Archives of Academic Emergency Medicine. 2026; 14(1): e31

REVIEW ARTICLE

Explainable Artificial Intelligence in Non-Contrast Brain Computed Tomography Scan for Intracerebral Hemorrhage: A Scoping Review

Hamed Ghorani[1,2], Seyedeh Toktam Masoumian Hosseini[3,4], Matin Noroozi[5], Mohammad Hossein Imani[6], Iman Kiani[7], Nastaran Sadat Mahdavi[8], Soleiman Ahmady[9]*

1. Advanced Diagnostic and Interventional Radiology Research Center (ADIR), Tehran University of Medical Science, Tehran, Iran
2. Department of Radiology, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran
3. Department of Nursing, School of Nursing and Midwifery, Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran
4. Department of Medicine in Canadian Virtual Medical University, Vancouver, Canada
5. School of Medicine, Isfahan University of Medical Science, Isfahan, Iran
6. Student Research Committee, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
7. Students' Scientific Research Center, Tehran University of Medical Sciences, Tehran, Iran
8. Department of Anesthesiology, Mofid Children's Hospital, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
9. Department of Medical Education, Virtual School of Medical Education & Management, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Received: May 2026; Accepted: June 2026; Published online: 26 July 2026

Abstract: **Introduction:** Artificial intelligence (AI) models applied to non-contrast brain computed tomography (CT) scan have demonstrated promising performance in hematoma detection, segmentation, and outcome prediction. Explainable artificial intelligence (XAI) has been proposed as a strategy to improve transparency and facilitate clinical integration. This study aimed to map and summarize XAI methods used in CT scan-based studies of intracranial hemorrhage (ICH), and to assess their validation approaches, transparency, and clinical relevance. **Methods:** The scoping review was conducted in accordance with the PRISMA-ScR guidelines and was registered on the Open Science Framework (osf.io/5kxbt). In this review, original studies of adult and pediatric patients with spontaneous ICH that used AI or deep learning on non-contrast brain CT scan and included an explainability or interpretability method were eligible. Reviews, editorials, conference abstracts without full text, and studies without XAI components were excluded. PubMed, Embase, and Scopus were searched from inception to September 2025. Data were extracted on study design, imaging inputs, AI task, explainability technique, validation strategy, and reported clinical relevance. **Results:** A total of twelve studies met inclusion criteria. Most investigations focused on hematoma detection, segmentation, or prediction of functional outcome or mortality. Gradient-based visualization techniques, particularly Grad-CAM and saliency maps, were the most commonly used explainability approaches. XAI outputs predominantly highlighted hematoma regions and perihematomal tissue; however, reporting and interpretation of explainability varied substantially across studies. External validation was infrequent, and few studies formally assessed alignment between XAI outputs and established radiological or clinical reasoning. The evidence was small, heterogeneous, and largely retrospective, with inconsistent reporting of explainability quality and limited external validation. **Conclusion:** Explainable AI applied to non-contrast brain CT scan in ICH cases, is an evolving field with growing methodological diversity but limited standardization. While XAI techniques offer potential to enhance transparency and clinician confidence, current evidence remains insufficient to support routine clinical implementation. Future studies should emphasize standardized reporting, external validation, and clinically grounded evaluation of explainability outputs.

Keywords: Artificial intelligence; Cerebral hemorrhage; Brain computed tomography; Medical imaging

Cite this article as: Ghorani H, Masoumian Hosseini ST, Noroozi M, et al. Explainable Artificial Intelligence in Non-Contrast Brain Computed Tomography Scan for Intracerebral Hemorrhage: A Scoping Review. Arch Acad Emerg Med. 2026; 14(1): e31. https://doi.org/10.22037/aaem.v14i1.2997.

***Corresponding Author:** Soleiman Ahmady; Department of Medical Education, Virtual School of Medical Education & Management, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: Soleiman.Ahmady@gmail.com, Tel: +98 912 675 21 82, ORCID: 0000-0003-0551-6068

1. Introduction

Intracerebral hemorrhage (ICH) remains one of the most severe cerebrovascular events, representing a substantial portion of acute strokes worldwide and causing high mortality and morbidity (1). Spontaneous ICH often leads to poor functional outcomes and significant early fatality, largely because of both the initial hemorrhage and subsequent risks such as hematoma expansion and mass effect (2). Timely recognition is therefore critical; in clinical practice, non-contrast brain computed tomography (CT) scan serves as the backbone of acute ICH diagnosis due to its rapid availability, sensitivity in detecting hemorrhage, and widespread accessibility even in resource-limited settings (2, 3).

However, interpretation of non-contrast brain CT scan can be challenging, especially under emergency conditions or in centers with limited neuroradiological expertise, which may lead to misdiagnosis or delay in detection of hemorrhage or hematoma expansion (4, 5). In recent years, advances in artificial intelligence (AI), particularly deep learning (DL), have opened new avenues for automated detection, segmentation, volumetric quantification, and prognostication of ICH on CT scan, showing pooled diagnostic performance levels that in many studies approach expert human readers (5). Such AI-based tools promise improved triage, standardized interpretation, and more efficient workflow in emergency neuroimaging settings (5).

Despite recent advances, most deep-learning (DL) models in medical imaging remain 'black boxes', delivering high predictive performance but offering little transparency about how predictions are made, a fundamental challenge for clinical trust, validation, and regulatory acceptance (6-8). To address these concerns, explainable artificial intelligence (XAI) approaches have gained traction, offering methods such as saliency maps, class-activation mapping (CAM), and feature-importance scoring techniques specifically developed to render model decision-making transparent in medical imaging (9-11). In neuroimaging, explainability tools such as saliency-based maps, class–activation mapping, and attribution methods may facilitate error analysis, enhance model reliability, and even uncover imaging biomarkers correlating with pathological processes (10, 12, 13). Nevertheless, the application of XAI in medical-imaging DL studies remains inconsistent; while many report performance metrics, only a minority provide rigorous validation of interpretability outputs (e.g., overlap with ground-truth masks, robustness evaluation, or expert-segmentation comparison) (9, 13, 14).

Moreover, reproducibility in DL-based medical imaging studies is frequently limited due to lack of shared code/data and insufficient methodological transparency (e.g., preprocessing, segmentation pipelines) (15, 16)

Given the global burden of ICH, the growing number of AI-based imaging studies, and the urgent need for trustworthy, clinically applicable tools, a comprehensive synthesis of how explainability has been implemented, evaluated, and reported is timely. Deep learning has shown strong performance in detecting and characterizing intracerebral hemorrhage on non-contrast CT scan, but most studies still emphasize predictive accuracy rather than interpretability. This limits clinical adoption because clinicians need to know whether the model is relying on meaningful radiological features or on spurious patterns. Although explainable AI methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), attention maps, SHapley Additive exPlanations (SHAP), and Integrated Gradients are increasingly used, their application in ICH research has not been systematically reviewed. This scoping review aims to (1) identify and categorize XAI techniques applied to AI models for ICH on non-contrast brain CT scan; (2) appraise their methodological rigor and validation strategies; (3) assess transparency and reproducibility; and (4) discuss clinical implications and provide recommendations for future neuro-AI research.

2. Methods

2.1. Study design and setting

This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR). The protocol for this study was registered at osf.io/5kxbt. A scoping review methodology was selected to comprehensively map the existing literature on the application of XAI techniques in deep-learning models for ICH assessed on non-contrast brain CT scan. This approach was chosen to identify the breadth of available evidence, characterize methodological trends, and highlight gaps in current research rather than to quantitatively synthesize effect sizes.

2.2. Search strategy

A comprehensive literature search was performed across PubMed/MEDLINE, Scopus, and Embase from database inception to September 2025. Search strategies combined controlled vocabulary and free-text terms related to intracerebral hemorrhage, non-contrast CT scan, deep learning, artificial intelligence, and explainable or interpretable models. Reference lists of included studies were manually screened to identify additional relevant publications that may not have been captured through database searches (Supplementary Material). The population of interest comprised adult and pediatric patients with ICH evaluated on non-contrast CT (NCCT). The intervention encompassed deep learning models that explicitly incorporated explainable or interpretable artificial intelligence (XAI) techniques, including but not limited to Grad-CAM and its variants, attention-based mechanisms, SHAP, integrated gradients, and occlusion or perturbation analyses. Since this was a scoping review, no formal comparator was defined. Outcomes of interest included model performance metrics and the nature and quality of explainability outputs across diagnostic, segmentation, volumetric, prognostic, and outcome-prediction tasks.

All identified records were imported into a reference man-

agement system and duplicates were removed. Two reviewers independently screened titles and abstracts for relevance. Full-text articles were subsequently reviewed to determine final eligibility. Any disagreements during the screening or inclusion process were resolved through discussion and, when necessary, consultation with the senior reviewer.

2.3. Study selection

Eligible studies included original research articles that investigated artificial intelligence or deep-learning models applied to non-contrast brain CT scan in patients with ICH and explicitly incorporated explainability or interpretability methods. Both adult and pediatric populations were eligible. Studies were required to report diagnostic, segmentation, volumetric, prognostic, or outcome-prediction tasks. Explainability techniques included, but were not limited to, CAM, Grad-CAM and its variants, attention-based mechanisms, SHAP, Integrated Gradients, and occlusion or perturbation analyses. Review articles, editorials, conference abstracts without full text, non-imaging AI studies, studies not using NCCT, and studies without an explainability component were excluded, as were case reports or studies with n<5. Non-English studies were also excluded.

2.4. Data extraction

Data extraction was performed using a form designed prior to review. Extracted variables included study characteristics such as publication year, country, study design, and sample size; patient population and ICH subtype; imaging acquisition and preprocessing methods; deep-learning architecture and task; explainability technique(s) employed; model performance metrics; validation strategy; and any reported assessment of explainability outputs. Information on code or data availability and other indicators of reproducibility were also recorded when reported. Because of heterogeneity in design and reporting, no formal pooled quality score was applied.

2.5. Statistical analysis

The results were synthesized using descriptive and narrative approaches. Studies were grouped according to their primary task, AI architecture, and explainability method. Given the heterogeneity in study designs, outcomes, and reporting of explainability metrics, no quantitative meta-analysis was performed. Instead, patterns in methodological choices, validation strategies, and explainability practices were summarized.

Finally, transparency and reproducibility were qualitatively assessed by examining the reporting of preprocessing pipelines, model training and validation procedures, availability of code or datasets, and clarity in the interpretation of explainability outputs. These aspects were synthesized to identify common methodological limitations and areas requiring standardization to facilitate clinical translation. Also, quality assessment of the included studies was conducted using the METRICS v1.0 tool.

3. Results

3.1. Study selection and study characteristics

A systematic literature search identified 300 records from electronic databases. After removal of 73 duplicate records, 227 unique records were screened based on titles and abstracts. Of these, 213 records were excluded for not meeting the inclusion criteria. Sixteen full-text articles were assessed for eligibility, 4 of which were excluded because they did not assess a deep learning–based model. Ultimately, 12 studies met the predefined inclusion criteria and were included in the qualitative synthesis and review (Figure 1).

The included studies were retrospective, single center or multicenter cohorts of patients with non traumatic ICH, spanning the period from 2019 to 2025. Populations consisted of adults presenting with acute ICH, including intraparenchymal, intraventricular, subarachnoid, epidural, and subdural hemorrhages as reported in the original studies. The predictor was a deep learning algorithm applied to non-contrast brain CT scan acquired according to standard clinical imaging protocols; the primary outcomes were detection, segmentation, classification of ICH subtypes, hemorrhage stability, hematoma expansion, and prognose (17-28).

This review investigated studies evaluating patients with ICH using deep learning models. The datasets were retrospective and contained non-contrast brain CT images, which were taken according to standard clinical imaging protocols. Sample sizes varied from more than 80 to 1000 patients, with most derived from hospital based imaging archives. Different deep learning models, including 2D and 3D convolutional neural networks, transformer-based classifiers, and ensemble Convolutional Neural Networks - Long Short-Term Memory (CNN-LSTM) architectures were used for different tasks like segmentation, classification, prognose and hematoma expansion. The images were preprocessed with normalization and augmentation techniques (17-28). Some studies also combined images with clinical parameters for enhanced prediction. Overall, the wide international and methodological range of the included studies demonstrates the developing use of deep learning techniques to improve diagnosis and clinical management of ICH.

Across the included studies, several consistent patterns emerged. Gradient-based saliency methods predominated Grad-CAM or class-activation mapping was used in most of the studies, while attention mechanisms, SHAP, Integrated Gradients, and occlusion sensitivity were each applied in only a minority, often in combination with Grad-CAM. Hematoma detection, segmentation, and subtype classification were the most common tasks, followed by prediction of hematoma expansion and functional outcome. Methodologically, the field remained immature. Formal evaluation of explainability outputs against ground-truth segmentations or expert reading was rare. METRICS scores spanned a wide

range (38.5%–87.3%), reflecting substantial heterogeneity in methodological quality. These shared gaps, infrequent external validation, scarce code and data sharing, and the absence of standardized explainability evaluation, rather than the performance of any individual model, define the principal limitations of the current evidence base and represent the clearest priorities for future work.

3.2. Risk of bias assessment

Studies clearly defined eligibility criteria for a representative population and used a high quality reference standard, with several employing multi center imaging data and transparent imaging protocols (17-28). However, reporting of segmentation methodology, feature extraction software, and robustness assessments were inconsistent, with only a few studies formally evaluating fully automated segmentation or end to end pipeline robustness. Data partitioning and performance metrics were generally appropriate, but external validation, handling of confounders, and uncertainty or calibration assessment were limited. Data, code, and model availability were also the same, with only Zhao et al., Yang et al., and Altuve et al. providing some level of information (20, 25, 26). Overall, METRICS scores ranged from 38.5% to 87.3%, indicating that while key methodological elements were often met, there are also needs for improvement in transparency, external validation, and reproducibility in deep learning based ICH studies (20, 25).

3.3. Descriptions of deep learning models

The studies used different kinds of deep learning model structures, evaluating the power of deep learning model in detection, segmentation and classification of ICH. In the case of segmentation and classifying ICH, early studies used basic CNN models, as Lee et al. (2019) developed a modified Visual Geometry Group (VGG)-like 2D CNN with an explainability module, which includes class activation maps. This enables the model to detect and determine different types of hemorrhages with more enhanced visualization (22). Another study by Altuve and Pérez (2022) used a 2D ResNet-18 convolutional neural network. The model was trained on retrospective CT scans for automatic detection of ICH (20). Rangaraj et al. (2022) also used a 3D attention U-Net for segmentation combined with an XGBoost classifier to evaluate hemorrhage stability and the explainability by attention maps and SHAP values (23). Umapathy et al. (2023) also developed a combined CNN-LSTM model for subtype classification, highlighting both temporal and spatial characteristics (19). Another study by Gong et al. (2023) designed a multitask 3D ResNet-34 model capable of predicting hemorrhage volume and patient prognosis, providing practically interpretable Grad-CAM visualizations to support the predictions (18). Recently, Ning et al. (2025) evaluated the performance of 2D and 3D CNNs, and showed the effects of 2D CNNs for volume assessment (24).

In recent years, studies have used more advanced models base on the transfer learning. Chitimireddy et al. (2023) developed a two-stage pipeline consisting of feature extraction by transformer encoder classifiers in order to directly identify ICH in the sinuous space before providing the image again (21). Liang et al. (2025) proposed an attention model that combines lesion location with radiomic features for the classification of etiology and prognosis (17). Yang et al. (2025) used transfer learning-based 3D U-Net models to distinguish cerebral venous sinus thrombosis-related intracerebral hemorrhage from spontaneous hemorrhages, providing accurate assessment (25). Zhao et al. (2025) developed HENet, which combines clinical and imaging data learning to predict hematoma expansion (26). These methods show the importance of development and evaluation of deep learning methods.

3.4. Explainability approach

Durning years of study, DL methods have been used in the case of ICH explainability. Chitimireddy et al. showed that Grad-CAM can be used to generate heatmaps, visualizing significant areas in model predictions (17-21). Also, Liang et al. (2025) used attention mechanisms to assess the importance of lesion location, which improved explainability in multi-modal radiomics (17). Ning et al. (2025) also applied Grad-CAM to detect hemorrhagic foci that indicate the potential for hematoma expansion (24). Zhao et al. (2025) even used Grad-CAM to confirm predictive regions associated with hematoma and intraventricular hemorrhage expansion (26). Lee et al. used class activation maps (CAM) to detect subtype-specific hemorrhagic regions (22). The above-mentioned methodologies have led to improved explainability.

Several methods have been also used combined with Grad-CAM, considering more features of ICH for interpretation. Rangaraj et al. combined attention map visualizations with SHAP assessment to evaluate the level of feature importance, especially recognizing hematoma shape and volume as important predictors (23). Yang et al. also used different attribute methods, such as Integrated Gradients and occlusion sensitivity, to determine the importance of the hemorrhage borders and perihematomal low-density areas (25). These areas align with pervious knowledge about pathological features. These methods localize model focus to important regions on CT slices or sinograms, which improves radiologists' understanding of model decisions and promotes integration into clinical practice.

3.5. Explainability result

Explainability results provided substantial evidence that the deep learning models are capable of showing and finding areas that were both pathologically and clinically important. Grad-CAM and CAM visualizations frequently detected hemorrhagic foci that was also detected by radiologists, as a result it provides visual approval of the model's accuracy. It also enabled hemorrhage border and perihematomal involvement

(17, 19, 25, 26). SHAP and Integrated Gradient analyses also detected important characteristics, including irregularities in hematoma shape, volume, and perihematomal edema, significant prognostic in clinical neuroradiology (23). Rangaraj et al. showed that the attention maps not only showed hemorrhagic regions but also identified geometrical features, which helps evaluating the risk of hemorrhage instability (23). These explainability methods provide better decision-making for clinicians in daily clinical practice.

3.6. Model performance outcomes

The performance of deep learning models in detecting and characterizing ICH was significant with high Area Under the Receiver Operating Characteristic Curve (AUC). Lee et al. achieved a high AUC of 0.99(95% CI: 0.97–1.00) in classifying ICH subtypes, showing a very potent diagnostic ability (22). Altuve and Pérez reported an AUC of 0.987 for binary ICH detection, outperforming several models such as VGG16 and DenseNet (20). Rangaraj et al. achieved an AUC of 0.94 in stability prediction of ICH; also, Gong et al. achieved an AUC of approximately 0.936 (95% CI: 0.914–0.954) for prognostic classification based on volume and clinical outcomes (23). Yang et al.'s model showed robust internal and external validation results with AUCs around 0.94 (95% CI: 0.90–0.97) and 0.85 (95% CI: 0.79–0.90), respectively, in classifying cerebral venous sinus thrombosis–related hemorrhages (25). These results demonstrate the ability of deep learning, surpassing traditional classifiers and even experts' interpretation.

4. Discussion

Explainability approaches used with DL models have been correlated with diagnostic imaging indicators of ICH progression and outcome. In this work, the predominant types of post hoc activation mapping methods were Grad CAM and class activation maps, which could enhance regions on non-contrast CT scans from the interpretation of deep architectures trained for ICH detection and prognosis (22, 26). Attention enhanced U Nets and transformer-based networks incorporate attention weights as a visualization tool in addition to being an intrinsic form of explanation, providing both high segmentation and classification accuracy (21, 23). In addition, some incorporated the use of feature attribution methods such as SHAP, Integrated Gradients, or occlusion sensitivity with Grad CAM or attention maps to apply quantitative evaluations of hemorrhage shape, volume, and surrounding hypo-attenuated tissue (23, 25). Furthermore, other methods of explainable AI in imaging could be used such as LIME, concept activation vectors, prototype and case-based reasoning, and counterfactual explanations. Unlike previous methods, these generally use clinically meaningful structures (13). Incorporating these broader families of methods into future ICH models could enable multi-layered explanations that combine saliency maps, feature attribution, and concept- or prototype-level reasoning, thereby aligning model behavior more closely with radiological reasoning and facilitating clinical adoption (10).

DL models along with explainability approaches have enabled better interpretation of CT scans of ICH patients in a way similar to common signs used by radiologists, as Grad-CAM visualization showed areas of hemorrhage that are connected with clinically defined high-risk characteristics of ICH, such as the swirl sign and black hole sign representing heterogeneous density and consistent hemorrhage (26). Furthermore, maps frequently visualized perihematoma edema, showing the pathologic importance of the volume of edema and the irregularity of hematoma borders, including the island sign, as vital predictors of hematoma ICH progression (27, 28). Comprehensive approaches, such as SHAP, have also confirmed these relationships. This method demonstrated that DL models are weighing shape irregularity and textural heterogeneity to the same extent as expert human evaluations (23). The relation of both approaches confirms that DL models are using biologically significant, robust features rather than merely learning to produce results based on artifacts or noise (26).

DL models first detect and segment in non-contrast CT images to quickly determine the sizes and density of ICH lesions (20, 22, 29). DL models have also been used for classifying subtypes of ICH and identifying the causes, such as cerebral venous sinus thrombosis, which may lead to more optimal, cause-specific treatment plans (19, 25). Recent models are used in order to predict growth of the hematoma, resulting in important decisions for the treatment approach (26, 27). In addition, early detection of the period from the onset of injury to the time, is the ability to evaluate hemorrhage volume and to predict the perihematomal edema progression and clinically long-term function outcomes, thereby evaluating the severity of the patient, rather than simply diagnosing them (18, 30).

We confirmed the findings of earlier systematic reviews about ICH through additional findings; however, our review emphasizes not only the performance of the model, but also the importance of interpretability of ICH models as the primary reason for their usability. A recent systematic review has evaluated a number of different general techniques for explainability including Grad-CAM and SHAP considering different medical imaging techniques, but they have not provided the clinical context for neurologic disorders (31). Similarly, another review that focused on the use of AI in the care of ICH has provided a general overview of the use of AI in the diagnosis and prognosis of ICH (5). A meta-analysis study has also recently demonstrated the high accuracy of segmentation in ICH patients (5). However, in these reviews explainability is generally considered a sub-section, with little or no critical assessment of the fidelity of the explanation of the model used and without assessment of the model power and the way of use in daily clinical practice. Our review investigates the gap between the technical XAI methods and the clinical validation of diagnostic accuracy.

5. Limitations

Several limitations should be considered when interpreting the findings of this scoping review. First, while explanation methods such as Grad-CAM often highlighted hemorrhagic and perihematomal regions that appear clinically relevant, these visual overlaps should not be taken as definitive proof of biological validity or faithful model reasoning. Due to the known limitations of post hoc saliency methods, including instability, sensitivity to input perturbations, and incomplete faithfulness to model internals, the explanatory outputs should be interpreted as supportive evidence rather than confirmation of mechanistic understanding. Second, the majority of included studies relied on retrospective datasets, which introduces a significant risk of bias and limits the generalizability of model performance to real-world clinical settings. The lack of prospective validation undermines confidence in the reproducibility and robustness of reported XAI-enhanced models. Third, there is currently no standardized framework for measuring the quality, fidelity, or clinical utility of explanations generated by AI models in ICH imaging. This absence of standardization creates a tangible risk of eliciting misplaced clinician trust in potentially misleading or oversimplified visualizations, which could adversely affect diagnostic decision-making. Fourth, although this review emphasizes clinical interpretability alongside model performance, the existing literature rarely provides critical assessments of explanation fidelity or rigorous evaluation of how XAI outputs influence clinical practice. Most studies focus on technical development rather than user-centric or outcome-based validation. Finally, the heterogeneity of XAI methods, evaluation metrics, and reporting standards across studies precludes direct comparison and meta-analysis. Future prospective research should prioritize external validation on multi-center datasets, integration of concept-based or counterfactual explainability approaches, and standardized reporting of both model performance and explanation quality to facilitate clinical adoption.

6. Conclusions

Explainable AI applied to non-contrast brain CT scan in ICH is an evolving field with growing methodological diversity but limited standardization. While XAI techniques offer potential to enhance transparency and clinician confidence, current evidence remains insufficient to support routine clinical implementation. Future studies should emphasize standardized reporting, external validation, and clinically grounded evaluation of explainability outputs.

7. Declarations

7.1. Acknowledgments

This article is based on the author's Master's thesis.

7.2. Authors' contributions

H.G.: Conceptualization; Methodology; Software; Formal analysis; Investigation; Data curation; Writing – original draft; Writing – review & editing; Visualization; Project administration.
S.T.M.H: Conceptualization; Methodology; Validation; Formal analysis; Investigation; Data curation; Writing – original draft; Writing – review & editing.
M.N.: Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing – original draft; Visualization.
M.H.I.: Investigation; Data curation; Writing – original draft; Writing – review & editing.
I.K.: Investigation; Resources; Writing – review & editing; Supervision.
S.A.: Conceptualization; Methodology; Writing – review & editing; Supervision; Project administration; Funding acquisition (if any).
All authors read and approved the final version of manuscript.

7.3. Funding/Support

None.

7.4. Consent to publish

Not applicable.

7.5. Availability of data

All data supporting the findings of this review are available within the article and its supplementary materials.

7.6. Conflict of interest

The authors declare no conflicts of interest.

7.7. Using artificial intelligence chatbots

ChatGPT (OpenAI) was used during the preparation of this manuscript solely for language editing and improving the clarity of the text. All scientific content, interpretation, and conclusions were generated and verified by the authors.

7.8. Ethical considerations

This study was carried out with ethical approval granted by the Shahid Beheshti University of Medical Sciences Institutional Review Board (IR.SBMU.SME.REC.1404.024).

References

1. An SJ, Kim TJ, Yoon BW. Epidemiology, Risk Factors, and Clinical Features of Intracerebral Hemorrhage: An Update. J Stroke. 2017;19(1):3-10.
2. Hillal A, Ullberg T, Ramgren B, Wassélius J. Computed tomography in acute intracerebral hemorrhage: neuroimaging predictors of hematoma expansion and outcome. Insights into Imaging. 2022;13(1):180.

3. Sudarshan VK, Raghavendra U, Gudigar A, Ciaccio EJ, Vijayananthan A, Sahathevan R, et al. Assessment of CT for the categorization of hemorrhagic stroke (HS) and cerebral amyloid angiopathy hemorrhage (CAAH): A review. Biocybernetics and Biomedical Engineering. 2022;42(3):888-901.
4. Liu J, Fan W, Yang Y, Peng Q, Ji B, He L, et al. Deep learning-based identification and localization of intracranial hemorrhage in patients using a large annotated head computed tomography dataset: A retrospective multicenter study. Intelligent Medicine. 2025;5(1):14-22.
5. Karamian A, Seifi A. Diagnostic Accuracy of Deep Learning for Intracranial Hemorrhage Detection in Non-Contrast Brain CT Scans: A Systematic Review and Meta-Analysis. Journal of Clinical Medicine [Internet]. 2025; 14(7):[2377 p.].
6. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". Intelligent Medicine. 2024;4(1):52-7.
7. ŞAhiN E, Arslan NN, Özdemir D. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. Neural Computing and Applications. 2025;37(2):859-965.
8. Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. Journal of Medical Ethics. 2022;48(10):764.
9. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches. European Journal of Radiology. 2023;162:110787.
10. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis. 2022;79:102470.
11. Bhati D, Neha F, Amiruzzaman M. A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging. Journal of Imaging [Internet]. 2024; 10(10):[239 p.].
12. Munroe L, da Silva M, Heidari F, Grigorescu I, Dahan S, Robinson EC, et al. Applications of interpretable deep learning in neuroimaging: A comprehensive review. Imaging Neuroscience. 2024;2:imag-2-00214.
13. Muhammad D, Bendechache M. Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. Computational and Structural Biotechnology Journal. 2024;24:542-60.
14. Houssein EH, Gamal AM, Younis EMG, Mohamed E. Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Computing. 2025;28(7):469.
15. Moassefi M, Rouzrokh P, Conte GM, Vahdati S, Fu T, Tahmasebi A, et al. Reproducibility of Deep Learning Algorithms Developed for Medical Imaging Analysis: A Systematic Review. Journal of Digital Imaging. 2023;36(5):2306-12.
16. Kocak B, Yardimci AH, Yuzkan S, Keles A, Altun O, Bulut E, et al. Transparency in Artificial Intelligence Research: a Systematic Review of Availability Items Related to Open Science in Radiology and Nuclear Medicine. Academic Radiology. 2023;30(10):2254-66.
17. Liang J, Tan W, Xie S, Zheng L, Li C, Zhong Y, et al. Prediction of etiology and prognosis based on hematoma location of spontaneous intracerebral hemorrhage: a multicenter diagnostic study. Neuroradiology. 2025;67(7):1761-72.
18. Gong K, Dai Q, Wang J, Zheng Y, Shi T, Yu J, et al. Unified ICH quantification and prognosis prediction in NCCT images using a multi-task interpretable network. Front Neurosci. 2023;17:1118340.
19. Umapathy S, Murugappan M, Bharathi D, Thakur M. Automated Computer-Aided Detection and Classification of Intracranial Hemorrhage Using Ensemble Deep Learning Techniques. Diagnostics (Basel). 2023;13(18).
20. Altuve M, Pérez A. Intracerebral hemorrhage detection on computed tomography images using a residual neural network. Phys Med. 2022;99:113-9.
21. Sindhura C, Al Fahim M, Yalavarthy PK, Gorthi S. Fully automated sinogram-based deep learning model for detection and classification of intracranial hemorrhage. Med Phys. 2024;51(3):1944-56.
22. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng. 2019;3(3):173-82.
23. Rangaraj S, Islam M, Vs V, Wijethilake N, Uppal U, See AAQ, et al. Identifying risk factors of intracerebral hemorrhage stability using explainable attention model. Med Biol Eng Comput. 2022;60(2):337-48.
24. Ning Y, Yu Q, Fan X, Jiang W, Chen X, Jiang H, et al. Non-contrast CT-based deep learning for predicting intracerebral hemorrhage expansion incorporating growth of intraventricular hemorrhage. Sci Rep. 2025;15(1):32021.
25. Yang KC, Xu Y, Lin Q, Tang LL, Zhong JW, An HN, et al. Explainable deep learning algorithm for identifying cerebral venous sinus thrombosis-related hemorrhage (CVST-ICH) from spontaneous intracerebral hemorrhage using computed tomography. EClinicalMedicine. 2025;81:103128.
26. Zhao X, Zhang Z, Shui J, Xu H, Yang Y, Zhu L, et al. Explainable CT-based deep learning model for predicting hematoma expansion including intraventricular hemorrhage growth. iScience. 2025;28(7):112888.
27. Lu M, Wang Y, Tian J, Feng H. Application of deep learning and radiomics in the prediction of hematoma expansion in intracerebral hemorrhage: a fully automated hybrid approach. Diagn Interv Radiol. 2024;30(5):299-312.
28. Huang YW, Huang HL, Li ZP, Yin XS. Research advances in

imaging markers for predicting hematoma expansion in intracerebral hemorrhage: a narrative review. Front Neurol. 2023;14:1176390.

29. Wang X, Shen T, Yang S, Lan J, Xu Y, Wang M, et al. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. NeuroImage: Clinical. 2021;32:102785.
30. Pérez Del Barrio A, Esteve Domínguez AS, Menéndez Fernández-Miranda P, Sanz Bellón P, Rodríguez González D, Lloret Iglesias L, et al. A deep learning model for prognosis prediction after intracranial hemorrhage. J Neuroimaging. 2023;33(2):218-26.
31. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. 2021;32(11):4793-813.

Table 1: Summary of study characteristics

Authors	Data Origin	ICH-subtype	Sample Size	DL Model	ICH Classification Approach	Validation Approach	Interpretability Approach	Explainability (Interpretability) Results	Model Performance
Altuve et al. 2022	Kaggle Database	All types of ICH (not subtype-specific)	200	2D ResNet-18	Binary classification (ICH vs non-ICH)	HOIV	Grad-CAM	Grad-CAM heatmaps correctly focused on hemorrhagic areas matching radiologist findings	Accuracy of 0.959
Chen et al. 2025	USA	All types of ICH (not subtype-specific)	1,028	fICHNet	Binary: prognosis classification	External validation	Saliency map	Saliency mapping reveals that fICHnet bases its post-ICH outcome predictions on both the hemorrhage and surrounding healthy tissue. In the training set, the model prioritizes the hemorrhage itself for infratentorial bleeds, but shifts focus outside the bleed for deep and lobar hemorrhages.	C-index of 0.712 (0.674–0.752) for mRS>2
Gong et al. 2023	China	sICH	258	Multi-task 3D CNN based on 3D ResNet-34 architecture	Binary: prognosis classification; regression for ICH volume estimate	HOIV	Grad-CAM	heatmaps correspond to bleeding areas annotated by clinicians, confirming model's clinical plausibility.	AUC of 0.936
Lee et al. 2019	USA	Intraparenchymal, Intraventricular, Subarachnoid, Epidural, Subdural	904	2D CNN (modified VGG-like) with explainability module	Multi-class classification (5 ICH subtypes + normal)	Cross-validation	Grad-CAM	CAMs localized hemorrhagic regions matching radiologist annotations; with great Cohen's kappa values.	AUC of 0.99
Liang et al. 2025	China	sICH etiologies categorized as hypertension, aneurysm rupture, vascular malformation, and unknown	1162	Deep learning multi-locus attention model algorithm	Multiclassification approach to discriminate among hypertension, aneurysm rupture, and vascular malformation	Fivefold cross-validation on training set Performance tested on independent datasets (2 and 3)	Feature importance analysis	Feature importance plots highlight brain regions critical for classification and prognosis.	Etiology classification: Fusion model (clinical + location) AUC :0.915 (0.909-0.923)
Ning et al. 2025	China	sICH including IVH	775 total (Training: 389, Internal Testing: 167, External Testing: 219)	CNNs: ResNet-101, ResNet-152, DenseNet-121, DenseNet-201 (2D and 3D)	Binary classification: high-risk rHE vs. non-rHE	Internal and external testing sets	Grad-CAM	Highlighted key attention areas in the hematoma and periphery, consistent with NCCT markers of active multifocal bleeding.	AUC: 0.777 (0.716-0.83)
Rangaraj et al. 2022	Singapore	sICH	79	3D Attention U-Net for segmentation + MLP	Binary classification (stable vs unstable) alongside segmentation	Five-fold cross-validation	Integrated attention maps + SHAP analysis	Attention maps localized hemorrhage; SHAP highlighted shape/volume as key predictors.	AUC of 0.94

Table 1: Summary of study characteristics

Authors	Data Origin	ICH-subtype	Sample Size	DL Model	ICH Classification Approach	Validation Approach	Interpretability Approach	Explainability (Interpretability) Results	Model Performance
Sindhura et al. 2023	USA	Intraparenchymal, Intraventricular, Subarachnoid, Subdural, Epidural	8652	Intensity Transformed Sinogram Synthesizer + CNN-RNN	Multi-class (5 ICH subtypes + normal)	Hold-out test set + 5-fold cross-validation (averaged performance)	Grad-CAM visualizations on sinogram domain	Grad-CAM emphasized sinogram angles corresponding to hemorrhage positions.	AUC of 80.7
Umapathy et al. 2023	USA, Brazil	Epidural (EDH), Intraparenchymal (IPH), Intraventricular (IVH), Subarachnoid (SAH), Subdural (SDH)	Training 133,709 slices, testing 14,600 slices; CQ500: 171,390 images from 491 CT scans	3D SE-ResNeXT and LSTM models	Multi-class classification for 5 ICH subtypes using ensemble CNN-LSTM	10-fold cross-validation	Grad-CAM	Grad-CAM accurately localized ICH subtypes in specific lobes; visualized classified regions aiding interpretability.	AUC of 0.99
Yang et al. 2025	China	Cerebral Venous Sinus Thrombosis-related ICH (CVST-ICH) vs sICH	Internal: 408 (102 CVST-ICH + 306 sICH); External: 157 (38 CVST-ICH + 119 sICH)	Transfer learning–based 3D U-Net with segmentation and classification networks	Binary classification (CVST-ICH vs sICH)	Internal testing, external evaluation, and five-fold cross-validation	Grad-CAM, SHAP, Integrated Gradients (IG), and Occlusion	Attention focused on hemorrhage border and perihematomal low-density areas in CVST-ICH; SHAP/IG showed edge and parenchymal region importance.	AUC of 0.94 (0.87-0.98)
Zhang et al. 2024	China	All types of ICH	222	EfficientNetV2	Binary: prognosis classification	External validation	Grad-CAM	Grad-CAM and Guided Grad-CAM visualizations revealed that the deep learning model primarily attended to highly heterogeneous regions within the hematoma ROI.	AUC of 0.85 for radiomics-clinical model (SE: 0.06)
Zhao et al. 2025	China	Conventional hematoma expansion (CHE), Revised hematoma expansion (RHE1, RHE2) including IVH growth	718	HENet (2.5D Deep Convolutional Neural Network using pretrained ResNet101 backbone)	Binary classification for predicting expansion (RHE1, RHE2, CHE)	External validation on two independent datasets (YY, LA); ROC, calibration, and decision curve analyses used	Grad-CAM	Heatmaps showed model attention on hematoma periphery and ventricular regions, supporting biological plausibility via "avalanche model".	AUC: RHE1 (YY=0.882 (0.808-0.956) , LA=0.854 (0.739-0.970))

The abbreviations: AUC: area under the curve; CAM: class activation map; CHE: conventional hematoma expansion; CNN: convolutional neural network; CVST-ICH: cerebral venous sinus thrombosis-related intracerebral hemorrhage; DL: deep learning; EDH: epidural hematoma; Grad-CAM: gradient-weighted class activation mapping; HOIV: hold-out internal validation; ICH: intracerebral hemorrhage; IG: Integrated Gradients; IPH: intraparenchymal hemorrhage; IVH: intraventricular hemorrhage; LIME: local interpretable model-agnostic explanations; LSTM: long short-term memory; MLP: multilayer perceptron; NCCT: non-contrast computed tomography; RHE: revised hematoma expansion; rHE: radiological hematoma expansion; RNN: recurrent neural network; SAH: subarachnoid hemorrhage; SDH: subdural hematoma; SHAP: Shapley additive explanations; sICH: spontaneous intracerebral hemorrhage; XAI: explainable artificial intelligence; mRS: Modified Rankin Scale; VGG: Visual Geometry Group; CT: computed tomography; ROI: region of interest; SE: standard error; ROC: receiver operating characteristic curve.

Table 2: Quality assessment of the included studies

METRICS Items /Article	Liang et al.	Gong et al.	Umapathy et al.	Zhao et al.	Ning et al.	Yang et al.	Rangaraj et al.	Altuve et al.	Lee et al.	Sindhura et al.	Zhang et al.	Chen et al.
Adherence to radiomics / ML-specific checklists or guidelines	No	No	No	No	No	No	No	No	No	No	No	Yes
Eligibility criteria describing a representative study population	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
High-quality reference standard with clear definition	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Multi-center imaging data	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Clinical translatability of imaging data source	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Imaging protocol with acquisition parameters	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No
Interval between imaging and reference standard	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
Transparent description of segmentation methodology	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Formal evaluation of fully automated segmentation	No	No	No	No	No	No	Yes	No	No	No	N/A	Yes
Test-set segmentation masks (single reader/automated tool)	Yes	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No	Yes
Appropriate use of image preprocessing techniques	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Use of standardized feature extraction software	Yes	No	No	No	Yes	No	No	No	No	No	Yes	N/A
Transparent reporting of feature extraction parameters	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes
Removal of non-robust features	Yes	No	No	No	Yes	No	Yes	No	No	No	No	No
Removal of redundant features	Yes	No	No	No	Yes	No	Yes	No	No	No	No	No
Appropriateness of dimensionality vs. data size	Yes	No	No	Yes	Yes	No	Yes	No	No	No	No	Yes
Robustness assessment of end-to-end deep learning pipelines	Yes	No	No	No	No	No	No	No	No	No	No	Yes
Proper data partitioning process	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Handling of confounding factors	No	No	No	No	No	Yes	No	No	No	No	No	No
Use of appropriate performance evaluation metrics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Consideration of uncertainty	No	No	No	Yes	No	Yes	No	No	No	No	Yes	Yes
Calibration assessment	No	No	No	Yes	No	No	No	No	No	No	Yes	Yes
Uni-parametric imaging or proof of inferiority	No	No	No	No	No	No	No	No	No	No	No	No
Comparison with non-radiomic approach / added clinical value	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Comparison with classical statistical models	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	No	Yes
Internal testing	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
External testing	Yes	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Data availability	No	Yes	No	No	No	No	No	Yes	No	No	No	No
Code availability	No	No	No	Yes	No	Yes	No	Yes	No	No	No	Yes
Model availability	No	No	No	Yes	No	Yes	No	Yes	No	No	No	No

ML: machine learning.

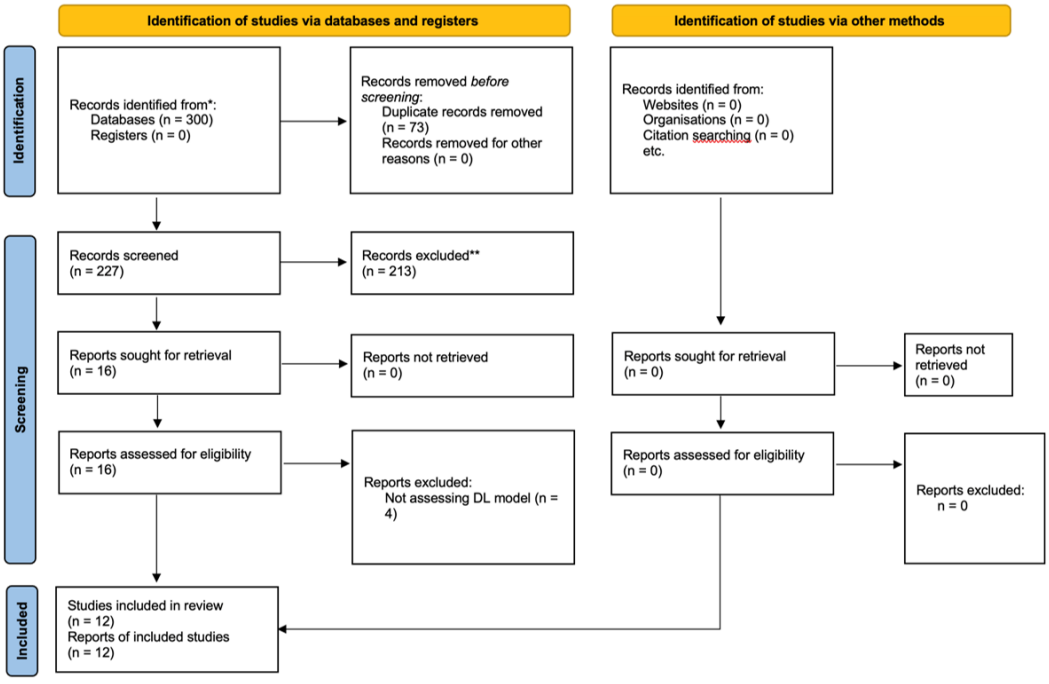

Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart of the included studies. DL: deep learning.

Supplementary table 1: Search strategy in different databases

Database/Search terms
PubMed
("ICH"[tiab] OR "intracerebral haemorrhage"[tiab] OR "intraparenchymal hemorrhage"[tiab] OR "spontaneous intracerebral hemorrhage"[tiab] OR "primary intracerebral hemorrhage"[tiab]) AND ("non-contrast CT"[tiab] OR "noncontrast CT"[tiab] OR "NCCT"[tiab] OR "plain CT"[tiab] OR "unenhanced CT"[tiab] OR "Computed tomography"[tiab] OR "CT scan"[tiab]) AND ("explainable artificial intelligence"[tiab] OR "XAI"[tiab] OR "interpretable artificial intelligence"[tiab] OR "model interpretability"[tiab] OR "explainable AI"[tiab] OR "saliency map"[tiab] OR "saliency maps"[tiab] OR "Grad-CAM"[tiab] OR "gradient-weighted class activation mapping"[tiab] OR "class activation map"[tiab] OR "CAM"[tiab] OR "SHAP"[tiab] OR "Shapley values"[tiab] OR "LIME"[tiab] OR "local interpretable model-agnostic explanations"[tiab] OR "Integrated Gradients"[tiab] OR "occlusion sensitivity"[tiab] OR "attention map"[tiab] OR "attention mechanism"[tiab] OR "feature attribution"[tiab] OR "concept activation vector"[tiab] OR "counterfactual explanation"[tiab] OR "prototype-based explanation"[tiab] OR "case-based reasoning"[tiab] OR "interpretable deep learning"[tiab] OR "explainable deep learning"[tiab] OR "explainability"[tiab] OR "interpretability"[tiab])
Web of Science
TS= (("ICH" OR "intracerebral haemorrhage" OR "intraparenchymal hemorrhage" OR "spontaneous intracerebral hemorrhage" OR "primary intracerebral hemorrhage") AND ("non-contrast CT" OR "noncontrast CT" OR "NCCT" OR "plain CT" OR "unenhanced CT" OR "computed tomography" OR "CT scan") AND ("explainable artificial intelligence" OR "XAI" OR "interpretable artificial intelligence" OR "model interpretability" OR "explainable AI" OR "saliency map" OR "saliency maps" OR "Grad-CAM" OR "gradient-weighted class activation mapping" OR "class activation map" OR "CAM" OR "SHAP" OR "Shapley values" OR "LIME" OR "local interpretable model-agnostic explanations" OR "Integrated Gradients" OR "occlusion sensitivity" OR "attention map" OR "attention mechanism" OR "feature attribution" OR "concept activation vector" OR "counterfactual explanation" OR "prototype-based explanation" OR "case-based reasoning" OR "interpretable deep learning" OR "explainable deep learning" OR "explainability" OR "interpretability"))
Scopus
TITLE-ABS-KEY (("ICH" OR "intracerebral haemorrhage" OR "intraparenchymal hemorrhage" OR "spontaneous intracerebral hemorrhage" OR "primary intracerebral hemorrhage") AND ("non-contrast CT" OR "noncontrast CT" OR "NCCT" OR "plain CT" OR "unenhanced CT" OR "computed tomography" OR "CT scan") AND ("explainable artificial intelligence" OR "XAI" OR "interpretable artificial intelligence" OR "model interpretability" OR "explainable AI" OR "saliency map" OR "saliency maps" OR "Grad-CAM" OR "gradient-weighted class activation mapping" OR "class activation map" OR "CAM" OR "SHAP" OR "Shapley values" OR "LIME" OR "local interpretable model-agnostic explanations" OR "Integrated Gradients" OR "occlusion sensitivity" OR "attention map" OR "attention mechanism" OR "feature attribution" OR "concept activation vector" OR "counterfactual explanation" OR "prototype-based explanation" OR "case-based reasoning" OR "interpretable deep learning" OR "explainable deep learning" OR "explainability" OR "interpretability"))