment, and future directions. *Circ Res*. 2022;130(8):1204-1229. [Crossref]
36. Chaidee S, Inkeaw P, Makee T, et al. Comparative analysis between different volumetric methods on measuring intracranial hemorrhage incorporating roundness index. *PLoS One*. 2023;18(10):e0292092. [Crossref]
37. Martín Noguerol T, Paulano-Godino F, Martín-Valdivia MT, Menias CO, Luna A. Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. *J Am Coll Radiol*. 2019;16(9 Pt B):1239-1247. [Crossref]
38. Fathi M, Eshraghi R, Behzad S, et al. Potential strength and weakness of artificial intelligence integration in emergency radiology: a review of diagnostic utilizations and applications in patient care optimization. *Emerg Radiol*. 2024;31(6):887-901. [Crossref]
39. Sourlos N, Vliegenthart R, Santinha J, et al. Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. *Insights Imaging*. 2024;15(1):248. [Crossref]
40. Suleman MU, Mursaleen M, Khalil U, et al. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. *Ann Med Surg (Lond)*. 2025;87(12):8803-8811. [Crossref]
41. Musa A, Prasad R, Hernandez M. Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation. *Sci Rep*. 2025;15(1):11383. [Crossref]
42. Delcourt C, Sato S, Zhang S, et al. Intracerebral hemorrhage location and outcome among INTERACT2 participants. *Neurology*. 2017;88(15):1408-1414. [Crossref]
43. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol*. 2021;31(6):3786-3796. [Crossref]
44. Alis D, Tanyel T, Meltem E, et al. Choosing the right artificial intelligence solutions for your radiology department: key factors to consider. *Diagn Interv Radiol*. 2024;30(6):357-365. [Crossref]