

External Validation of BrainScan AI for Detection of Hemorrhagic Stroke on Non-Contrast Head CT

Ivan Sitnikov^{1,2,5*}, Eray Halil^{1,2,5}, Neli Atanasova^{3,5}, Petra Popova^{3,5}, Kostadin Kostadinov⁴, Fares Ezeldin⁶, Edison Judge⁷, Penka Atanassova^{1, 2, 5}

Abstract

Background and Purpose: Rapid and accurate detection of intracranial hemorrhage on non-contrast head CT (NCCT) is critical for acute stroke triage and safe thrombolysis. AI tools may support high-volume emergency workflows, but independent real-world validation remains limited. This study externally validated BrainScan CT (v1.4) for hemorrhage detection, with secondary evaluation of ischemia detection.

Methods: In this retrospective single-centre study at UMHAT “St. George” (Plovdiv, Bulgaria), 1,419 consecutive NCCT scans from 1,260 adults (January–December 2025) were analysed. The reference standard was routine radiology reports. Hemorrhage was present in 120 scans (8.5%) and ischemia in 201 (14.2%). Performance metrics included sensitivity, specificity, predictive values, likelihood ratios, AUC, and McNemar testing. Subgroup analyses examined hemorrhage subtype (ECASS) and anatomical location. Sensitivity analyses assessed threshold variation and repeated scans.

Results: BrainScan CT showed high accuracy for hemorrhage detection: sensitivity 96.7% (95% CI 91.7–99.1), specificity 99.8% (99.4–100.0), PPV 98.3% (94.0–100.0%), NPV 99.7% (99.3–100.0%) and AUC 0.983 (0.967–0.999). Likelihood ratios were 627.9 (positive) and 0.03 (negative), with no systematic disagreement versus the reference standard ($p=0.68$). Sensitivity reached 100% for parenchymal hemorrhage, while false negatives occurred only in hemorrhagic infarction (HI-1 86.7%; HI-2 87.5%). Performance was consistent across regions, lowest for occipital hemorrhage (84.6%). Results were invariant across thresholds and unchanged when restricted to first scans (AUC 1.000). For ischemia, sensitivity was lower (59.2%) despite high AUC (0.930), significantly inferior to hemorrhage detection ($p<0.001$).

Conclusions: BrainScan CT demonstrated excellent real-world performance for hemorrhage detection on NCCT, supporting its role in triage, while expert oversight remains essential, particularly for subtle hemorrhagic infarction.

Keywords: Artificial Intelligence (AI); Computed tomography (CT); Hemorrhagic stroke; Intracranial hemorrhage

Introduction

Intracerebral hemorrhage is a subtype of stroke characterized by bleeding into the brain parenchyma or adjacent intracranial compartments. Although hemorrhagic stroke accounts for approximately one tenth of all stroke cases, it contributes disproportionately to mortality and long-term disability, with in-hospital mortality rates approaching 40–50% [1]. Clinical outcomes are strongly associated with hemorrhage volume, anatomical location, and the timeliness of treatment initiation. Early and accurate diagnosis is therefore critical. Non-contrast head CT (NCCT) remains the imaging modality of choice in the acute evaluation of suspected intracranial hemorrhage, as it is rapid, widely available, and highly sensitive for the detection of acute blood products. As imaging volumes increase, recent deep-learning approaches have sought to support rapid hemorrhage recognition on emergency head CT, with high-sensitivity ensemble methods representing one contemporary example [2]. Once hemorrhage is identified, urgent management measures, such as blood pressure control, anticoagulation reversal, and neurosurgical consultation, can be initiated promptly, potentially improving clinical outcomes.

In routine clinical practice, the detection of intracranial hemorrhage is not always straightforward. Small-volume or subtle hemorrhages—such as punctate intraparenchymal bleeds or thin subdural collections—may be difficult to identify, particularly in early stages or under suboptimal viewing conditions. These findings are especially prone to

¹Department of Neurology, Medical University of Plovdiv, 15A Vasil Aprilov Blvd., 4002 Plovdiv, Bulgaria

²Clinic of Neurology, UMHAT “St George”, 66 Peshtersko Shose Blvd., 4001 Plovdiv, Bulgaria

³Department of Radiology, UMHAT “St George”, 66 Peshtersko Shose Blvd., 4001 Plovdiv, Bulgaria

⁴Department of Social Medicine and Public Health, Medical University of Plovdiv, Plovdiv, Bulgaria

⁵Strategic Research and Innovation Program for the

⁶Development of MU-PLOVDIV–(SRIPD-MUP),

European Union-NextGeneration EU, BG-RRP-2.004-

0007-C01, 4002 Plovdiv, Bulgaria

⁷Department of Forensic Medicine and Deontology,

Medical University of Plovdiv, Bulgaria

⁸EPAM Systems Ltd. 9 Devonshire Square, London EC2M 4YE, United Kingdom

*Corresponding author:

Ivan Sitnikov

E-mail: ir.sitnikov@gmail.com

DOI: 10.2478/ebtj-2026-0019

© 2026 Authors. This work was licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

being overlooked by non-specialist readers or during overnight reporting, when workload and fatigue may further impact diagnostic accuracy. In parallel, increasing imaging volumes in emergency settings introduce an additional challenge related not only to image interpretation but also to the prioritisation of radiology worklists. In this context, delayed recognition of clinically significant hemorrhage may lead to delays in critical interventions. These factors highlight the need for high-sensitivity automated tools that can function as a safety net in acute imaging workflows, complementing radiologist interpretation rather than merely accelerating it.

Despite the central role of CT, radiology departments face increasing imaging volumes and persistent shortages of specialist neuroradiologists. In response, several artificial intelligence (AI) systems have been developed to assist with automated hemorrhage detection and classification. BrainScan CT is one such platform, designed to analyse NCCT examinations and generate automated alerts when hemorrhage is suspected. Although vendor-reported data suggest efficient processing and integration into clinical workflows, independent external validation is essential to determine whether performance generalises beyond the training environment.

Previous studies evaluating commercial AI tools for hemorrhage detection have reported variable results. External validation of the Caire ICH device demonstrated strong performance for intracranial hemorrhage detection [3], whereas other investigations have highlighted site-specific degradation and unresolved failure modes despite promising development metrics [4]. Real-world validation of VeriScout also showed clinically useful performance but persistent false-positive behavior in scans with postoperative change or calcification [5], while a recent post-traumatic evaluation of qER.ai suggested that the software is best interpreted as a supportive rather than stand-alone tool [6]. These findings highlight variability in real-world performance and underscore the need for independent validation across diverse clinical settings. Taken together, these findings suggest that while AI systems for intracranial hemorrhage detection often demonstrate high performance in controlled settings, their real-world effectiveness may be influenced by dataset heterogeneity, imaging variability, and unresolved failure modes.

In the acute evaluation of suspected stroke, the primary role of NCCT extends beyond hemorrhage detection to include the rapid exclusion of bleeding in order to guide treatment decisions. In particular, thrombolytic therapy for acute ischemic stroke can only be initiated once intracranial hemorrhage has been reliably excluded. However, early ischemic changes on NCCT are often subtle or even absent, creating a diagnostic scenario frequently referred to as “negative CT” or “stroke mimic.” This highlights the dual challenge of acute stroke imaging: the need for highly reliable hemorrhage detection alongside the recognition that ischemia detection on NCCT remains inherently limited. Within this context, evaluation of AI performance for ischemia detection, even as a secondary objective, provides additional insight into the broader applicability and limitations of such

tools in real-world stroke workflows.

In this study, we perform an external validation of BrainScan CT using an independent dataset, focusing on intracranial hemorrhage detection on NCCT. We evaluate diagnostic performance using a radiologist consensus reference standard and conduct subgroup analyses by anatomical location and ECASS classification.

We hypothesise that BrainScan CT achieves performance comparable to expert radiologist interpretation for clinically relevant hemorrhage detection and may support triage workflows in acute clinical settings.

The study was designed and reported in alignment with contemporary reporting standards for artificial intelligence in medical imaging, including the TRIPOD+AI framework.

Materials-Methods

Study Design and Setting

This study was conducted as a retrospective, single-centre external validation analysis of an artificial intelligence-based tool (BrainScan CT) for the detection of intracranial hemorrhage on non-contrast head CT (NCCT).

NCCT examinations were acquired using standard institutional head CT protocols on two multidetector CT systems: a 64-slice SOMATOM Definition AS and a 128-slice SOMATOM X.cite (Siemens Healthineers, Erlangen, Germany). Scans were acquired from the skull base to the vertex using scanner-specific acquisition protocols. Acquisition parameters were generally comparable across protocols, with tube voltage ranging from 100 to 120 kVp and automated tube current modulation (CARE Dose4D). Images were reconstructed using standard soft-tissue brain kernels, with a primary slice thickness of 5 mm for routine diagnostic interpretation and additional thin-section reconstructions of 1–2 mm available for multiplanar reconstruction and post-processing. Iterative reconstruction was applied according to the scanner-specific protocol. Imaging protocols remained consistent throughout the study period, with minor scanner- and protocol-specific variations reflecting routine clinical practice.

All NCCT examinations included in the study were obtained consecutively over a 12-month period (January 1 to December 31, 2025) as part of standard emergency and inpatient diagnostic workflows. The dataset reflects real-world clinical practice and encompasses a heterogeneous spectrum of indications, including suspected stroke, head trauma, acute neurological deficit, and altered mental status.

The BrainScan CT software evaluated in this study was developed independently by a third-party vendor. The study investigators had no involvement in the development, training, or optimisation of the algorithm. AI analysis was performed offline on anonymised DICOM datasets to simulate real-world deployment without influencing clinical workflow; the algorithm output did not influence clinical reporting or patient management.

The primary objective of the study was to assess the

diagnostic accuracy of BrainScan CT for the detection of intracranial hemorrhage using an external dataset independent of the algorithm's development environment. Secondary analyses explored performance variability across anatomical regions, hemorrhage subtypes, and laterality.

Reference Standard

The reference standard for the presence or absence of intracranial hemorrhage was established based on radiological diagnosis recorded in the official clinical report, with access to full clinical information including presenting symptoms, neurological examination findings, and follow-up imaging where available. Reports were generated by board-certified radiologists with experience in neuroimaging. Radiologists were blinded to the AI output at the time of reporting.

Intracranial hemorrhage was defined as the presence of hyperattenuating blood products on NCCT, encompassing intraparenchymal, intraventricular, subdural, epidural, and subarachnoid hemorrhage. Classification was binary (hemorrhage present/absent), with positive cases further characterised by hemorrhage type and anatomical location.

The reference standard diagnosis was used directly as the comparator for all diagnostic accuracy analyses. No additional adjudication procedure was applied beyond routine clinical reporting practice. Non-contrast head CT is widely regarded as the first-line imaging modality for the detection of acute intracranial hemorrhage; however, it is not a perfect reference standard, particularly for small-volume hemorrhage or hemorrhagic transformation within ischemic infarction. No systematic confirmation with follow-up imaging or alternative modalities was performed, reflecting real-world clinical practice. This approach may introduce a degree of reference standard misclassification, particularly in subtle cases, and should be considered when interpreting the diagnostic accuracy estimates. The reference standard is accordingly best characterised as a high-quality clinical standard rather than an independent gold standard.

Hemorrhagic cases were further subclassified using an adapted morphological grading based on ECASS criteria: parenchymal hemorrhage categories (PH-1: small parenchymal haematoma occupying less than 30% of the infarcted area with mild mass effect; PH-2: large parenchymal haematoma exceeding 30% of the infarcted area with substantial mass effect) and hemorrhagic infarction categories (HI-1: small petechiae along the margins; HI-2: confluent petechiae within the infarcted area without mass effect). In this dataset, these morphological grades were applied to all hemorrhage-positive cases regardless of confirmed ischaemic aetiology, serving as an institutional severity grading of hemorrhage morphology rather than a strictly etiological ECASS classification. This is acknowledged as a limitation of the subgroup analysis.

Eligibility Criteria

All consecutive NCCT examinations performed during the study period were screened for eligibility. Examinations were

included if they met the following criteria: patient age 18 years or older; availability of a non-contrast head CT acquired at presentation as part of routine clinical care; and sufficient image quality to permit reliable radiological interpretation.

Examinations were excluded if image quality was deemed non-diagnostic due to severe motion artefacts, incomplete intracranial coverage, or major technical acquisition errors. Scans obtained for post-operative evaluation or follow-up of previously known intracranial pathology were also excluded to avoid overrepresentation of chronic or treated conditions that could confound acute hemorrhage detection.

No restrictions were applied based on clinical indication, permitting inclusion of a heterogeneous real-world population encompassing suspected stroke, head trauma, headache, altered consciousness, and other acute neurological presentations. Eligibility assessment was performed prior to AI analysis, and all included examinations were processed irrespective of the presence or absence of pathology, ensuring an unbiased representation of the clinical case mix. No case enrichment, oversampling, or artificial class balancing was performed; the dataset reflects the natural prevalence of intracranial hemorrhage in the studied population.

BrainScan AI Analysis

The BrainScan CT artificial intelligence software (version 1.4, release date 8 July 2025) was used for automated analysis of all NCCT examinations included. All 1,419 examinations were processed offline in a single retrospective batch using version 1.4, irrespective of original acquisition date; 659 scans (46.4%) had been acquired prior to the software release date (January–July 2025) and were analysed post hoc, ensuring uniform application of the same algorithm version across the full dataset. The algorithm processes axial CT images and generates a continuous probability score (range 0–1) reflecting the likelihood of intracranial hemorrhage.

For the primary analysis, examinations were classified as positive or negative using the manufacturer-defined decision threshold of 0.25, representing the default operating point of the algorithm. This threshold was applied uniformly across all cases and was not modified or recalibrated using the study dataset. The use of a fixed, pre-specified threshold without dataset-specific optimisation was intended to reflect real-world deployment conditions and to avoid the optimistic bias associated with post hoc threshold tuning.

AI analysis was performed offline on anonymised DICOM datasets. The software operated fully automatically, without manual interaction or case selection by the investigators. The AI output was not available to the reporting radiologists and did not influence the reference standard assessment. In addition to binary classification, the continuous probability scores generated by the algorithm were retained for receiver operating characteristic (ROC) analysis and threshold-independent evaluation of discriminative performance.

The mean end-to-end study processing time, measured from retrieval of the examination from the picture archiving

and communication system (PACS) to return of the AI-generated result to PACS, was 3 min 35 s. Component-level mean processing times were approximately 1 min 15 s for receiving the study from PACS and processing it on the proxy server (Stage 1), 2 min 16 s for transmitting the study to and processing it through the application programming interface (API) (Stage 2), and 2 s for returning the results to PACS (Stage 3). This short end-to-end processing time supports the technical feasibility of incorporating BrainScan CT into time-critical acute neuroimaging workflows, although it should not be interpreted as equivalent to radiologist report turnaround time or as evidence of improved patient outcomes.

Reporting Standards and Model Transparency

This study was conducted and reported in accordance with the TRIPOD+AI guidelines for reporting of studies evaluating artificial intelligence-based prediction models. The evaluated BrainScan CT algorithm is a proprietary, third-party deep learning system; full technical details regarding model architecture, training data composition, and internal validation procedures were therefore not accessible to the investigators. All aspects of the external validation, including dataset definition, eligibility criteria, reference standard construction, and statistical analysis, were pre-specified and are transparently reported herein.

Input data consisted of standard NCCT DICOM series acquired in routine clinical practice. No additional preprocessing, case selection, or manual data manipulation was performed prior to AI analysis. Missing or non-diagnostic studies were handled at the eligibility stage and excluded prior to AI analysis according to pre-defined criteria; no imputation or data augmentation was applied.

Performance evaluation encompassed discrimination metrics (sensitivity, specificity, AUC) and threshold-dependent classification measures. Calibration and clinical utility analyses, such as calibration plots or decision curve analysis, were not performed, as the study was designed to assess diagnostic accuracy rather than predictive risk estimation; this is acknowledged as a limitation.

Statistical Analysis

Statistical analysis was performed using R (version 4.5.1; R Foundation for Statistical Computing, Vienna, Austria), with the following packages: *epiR* (diagnostic accuracy metrics), *pROC* (ROC analysis), *gtsummary* (descriptive tables), *ggplot2* and *kkstatfun* (figures and analytical functions). All analyses were conducted at the examination level.

Diagnostic performance was evaluated by constructing a 2×2 contingency table comparing the binary AI classification with the reference standard. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall correctly classified proportion were calculated with corresponding 95% confidence intervals using the exact (Clopper–Pearson) method. Positive and negative likelihood ratios (LR+ and LR–), the diagnostic odds ratio (DOR), and

the Youden index were additionally derived; confidence intervals for likelihood ratios were estimated using standard log-transformation. Systematic disagreement between AI classification and the reference standard was assessed using the McNemar test for paired binary outcomes.

ROC analysis was performed using the continuous probability scores generated by the AI algorithm. The area under the ROC curve (AUC) and corresponding 95% confidence intervals were estimated using the DeLong method. The Youden-optimal threshold was identified as an exploratory analysis and did not influence the primary classification results, which were based exclusively on the pre-specified manufacturer-defined threshold of 0.25.

Subgroup analyses assessed sensitivity across anatomical regions, hemispheric laterality, and ECASS hemorrhage subtypes, restricted to reference-standard-positive cases. Sensitivity estimates for subgroups were calculated with Wilson score 95% confidence intervals.

All statistical tests were two-sided, with a significance threshold of $\alpha = 0.05$. No correction for multiple comparisons was applied to subgroup analyses, which are considered exploratory.

Sample Size and Precision Considerations

No prospective sample size calculation was performed, as the study was based on retrospective analysis of consecutively acquired clinical data. Sample size adequacy was assessed post hoc using a precision-based framework. Acceptability thresholds were pre-specified as a 95% confidence interval half-width not exceeding ± 0.05 for sensitivity, specificity, PPV, and NPV, and ± 0.03 for AUC. All primary hemorrhage detection estimates met these thresholds, confirming adequate precision of the reported performance metrics.

Two pre-specified sensitivity analyses were conducted. First, to assess robustness of binary classification performance to threshold choice, diagnostic accuracy was recomputed at six alternative confidence score thresholds ranging from 0.10 to 0.50. Second, to address potential non-independence arising from repeated examinations contributed by the same patient (1,419 scans from 1,260 patients; 159 patients contributing more than one examination), all primary accuracy analyses were repeated after restricting to the earliest scan per patient ($n = 1,260$). Concordance between full-cohort and restricted-sample estimates was assessed by inspection of confidence interval overlap.

No formal adjustment was made for within-patient clustering, and all primary analyses were conducted at the examination level, reflecting the intended per-scan use of the algorithm in clinical practice. The potential for modest underestimation of variance due to repeated measurements is acknowledged as a methodological limitation.

Ethical Considerations

The study was conducted in accordance with the principles of the Declaration of Helsinki and applicable national regulations.

Ethical approval was obtained from the institutional review board of the University Hospital "St. George", Plovdiv, Bulgaria. Given the retrospective nature of the study and the use of fully anonymised imaging and clinical data, the requirement for individual informed consent was waived by the ethics committee. All data were de-identified prior to analysis, and no patient-identifiable information was accessible to the investigators at any stage of the study. The BrainScan CT algorithm was evaluated in a research context only; the AI output was not available to treating clinicians and did not influence clinical decision-making or patient management.

Results

Cohort Characteristics

A total of 1,419 consecutive non-contrast head CT (NCCT) examinations from 1,260 adult patients were included in the analysis. The mean patient age was 66.2 years (SD 15.4), and 703 patients (49.6%) were male. The dataset reflected a heterogeneous real-world clinical population, with examinations performed for a broad range of acute indications including suspected stroke, head trauma, headache, syncope, and altered mental status. A total of 159 patients contributed more than one examination, reflecting repeat imaging in routine clinical care.

Of the 1,419 included examinations, 120 (8.5%) were positive for intracranial hemorrhage according to the reference standard, 201 (14.2%) demonstrated acute ischemia, and 19 (1.3%) showed evidence of both. The remaining 1,117 examinations (78.7%) showed no acute intracranial pathology. The distribution of scan volume by referring department and pathology category is illustrated in Supplement File, Figure S1, and the age distribution across pathology categories is presented in Supplement File, Figure S2.

Among hemorrhage-positive cases, intraparenchymal hemorrhage was the most frequent subtype (45%), followed by subdural hematoma (20%), subarachnoid hemorrhage (15%), intraventricular hemorrhage (10%), and mixed hemorrhage patterns (10%). According to ECASS classification, 49 cases (40.8%) were classified as PH-1, 40 (33.3%) as PH-2, 15 (12.5%) as HI-1, and 16 (13.3%) as HI-2. Anatomically, hemorrhages were most frequently located in the basal ganglia ($n = 53$; 44.2%) and frontal lobe ($n = 35$; 29.2%), with lower frequencies in the parietal lobe ($n = 30$; 25.0%), temporal lobe ($n = 22$; 18.3%), cerebellum ($n = 17$; 14.2%), and occipital lobe ($n = 13$; 10.8%). Regarding laterality, 49 cases were left-sided, 55 were right-sided, and 16 demonstrated bilateral involvement. The reference standard was derived from routine clinical radiological reporting. A formal inter-rater reliability analysis was not conducted as part of this study, which is acknowledged as a limitation discussed in the Limitations section.

Overall Diagnostic Performance

BrainScan CT demonstrated high diagnostic accuracy for the detection of intracranial hemorrhage on NCCT. Of the 1,419

examinations, the algorithm classified 118 as positive and 1,301 as negative. Compared with the reference standard, this yielded 116 true positives, 2 false positives, 4 false negatives, and 1,297 true negatives, as illustrated in Figure 1. A comprehensive summary of all diagnostic performance metrics for both hemorrhage (primary outcome) and ischemia (secondary outcome) is provided in Table 1.

The sensitivity of the algorithm was 96.7% (95% CI: 91.7–99.1%) and specificity was 99.8% (95% CI: 99.4–100.0%). The PPV was 98.3% (95% CI: 94.0–100.0%) and the NPV was 99.7% (95% CI: 99.3–100.0%). The overall correctly classified proportion was 99.6% (95% CI: 99.1–99.8%).

The positive likelihood ratio (LR+) was 627.9 (95% CI: 157.1–2,508.7), indicating a very large increase in post-test probability of hemorrhage following a positive AI result. The negative likelihood ratio (LR–) was 0.03 (95% CI: 0.01–0.09), indicating a marked reduction in post-test probability following a negative result. The diagnostic odds ratio was 18,806.5 and the Youden index was 0.965 (95% CI: 0.911–0.991).

ROC analysis yielded an AUC of 0.983 (95% CI: 0.967–0.999), confirming excellent discriminative performance, as shown in Figure 2. The combined ROC curves for hemorrhage and ischemia detection are presented for comparison in Supplement File, Figure S3. The Youden-optimal threshold was 0.255, essentially identical to the manufacturer-defined primary threshold of 0.25, yielding a sensitivity of 96.7% and specificity of 99.8%, confirming the appropriateness of the pre-specified operating point. McNemar's test showed no statistically significant systematic disagreement between BrainScan CT and the reference standard ($p = 0.68$).

The distribution of AI confidence scores among examinations flagged as positive, stratified by classification outcome, is shown in Figure 3. True-positive cases were associated with consistently high confidence values, while the two false-positive cases received markedly lower scores, indicating that model uncertainty was appropriately reflected in the continuous output.

Performance by Anatomical Location

BrainScan CT demonstrated high sensitivity across anatomical regions, as shown in Figure 4 and summarised in Supplement File, Table S1-2. The highest detection rates were observed for cerebellar hemorrhage (100.0%; 95% CI: 77.1–100.0%; 17/17) and parietal lobe hemorrhage (100.0%; 95% CI: 85.9–100.0%; 30/30), followed by basal ganglia (98.1%; 95% CI: 88.6–99.9%; 52/53) and frontal lobe hemorrhage (97.1%; 95% CI: 83.4–99.9%; 34/35). Temporal lobe hemorrhages were detected with a sensitivity of 95.5% (95% CI: 75.1–99.8%; 21/22). The lowest regional sensitivity was observed for occipital lobe hemorrhages at 84.6% (95% CI: 53.7–97.3%; 11/13), likely reflecting smaller lesion volumes and beam-hardening artefacts characteristic of this anatomical location.

Detection performance was consistently high across hemispheric distributions. Bilateral hemorrhages were identified with 100.0% sensitivity (95% CI: 75.9–100.0%; 16/16).

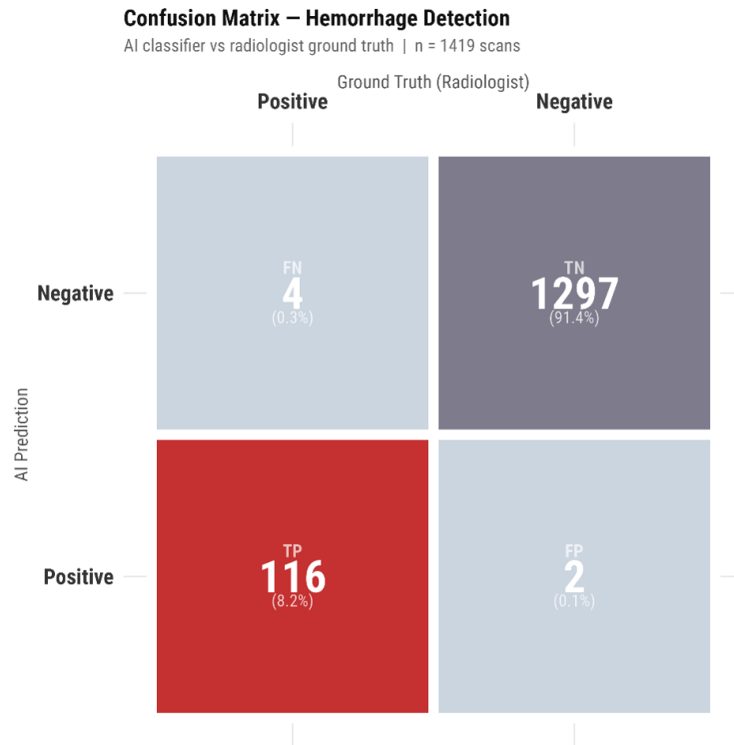


Figure 1. Confusion matrix comparing binary BrainScan CT classifications with the radiologist reference standard for intracranial hemorrhage detection across 1,419 non-contrast head CT examinations. Cell labels indicate absolute counts and corresponding proportions of total examinations. TP = true positive; FP = false positive; FN = false negative; TN = true negative.

Table 1. Diagnostic accuracy of BrainScan CT for intracranial hemorrhage and ischemia detection compared with the radiologist reference standard

Statistic	Hemorrhage	Ischemia
Sensitivity	0.967 (0.917–0.991)	0.592 (0.521–0.661)
Specificity	0.998 (0.994–1.000)	0.998 (0.994–1.000)
Diagnostic odds ratio	18,806.5 (3,408.4–103,769.9)	882.3 (214.3–3,633.2)
Youden index	0.965 (0.911–0.991)	0.590 (0.515–0.660)
Positive likelihood ratio	627.85 (157.13–2,508.74)	360.55 (89.85–1,446.86)
Negative likelihood ratio	0.033 (0.013–0.087)	0.409 (0.346–0.483)

Right-sided hemorrhages were detected in 54 of 55 cases (98.2%; 95% CI: 89.0–99.9%) and left-sided hemorrhages in 46 of 49 cases (93.9%; 95% CI: 82.1–98.4%). All three false-negative cases among lateralised hemorrhages occurred in left-sided lesions.

Performance by ECASS Classification

When stratified by ECASS hemorrhage subtype, BrainScan CT demonstrated complete detection of parenchymal hemorrhage (Figure 5). All PH-1 cases (49/49) and all PH-2 cases (40/40) were correctly identified, corresponding to a sensitivity of 100.0% for both subtypes (Wilson 95% CI: 90.9–100.0% and 89.1–100.0%, respectively).

Sensitivity was lower for hemorrhagic infarction subtypes. For HI-1, 13 of 15 cases were detected (86.7%; 95% CI: 58.4–97.7%), and for HI-2, 14 of 16 cases were detected (87.5%; 95% CI: 60.4–97.8%). All four false-negative cases in the entire cohort occurred exclusively within the hemorrhagic infarction subgroups, with two missed cases in each subtype. This pattern is consistent with the inherently lower imaging conspicuity of petechial hemorrhagic transformation relative to frank parenchymal hematoma on non-contrast CT.

Sensitivity Analyses

To assess robustness of the primary classification threshold, hemorrhage detection performance was recomputed at six alternative confidence score thresholds ranging from 0.10 to 0.50. Sensitivity remained constant at 96.7% (95% CI: 91.7–99.1%) and specificity at 99.8% (95% CI: 99.4–100.0%) across all tested thresholds, with confidence intervals identical at each operating point. This finding indicates that AI confidence scores for hemorrhage are effectively bimodal, with no cases assigned

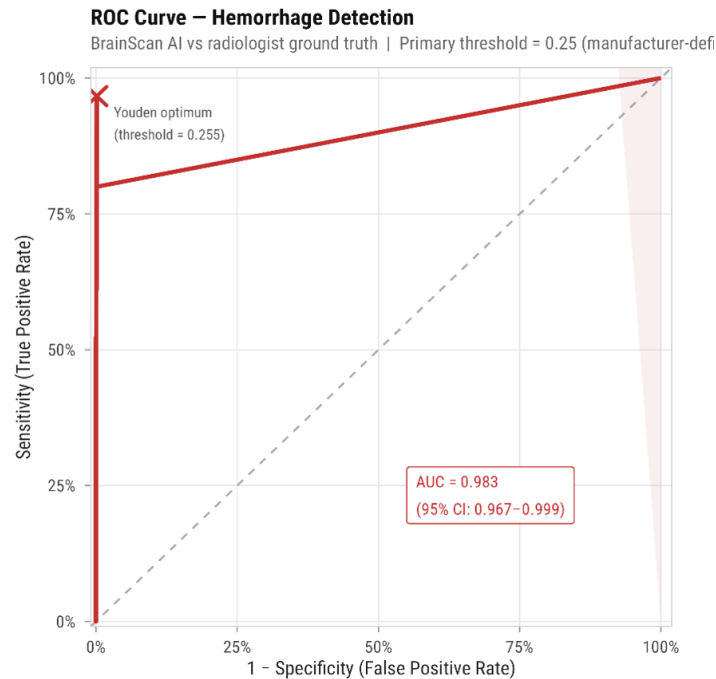


Figure 2. Receiver operating characteristic (ROC) curve for BrainScan CT hemorrhage detection, constructed from continuous AI confidence scores across 1,419 examinations. The AUC was 0.983 (95% CI: 0.967–0.999). The cross symbol denotes the Youden-optimal operating point (threshold = 0.255; sensitivity = 96.7%; specificity = 99.8%), which approximates the manufacturer-defined primary threshold of 0.25.

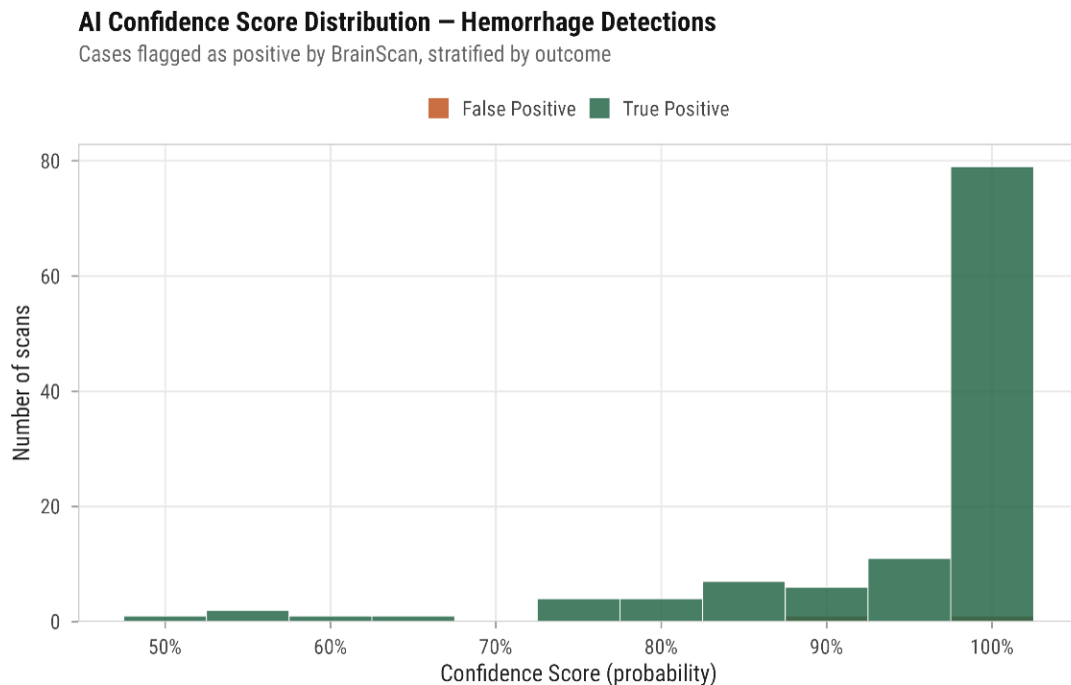


Figure 3. Distribution of AI confidence scores among non-contrast head CT examinations classified as positive by BrainScan CT (n = 118), stratified by classification outcome (true positive, n = 116; false positive, n = 2). Each bar represents the number of scans within a 0.05-wide confidence score interval. True-positive cases are concentrated at high confidence values; false-positive cases received markedly lower scores.

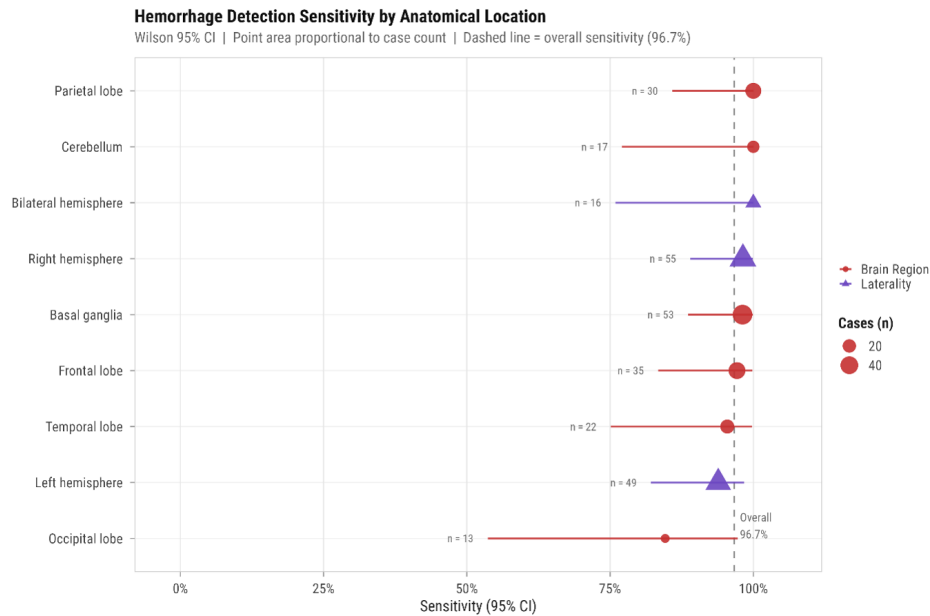


Figure 4. Forest plot of BrainScan CT hemorrhage detection sensitivity by anatomical brain region and hemispheric laterality. Circles represent lobe-level regional estimates; triangles represent laterality estimates. Point area is proportional to the number of reference-standard–positive cases in the subgroup. Error bars represent Wilson 95% confidence intervals. The dashed vertical line indicates the overall cohort sensitivity (96.7%).

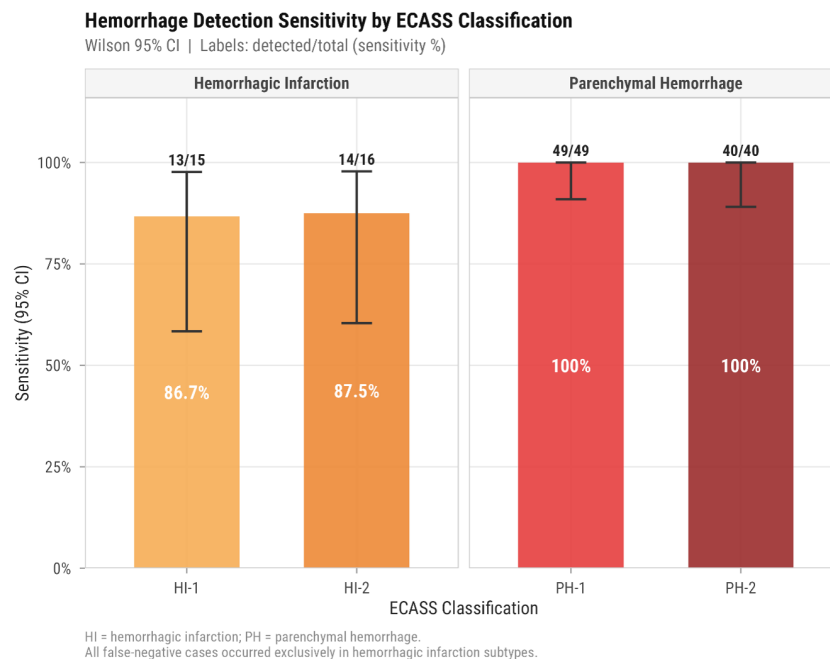


Figure 5. Bar chart of BrainScan CT hemorrhage detection sensitivity stratified by ECASS classification. Bars are grouped by hemorrhage category: hemorrhagic infarction (HI-1, HI-2) and parenchymal hemorrhage (PH-1, PH-2). Sensitivity percentages are displayed within each bar; detected/total case counts are displayed above the error bars. Error bars represent Wilson 95% confidence intervals. HI = hemorrhagic infarction; PH = parenchymal hemorrhage

intermediate probabilities between 0.10 and 0.50, and demonstrates that the primary classification result is entirely invariant to threshold choice within this range.

When the analysis was restricted to the first scan per patient ($n = 1,260$; 79 hemorrhage cases), diagnostic performance was fully concordant with the primary analysis. Sensitivity was 100.0% (95% CI: 95.5–100.0%), specificity 100.0% (99.7–100.0%), PPV 97.5% (91.3–99.7%), NPV 100.0% (99.7–100.0%), and AUC 1.000 (95% CI: 0.999–1.000). All estimates fell within the confidence intervals of the corresponding full-cohort results. Notably, none of the four false-negative cases identified in the full cohort were present in the first-scan-only analysis. Examination of the scan histories of the three affected patients confirmed that all four missed scans were repeat rather than index examinations (scan ranks 2 and 3), with index scans correctly classified as negative. Two patients each contributed one false-negative repeat scan, while one patient contributed two consecutive missed scans on days 2 and 5 after a negative index scan. All four false-negative scans were classified as hemorrhagic infarction (two HI-1, two HI-2). This pattern is most consistent with hemorrhagic transformation that developed or became detectable on follow-up imaging after a genuinely non-hemorrhagic index presentation, rather than reference standard misclassification. The AI was therefore presented with early or subtle hemorrhagic change of limited volume, consistent with the inherent imaging characteristics of HI subtypes, on scans not present in the index dataset. At the time of repeat scanning.

Discussion

Our external validation demonstrates that BrainScan CT achieves high diagnostic accuracy for the detection of intracranial hemorrhage on NCCT. The observed sensitivity (96.7%) and specificity (99.8%) are within the upper range reported for other commercially available AI systems and compare favorably with prior external validation work [3–6].

From a clinical perspective, the combination of very high sensitivity and extremely low negative likelihood ratio supports reliable exclusion of intracranial hemorrhage, while a positive result substantially increases the probability of true pathology. This is particularly relevant in acute stroke triage, where rapid identification of hemorrhage directly influences treatment decisions, including the initiation or withholding of thrombolytic therapy.

Within this context, BrainScan CT demonstrates performance comparable to that reported for currently available systems, with particularly high sensitivity observed for parenchymal hemorrhage subtypes. The AUC of 0.983 further supports its strong discriminative capability.

Beyond binary triage, the NCCT literature has increasingly explored automated hemorrhage delineation and volumetry. Prior studies have demonstrated feasible segmentation of hemorrhage-suspect hyperdensities [10] and deep-learning-based segmentation of intracerebral hemorrhage, intraventricular

hemorrhage, and perihematomal edema [11]. These approaches may contribute to severity assessment and longitudinal monitoring, although volumetric outputs were not evaluated in the present study.

Not all AI applications in stroke imaging demonstrate similarly high performance. AI tools developed for CT angiography (CTA), particularly for the detection of large-vessel occlusion, have shown more modest accuracy [7]. Likewise, external validation of e-ASPECTS has underscored that ischemia-oriented NCCT support does not generalize uniformly across all use contexts and should remain adjunctive to expert interpretation [12]. More recent ischemic-stroke NCCT studies, including multicentre early-ischemia detection models, two-stage deep-learning systems for initially invisible infarction, slice-level lesion classifiers, and infarct-estimation networks, have reported encouraging results, but performance remains heterogeneous and task-dependent [13–16]. Although discrimination for ischemia detection was relatively high (AUC 0.930), this did not translate into clinically sufficient sensitivity, underscoring the known limitations of NCCT for early ischemia detection.

The present study adds to the existing literature in several important ways. First, it provides independent external validation of a commercially available AI system using a consecutive, real-world clinical dataset without case enrichment or artificial class balancing, thereby reflecting routine clinical conditions. Second, the study offers detailed subgroup analyses across anatomical regions and hemorrhage subtypes, demonstrating consistently high performance for clinically relevant parenchymal hemorrhages while identifying specific limitations in detecting hemorrhagic infarction.

Third, by evaluating the model using a fixed, manufacturer-defined threshold without recalibration, the results provide a realistic estimate of performance under conditions that closely resemble actual clinical deployment.

Together, these aspects strengthen the evidence for the applicability of AI-based hemorrhage detection tools in real-world practice and help to clarify both their strengths and limitations. Several aspects of BrainScan CT performance warrant further consideration. The high positive likelihood ratio (>600) and very low negative likelihood ratio indicate that AI classification substantially modifies the post-test probability of intracranial hemorrhage, supporting its potential utility in clinical triage settings.

Performance varied across anatomical locations, with reduced sensitivity observed for occipital hemorrhages. These lesions are often smaller and located adjacent to the calvaria, where beam-hardening artefacts and reduced contrast may limit conspicuity on NCCT.

Hemorrhagic infarctions, particularly HI-1 and HI-2, accounted for the majority of false-negative cases. These lesions typically represent petechial hemorrhagic transformation within ischemic tissue and may be less visually distinct than parenchymal hematomas. Similar limitations have been reported for other AI-based hemorrhage detection systems, underscoring

the importance of interpreting AI outputs within the broader clinical and imaging context.

This observation may also be influenced by limitations of the reference standard, as subtle hemorrhagic transformation can be challenging to distinguish on NCCT and may be subject to inter-observer variability.

This study has several strengths. It includes a relatively large external cohort of 1,419 consecutive real-world examinations without case enrichment or class balancing, ensuring that findings reflect true clinical prevalence. The reference standard was derived from official radiological reports generated blind to the AI output. Diagnostic performance is reported comprehensively with exact 95% confidence intervals across all primary and subgroup metrics, and two pre-specified sensitivity analyses confirm the robustness of the primary findings. The AI analysis was performed offline in a single batch using a fixed manufacturer-defined threshold without post hoc recalibration, reflecting real-world deployment conditions.

Several limitations should also be considered. The dataset was derived from a single centre and included NCCT examinations acquired on a limited range of scanner platforms, which may affect generalisability to other clinical settings. In addition, the prevalence of intracranial hemorrhage in the study cohort (8.5%) may differ from other practice environments; while sensitivity and specificity are prevalence-independent, predictive values may vary accordingly.

BrainScan CT was evaluated as a standalone tool rather than within an integrated radiologist–AI workflow. Prior studies suggest that AI assistance may improve reader sensitivity, agreement, or workflow efficiency in acute stroke imaging, including real-world deployment of comprehensive decision-support platforms such as Brainomix e-Stroke [13,14,17]. However, such benefit is not universal: in a recent prospective real-world evaluation of AI triage for intracranial hemorrhage, radiologist diagnostic performance and report turnaround time were not significantly improved [9]. More broadly, recent research has highlighted the importance of continuous monitoring and confidence-aware workflows in the clinical deployment of AI for intracranial hemorrhage detection. An ensemble monitoring model (EMM) applied to an FDA-cleared algorithm demonstrated complete agreement with the primary model in just over half of cases, particularly when scans were clearly normal or showed obvious hemorrhage [8].

These findings support a confidence-stratified approach to AI-assisted interpretation, in which high-confidence cases may be prioritised differently from lower-confidence examinations requiring closer radiologist review [8]. Such strategies may improve diagnostic performance while balancing false-positive rates and cognitive workload.

A notable finding of the present study is the apparent invariance of classification performance across a range of decision thresholds (0.10–0.50), with sensitivity and specificity remaining unchanged. This reflects a bimodal distribution of confidence scores, with most negative cases assigned near-zero values and positive cases clustered at higher scores.

This suggests that the algorithm functions effectively as a near-binary classifier rather than a fully calibrated probabilistic model. While this enhances robustness of the triage decision, it limits the utility of intermediate confidence scores for graded risk stratification.

This framework is consistent with our observations, particularly regarding reduced sensitivity for subtle hemorrhagic infarctions. Together, these findings support the role of AI as an assistive tool within clinical workflows, complementing rather than replacing expert radiological interpretation.

In addition, the study did not include calibration or clinical utility analyses, such as decision curve analysis, which are recommended components of TRIPOD+AI-compliant evaluation frameworks. While discrimination performance was excellent, further work is needed to assess how model outputs translate into clinical decision-making and patient-level benefit.

An important limitation relates to the reference standard. Although radiologist consensus interpretation of NCCT reflects real-world clinical practice, it does not constitute a perfect gold standard. Small hemorrhages or subtle hemorrhagic transformation may be difficult to detect on NCCT alone, and no systematic confirmation with follow-up imaging or MRI was available. This introduces the possibility of misclassification bias, whereby some apparent false-negative or false-positive AI classifications may reflect uncertainty in the reference standard rather than true model performance.

Another limitation relates to the analysis being performed at the examination level without adjustment for potential within-patient clustering. Although the proportion of repeated scans was relatively small, such correlation may influence variance estimates and confidence intervals. Future studies using patient-level or hierarchical statistical approaches would provide more robust estimates.

The application of ECASS morphological grading to all hemorrhage-positive cases represents a further limitation, as these criteria were originally developed for hemorrhagic transformation in ischemic stroke. This limits direct comparability with published ECASS-based studies. Reduced sensitivity observed in hemorrhagic infarction suggests that performance limitations are concentrated in cases with an underlying ischemic substrate.

Future studies incorporating multimodal imaging or longitudinal clinical validation would provide a more robust reference framework.

Conclusion

BrainScan CT is a highly accurate AI-based tool for the detection of intracranial hemorrhage on non-contrast head CT. In this external validation cohort, it achieved sensitivity approaching 97%, specificity close to 100%, and an AUC of 0.983. Performance was highest for parenchymal hemorrhages and centrally located lesions, with lower sensitivity observed for occipital hemorrhages and hemorrhagic infarctions. Given its high diagnostic performance, BrainScan CT may

serve as an effective triage-support tool in acute clinical settings, particularly in environments with high imaging volumes or limited radiology resources. Its use may support the prioritisation and identification of clinically significant hemorrhage and facilitate timely treatment decisions, while maintaining the need for expert radiologist oversight.

Disclaimer

This research article represents the personal academic work and views of the authors. The findings, conclusions, and opinions expressed in this study are those of the authors and do not represent the views, positions, or endorsements of EPAM Systems, Inc. or any other institutions or organizations that the authors may be associated with in a professional or personal capacity. The evaluation of BrainScan CT software was conducted independently for academic purposes only. EPAM Systems, Inc. does not endorse, recommend, or make any representations regarding the BrainScan CT software or any other commercial products mentioned in this research. All information is provided on an as-is basis for academic and educational purposes.

Ethical Approval

The study was conducted in accordance with the principles of the Declaration of Helsinki and applicable national regulations. Ethical approval was obtained from the institutional review board of the University Hospital "St. George", Plovdiv, Bulgaria. Given the retrospective nature of the study and the use of fully anonymised imaging and clinical data, the requirement for individual informed consent was waived by the ethics committee. All data were de-identified prior to analysis, and no patient-identifiable information was accessible to the investigators at any stage of the study. The BrainScan CT algorithm was evaluated in a research context only; the AI output was not available to treating clinicians and did not influence clinical decision-making or patient management.

Funding

The authors received no financial support for the research.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

None.

References

1. Dagli C, Huang Z, Lin C. Intracerebral hemorrhage hospitalizations and outcomes: comparisons between institutional and national data. *Intern Emerg Med*. 2025;20(5):1455-1462. doi:10.1007/s11739-025-03977-5.
2. Takala J, Peura H, Pirinen R, et al. High sensitivity in spontaneous intracranial hemorrhage detection from emergency head CT scans using ensemble-learning approach. *Sci Rep*. 2025;15(1):29919. doi:10.1038/s41598-025-15835-7.
3. Neves G, Warman PI, Warman A, et al. External Validation of an Artificial Intelligence Device for Intracranial Hemorrhage Detection. *World Neurosurg*. 2023;173:e800-e807. doi:10.1016/j.wneu.2023.03.019.
4. Voter AF, Meram E, Garrett JW, Yu JPY. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage. *J Am Coll Radiol*. 2021;18(8):1143-1152. doi:10.1016/j.jacr.2021.03.005.
5. Wang D, Jin R, Shieh CC, et al. Real world validation of an AI-based CT hemorrhage detection tool. *Front Neurol*. 2023;14:1177723. doi:10.3389/fneur.2023.1177723.
6. Mabit L, Lepoittevin M, Valls M, et al. Real-Life Performance of a Commercially Available AI Tool for Post-Traumatic Intracranial Hemorrhage Detection on CT Scans: A Supportive Tool. *J Clin Med*. 2025;14(13):4403. doi:10.3390/jcm14134403.
7. Mair G, White P, Bath PM, et al. Accuracy of artificial intelligence software for CT angiography in stroke. *Ann Clin Transl Neurol*. 2023. doi:10.1002/acn3.51790.
8. Fang Z, Johnston A, Cheuy LY, et al. Automated real-time assessment of intracranial hemorrhage detection AI using an ensembled monitoring model (EMM). *npj Digit Med*. 2025;8(1):608. doi:10.1038/s41746-025-02007-0.
9. Savage CH, Tanwar M, Elkassem AA, et al. Prospective Evaluation of Artificial Intelligence Triage of Intracranial Hemorrhage on Noncontrast Head CT Examinations. *AJR Am J Roentgenol*. 2024;223(5):e2431639. doi:10.2214/AJR.24.31639.
10. Schmitt N, Mokli Y, Weyland CS, et al. Automated detection and segmentation of intracranial hemorrhage suspect hyperdensities in non-contrast-enhanced CT scans of acute stroke patients. *Eur Radiol*. 2022;32:2246-2254. doi:10.1007/s00330-021-08352-4.
11. Kok YE, Pszczolkowski S, Law ZK, et al. Semantic Segmentation of Spontaneous Intracerebral Hemorrhage, Intraventricular Hemorrhage, and Associated Edema on CT Images Using Deep Learning. *Radiol Artif Intell*. 2022;4(6):e220096. doi:10.1148/ryai.220096.
12. Mair G, White P, Bath PM, et al. External Validation of e-ASPECTS Software for Interpreting Brain CT in Stroke. *Ann Neurol*. 2022;92(6):943-957. doi:10.1002/ana.26495.
13. Chen W, Wu J, Wei R, et al. Improving the diagnosis of acute ischemic stroke on non-contrast CT using deep learning: a multicenter study. *Insights Imaging*. 2022;13(1):184. doi:10.1186/s13244-022-01331-3.
14. Lu J, Zhou Y, Lv W, et al. Identification of early invisible acute ischemic stroke in non-contrast computed tomography

- using two-stage deep-learning model. *Theranostics*. 2022;12(12):5564-5573. doi:10.7150/thno.74125.
15. Sahoo PK, Mohapatra S, Wu CY, et al. Automatic identification of early ischemic lesions on non-contrast CT with deep learning approach. *Sci Rep*. 2022;12(1):18054. doi:10.1038/s41598-022-22939-x.
 16. Gauriau R, Bizzo BC, Comeau DS, et al. Head CT deep learning model is highly accurate for early infarct estimation. *Sci Rep*. 2023;13(1):189. doi:10.1038/s41598-023-27496-5.
 17. Mallon D, Fallon M, Blana E, et al. Real-world evaluation of Brainomix e-Stroke software. *Stroke Vasc Neurol*. 2024;9(5):497-504. doi:10.1136/svn-2023-002859.
 18. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
 19. Moons KGM, Skoetz N, Riley RD, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.