Gete and Ayele *BMC Neurology* (2025) 25:507
https://doi.org/10.1186/s12883-025-04528-3

BMC Neurology

SYSTEMATIC REVIEW Open Access

Deep-learning–based non-contrast CT for detecting acute ischemic stroke: a systematic review and HSROC meta-analysis of patient-level diagnostic accuracy

Kalab Yigermal Gete[1,2*] and Asnakew Achaw Ayele[3,4]

Abstract

Background Non-contrast CT (NCCT) is first-line imaging for suspected acute ischemic stroke (AIS) but has limited early sensitivity; deep learning (DL) may improve patient-level detection.

Objectives To estimate the diagnostic accuracy of DL applied to NCCT for patient-level AIS detection and to examine prespecified sources of between-study heterogeneity.

Methods We searched MEDLINE, Embase, and Web of Science (January 2010–May 2025). Eligible prospective or retrospective diagnostic studies evaluated DL on NCCT against an appropriate reference standard and reported (or allowed reconstruction of) patient-level 2 × 2 data. Two-gate case–control and lesion-only reports were excluded. Dual reviewers screened/extracted data; risk of bias was assessed with QUADAS-2, and AI-reporting against items adapted from STARD-AI/CLAIM/CONSORT-AI. Bivariate random-effects/HSROC models summarized sensitivity and specificity. Prespecified moderators were posterior-fossa inclusion, reference-standard robustness, and validation type. Sensitivity analyses included external-only cohorts, robust standards, posterior-fossa inclusion, and a "Direct AIS" construct subset.

Results Of 1,899 records, 16 studies met inclusion; 13 contributed patient-level data to meta-analysis. Summary sensitivity was 0.91 (95% CI, 0.81–0.96) and specificity 0.90 (0.85–0.94). Sensitivity was lower for externally validated models than internally validated ones (0.82 [0.67–0.91] vs. 0.95 [0.89–0.98]) with similar specificity (0.88 [0.83–0.92] vs. 0.93 [0.82–0.97]). Findings were directionally robust across sensitivity analyses. QUADAS-2 frequently indicated concerns in patient selection and index-test domains; AI-reporting quality was mostly moderate, and explicit external validation remained uncommon.

Conclusions DL applied to NCCT shows high accuracy for patient-level AIS detection. However, generalizability is the principal gap; broader external validation and guideline-concordant reporting are needed to support safe clinical adoption.

*Correspondence:
Kalab Yigermal Gete
kalabgete@gmail.com

Full list of author information is available at the end of the article

Keywords Deep learning, Non-contrast computed tomography (NCCT), Acute ischemic stroke, Diagnostic test accuracy, Sensitivity and specificity, Hierarchical summary ROC (HSROC), Patient-level meta-analysis, Systematic review

Introduction

Acute ischemic stroke (AIS) is a leading cause of death and disability worldwide, and outcomes depend critically on rapid assessment and time-sensitive reperfusion therapies within organized systems of care [1]. In contemporary emergency pathways, non-contrast CT (NCCT) is the first-line imaging test for suspected stroke because it is fast, broadly available, and reliably excludes intracranial hemorrhage while offering an initial view of early ischemic change (EIC) [1]. However, for *ischemia detection per se*, NCCT is substantially less sensitive than diffusion-weighted MRI (DWI), particularly early after symptom onset; in a landmark prospective comparison, MRI identified acute ischemia far more often than CT across the emergency stroke population [2, 3]. These realities motivate methods that can enhance patient-level AIS detection on NCCT without slowing triage.

The diagnostic challenge is not uniform across AIS presentations. Sensitivity of NCCT is especially limited in the posterior circulation, where beam-hardening from the skull base, small lesion volumes, and slower radiographic evolution obscure EIC—factors that matter at the bedside [4, 5]. Even in the anterior circulation, the visibility of EIC depends on time-to-CT and infarct burden; very early scans (≤ 3–6 h) often show subtle gray–white blurring or loss of the insular ribbon that can be missed under routine windowing [2]. These disease- and physics-specific issues suggest that any technology designed to improve NCCT detection should be evaluated with attention to territory (posterior fossa inclusion), onset-to-scan times, and infarct extent.

It is also essential to distinguish three distinct tasks on NCCT that are frequently conflated in the literature. First, direct AIS detection is a patient-level classification ("AIS present/absent on NCCT") intended to reduce misses and accelerate care. Second, surrogate vascular signs such as the hyperdense artery sign (HAS) indicate large-vessel occlusion (LVO) risk rather than ischemic parenchyma itself; meta-analytic data show HAS is moderately accurate for arterial obstruction, but it is not a parenchymal AIS detector and its prevalence varies with timing and technique [6, 7]. Third, severity/topography scoring with the Alberta Stroke Program Early CT Score (ASPECTS) quantifies EIC in the MCA territory for prognostication and treatment selection; ASPECTS is complementary to, but conceptually distinct from, binary AIS detection [8, 9]. Conflating these constructs—e.g., pooling ASPECTS estimation, HAS detection, slice-level lesion finding, and patient-level AIS diagnosis—muddies clinical interpretability and can inflate performance estimates.

Deep-learning (DL) systems promise to enhance NCCT interpretation by amplifying subtle EIC cues that humans find difficult early after onset. Proof-of-concept studies using MRI-DWI as reference standard demonstrate that DL can localize and delineate early infarction on NCCT, explicitly addressing the modality's well-known insensitivity in the first hours [10]. At the same time, the broader AI-stroke literature remains heterogeneous in tasks (detection vs. ASPECTS vs. surrogates), reference standards (e.g., clinical follow-up vs. DWI), validation (internal vs. external), and operating-point selection—all of which shape measured sensitivity and specificity and hence transportability to real-world triage. Recent narrative and systematic reviews highlight growing capability but also emphasize the need for standardized reporting and external validation before clinical deployment, particularly for ischemic-stroke NCCT tasks where heterogeneity and generalizability remain key challenges [11]. These gaps make a disease-specific, construct-clean synthesis—focused on patient-level AIS detection on NCCT—both timely and clinically meaningful.

Accordingly, this systematic review and patient-level meta-analysis aims to provide robust, clinically interpretable estimates of diagnostic accuracy for DL applied to NCCT in suspected AIS at presentation. We prespecify moderators that reflect AIS pathophysiology and NCCT physics—not generic AI concerns—including time-to-CT, infarct burden (ASPECTS), and posterior fossa inclusion, alongside design features known to affect transportability (reference-standard robustness, validation type, and operating-point choice; DL used alone vs. as an assist). Locking the unit of analysis to the patient improves clinical interpretability and supports translation into emergency triage workflows. Methods and reporting follow PRISMA-DTA [12] with QUADAS-2 [13] risk-of-bias assessment; where applicable, we map AI-specific reporting items to CONSORT-AI [14] and CLAIM [15] to enhance transparency and reproducibility.

Methods

Eligibility criteria

We included prospective or retrospective cross-sectional and cohort studies evaluating the diagnostic accuracy of DL models applied to NCCT for detecting AIS. Eligible studies required both the index test (standalone DL or DL-assist) and a valid reference standard, with sufficient information to construct patient-level 2 × 2 tables (TP,

FP, FN, TN). We excluded two-gate case–control designs with non-representative controls, case reports, reviews, editorials, abstracts without full data, unverifiable accuracy data, and duplicates (retaining the most complete version). Lesion/image-level results were retained for narrative context but excluded from the patient-level meta-analysis.

Participants

Adults (≥ 18 years) undergoing NCCT for suspected AIS in emergency, inpatient, or acute outpatient settings were eligible. We excluded paediatric cohorts, hemorrhage-only cohorts lacking an AIS subgroup, non-acute indications only, and series where baseline scans could not be separated from post-reperfusion imaging.

Index test

The index tests were DL systems applied to NCCT yielding patient-level decisions (AIS present/absent), used as standalone DL or DL-assist. Any DL architecture and typical NCCT pre/post-processing were eligible. We excluded traditional ML/radiomics without DL, non-NCCT pipelines (e.g., CTA/CTP-only), or studies without a clear diagnostic operating point.

Target condition

AIS present at the time of the index NCCT. Subtypes (e.g., anterior/posterior, LVO/non-LVO, infarct size) were eligible if analyzable at the patient level. TIA without objective evidence of acute ischemia (e.g., DWI-positive lesion) was excluded.

Reference standard

Preferred standards were MRI-DWI for confirmation of AIS, with negatives verified by MRI or adequate clinical/imaging follow-up. We also accepted a clearly defined composite reference standard based on blinded expert consensus using all available clinical and imaging information; when MRI/follow-up were unavailable, we accepted expert interpretation independent of the index algorithm (single- or multi-reader, with or without consensus), coded as 'mixed/weaker' in moderator analyses.

Data and language restrictions

Human studies from 2010 onward were eligible. Non-English studies were included only if an accessible English translation was available. Studies had to report—or allow reconstruction of—sensitivity, specificity, or full patient-level 2 × 2 data.

Information sources and search strategy

We searched MEDLINE (Ovid), Embase (Ovid), and Web of Science (January 2010 to May 2025), combining controlled vocabulary and free-text for DL, AIS, NCCT, and diagnostic accuracy. Full strategies are provided in the Supplement (Tables S1–S3).

Study selection

Rayyan was used for deduplication and screening. Two reviewers independently screened titles/abstracts and full texts against prespecified criteria; disagreements were resolved by discussion or third-party adjudication. Reasons for exclusion were recorded.

Data extraction

Two reviewers independently extracted data using a piloted, standardised form. Items included study design/recruitment, participants and setting, index-test details (architecture; standalone vs. DL-assist; operating point; validation type), reference standards, flow/timing, and patient-level performance. One record was captured per unique patient-level arm (DL-alone; DL-assist). Consistency checks verified TP + FN and TN + FP totals, prevalence, and analysed N. Where only sensitivity/specificity with denominators or prevalence were available, we reconstructed 2 × 2 values with documented assumptions; authors were contacted as needed. Lesion-level results were recorded separately and not pooled.

When multiple candidate arms were reported, one primary patient-level estimate per track (DL-alone; DL-assist) was selected using a predefined hierarchy: external over internal test sets; pre-specified/locked over post-hoc thresholds; representative cohorts (ED all-comers; all territories) over enriched subsets; robust reference standards over mixed/weaker; multicentre over single-centre; and vendor-locked over statistically optimised operating points. All choices/ties were logged.

Data definitions

The index test was DL on NCCT (standalone or DL-assist). The target condition was AIS at index presentation. Sensitivity and specificity were TP/(TP + FN) and TN/(TN + FP). Prespecified covariates for subgroup/meta-regression were: time-to-CT (≤ 6 h vs. > 6–24 h/mixed), ASPECTS (≤ 7 vs. > 7), posterior fossa included (yes vs. anterior-only), reference-standard robustness (robust vs. mixed/weaker), validation type (external vs. internal), operating-point selection (pre-specified/locked vs. post-hoc optimised), and mode of use (DL-assist vs. DL-alone). Coding used majority rule (≥ 60%), fixed cut-points, and documented sources.

Risk of bias and reporting quality

Two reviewers assessed risk of bias with QUADAS-2 (Patient Selection, Index Test, Reference Standard, Flow/Timing); disagreements were resolved by consensus or third reviewer. AI-specific reporting was evaluated across seven items adapted from STARD-AI, CLAIM, and

CONSORT-AI—data source, training/test split, external validation, explainability, model availability, workflow integration, and regulatory status—and summarised; studies were classified as Good, Moderate, or Poor using a predefined rubric.

Statistical analysis

Primary outcomes were pooled sensitivity and specificity. We fitted hierarchical summary ROC (HSROC) and bivariate random-effects (Reitsma) models, selecting one operating point per study by hierarchy; independently tested, non-overlapping arms were treated as separate studies. Proportions were analysed on the logit scale with small continuity corrections for zero cells. Outputs included paired forest plots and an HSROC curve with summary point and 95% prediction region; diagnostic odds and likelihood ratios were derived when available. Subgroup analyses and univariable meta-regression were performed for prespecified covariates when both strata contained sufficient studies (≥ 4). Analyses used Stata v17 and R v4.5.1.

Additional Analyses Sensitivity analyses restricted to (i) external validation only, (ii) robust reference standards only (DWI-confirmed positives with adequately verified negatives), (iii) cohorts including the posterior fossa (excluding anterior-only studies), and (iv) a Direct AIS construct subset (patient-level AIS detection; excludes CTA-defined LVO/HAS and ASPECTS-only tools).

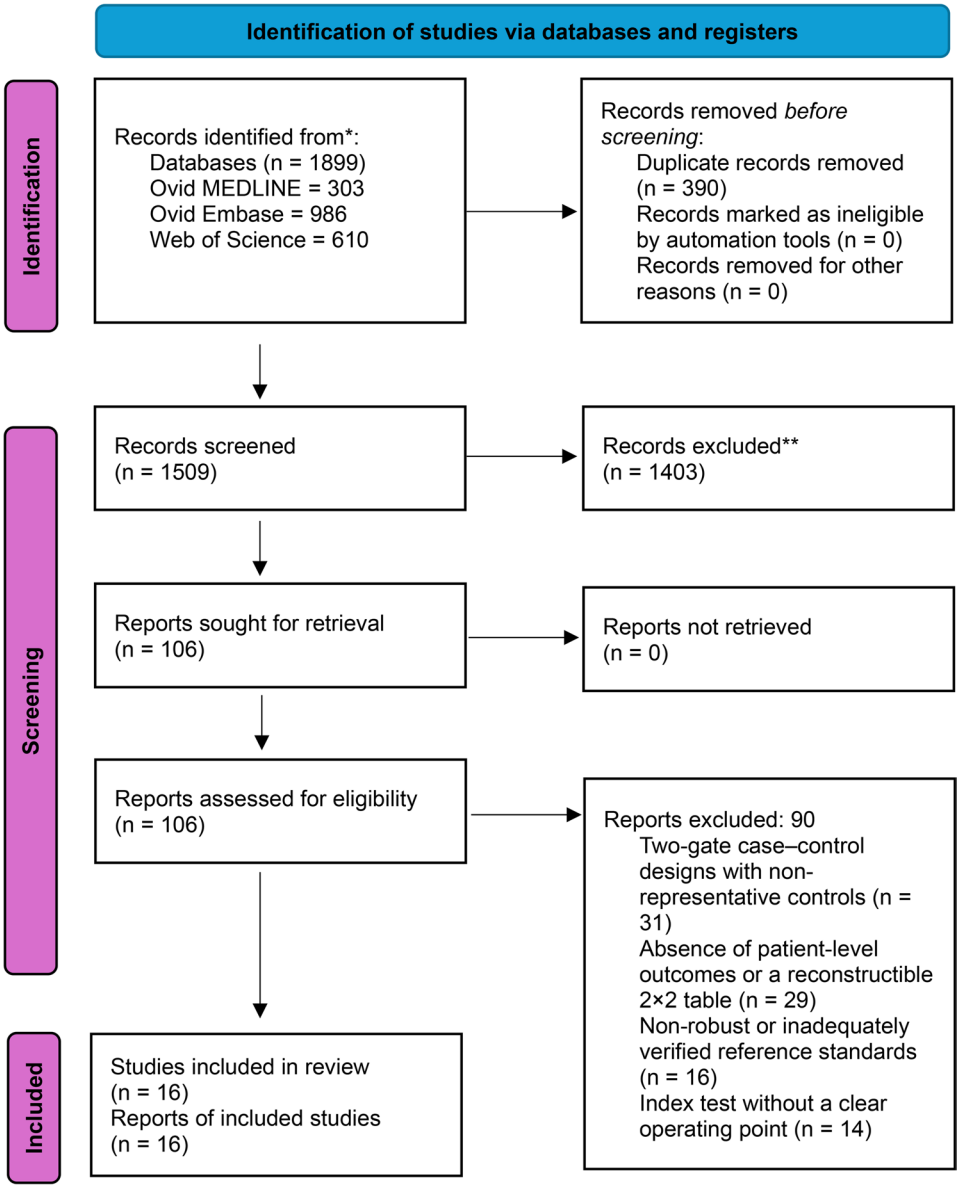

Fig. 1 PRISMA 2020 flow diagram

Table 1 Summary of study characteristics

Study ID	Design; Centers; Country	Inclusion criteria; Exclusion criteria	Unit of analysis; Sample size (*n*)	Index test name	Target condition	Reference standard
Babutain 2021 [16]	Retrospective; Single; Saudi Arabia	Adults with suspected acute stroke undergoing index NCCT; Poor-quality/artefact NCCT; hemorrhagic stroke on index CT.	Patient-level; *n* = 25	DL classifier for AIS detection on NCCT	AIS present at the time of index NCCT	Expert CT interpretation by medical team
Chen 2022 [17]	Retrospective; Single; Taiwan	Unenhanced brain NCCT images labeled normal/hemorrhage/infarct/other; Not explicitly stated	Image-level; *n* = 2,476	DL classifiers (CNN-2, VGG-16, ResNet-50; transfer learning) — multi-class NCCT categorization.	Infarct class on NCCT within a multi-class image-level task	Clinical labels with follow-up and expert radiologist review.
Chiramal 2024 [18]	Retrospective; Single; India	Consecutive NCCT scans for suspected stroke with both AI output and teleradiology reports; Missing AI or missing teleradiology report.	Patient-level; n= 864	qER (Qure.ai) — automated NCCT brain interpretation.	AIS (infarct) present at the time of index NCCT	Teleradiologist report used as ground truth
Dyer 2021 [19]	Retrospective; Multi; UK, USA, and India	NCCT head scans with FRCR radiologist ground truth; Corrupted/incomplete scans	Patient-level; *n* = 390	EfficientNet-based DL model — NCCT classification ± weak segmentation.	AIS present at the time of index NCCT	Expert radiologist ground truth on CT
Fuchigami 2020 [20]	Retrospective; Single; Japan	Adults with hyperacute ischemic stroke (≤ 4.5 h) undergoing NCCT; Not stated	Patient-level; n= 70	3D U-Net with left–right comparator — hyperacute ischemic lesion segmentation on NCCT	AIS present at the time of index NCCT in the hyperacute window	MRI-DWI ± FLAIR within ~ 1 h
Gauriau 2023 [10]	Retrospective; Multi; USA	Admission NCCT for suspected AIS with MRI-DWI ≤ 3 h (positives)/≤5 days (negatives); earliest NCCT used; Non-diagnostic/artefact CT; failed/imprecise registrations.	Patient-level; n= 364	3D U-Net–based DL model — early infarct detection/segmentation on NCCT.	AIS present at the time of index NCCT	MRI-DWI ± ADC with clinical correlation; positives verified ≤ 3 h, negatives within ≤ 5 days.
Herweh 2016 [21]	Retrospective; Single; Germany	Acute stroke patients with CT and DWI < 2 h apart; suspected unilateral anterior circulation ischemia; Poor-quality/artefact CT.	Lesion-level; n= 680	e-ASPECTS — automated ASPECTS scoring on NCCT.	Ischemic severity on NCCT (ASPECTS)	MRI-DWI consensus
Lo 2021 [22]	Retrospective; Multi; Taiwan	Adults with AIS ≤ 6 h and ICA/M1 occlusion on CTA; Vertebrobasilar occlusion; arrival > 6 h; motion artefacts	Lesion-level; n= 225	Slice-level DCNNs (AlexNet, Inception-v3, ResNet-101; transfer) — NCCT classification.	AIS—LVO subtype (arterial occlusion surrogate) on NCCT	Multiphase CTA (occlusion presence and site).
Lu 2022 [23]	Retrospective; Multi; China	Suspected AIS with NCCT ≤ 24 h and MRI (incl. DWI) ≤ 72 h; Head injury/brain tumor; NCCT–DWI mismatch; strong CT artefacts.	Patient-level; n= 150	Two-stage DL model — YOLOv3 (localization) + ResNet-50 (classification) for early/invisible AIS on NCCT.	AIS present at the time of index NCCT	MRI-DWI plus clinical diagnosis.
Mallon 2023 [24]	Prospective; Single; UK	Consecutive acute stroke admissions with NCCT (CTA/CTP per pathway); No index NCCT; mass lesion on initial CT	Patient-level; n= 492	Brainomix e-Stroke (e-ASPECTS) — automated ASPECTS analysis on NCCT.	AIS—severity subtype on NCCT using ASPECTS-based patient-level thresholding	Expert neuroradiologist ground truth using all imaging and clinical data
Pastor-Vargas 2025 [25]	Retrospective; Single; Spain	NCCT for suspected ischemia; Hemorrhage or non-ischemic conditions on NCCT.	Patient-level; n= 316	DenseNet-3D CNN — ischemia detection on NCCT (transfer learning).	AIS present at the time of index NCCT	MRI-DWI with clinical follow-up.
Sahoo 2022 [26]	Retrospective; Single; Taiwan	Adults with ischemic stroke onset < 12 h; DWI infarct < 0.5 cm; traumatic brain injury; brain malignancy; ICH	Patient-level; n= 58	Customized VGG-16 CNN — ischemic lesion identification on NCCT (2D slices).	AIS present at the time of index NCCT	MRI-DWI

Table 1 (continued)

Study ID	Design; Centers; Country	Inclusion criteria; Exclusion criteria	Unit of analysis; Sample size (*n*)	Index test name	Target condition	Reference standard
Wei 2025 [27]	Retrospective; Multi; China	Adults with suspected AIS, last-known-well ≤ 8 h, undergoing baseline NCCT (ASPECTS tool); Intracranial hemorrhage; prior craniotomy; traumatic ICH; poor-quality images precluding diagnosis	Patient-level; n= 220	DL-based automated ASPECTS system — 3D ResNet (localize) + 2.5D U-Net (segment) + 2.5D ResNet (score).	AIS—severity subtype on NCCT using ASPECTS-based patient-level thresholding	Expert consensus ASPECTS on NCCT (± CTP if available)
Weyland 2022 [7]	Retrospective; Multi; Germany, Greece, and UK	Acute stroke or stroke-like symptoms with both NCCT and CTA available (HAS/LVO task); Not explicitly stated	Patient-level; n= 154	Brainomix HAS detector — AI-driven hyperdense artery sign detection on NCCT.	AIS—LVO subtype on NCCT	CTA for LVO presence and side.
Wu 2021 [28]	Retrospective; Multi; China	Stroke cases with NCCT and DWI (onset-to-NCCT < 24 h); Not explicitly stated	Patient-level; n= 77	Two-stage CNN — U-Net (segmentation) + ResNet (patch classifier) with MAP post-processing on NCCT.	AIS present at the time of index NCCT	MRI-DWI for positives ± clinical confirmation
Yedavalli 2023 [29]	Retrospective; Multi; USA	Acutely presenting suspected stroke with index NCCT; LVO defined (distal intracranial ICA and/or M1 MCA); Not explicitly detailed.	Patient-level; n= 244	RAPID NCCT Stroke — automated NCCT analysis platform.	AIS—LVO subtype within suspected AIS cohort	CTA for LVO with consensus of two of three neuroradiologists.

NCCT non-contrast, *CT AIS* acute ischemic stroke;, *LVO* large-vessel occlusion, *HAS* hyperdense artery sign, *CTA* CT angiography, *MRI-DWI/ADC* diffusion-weighted MRI/apparent diffusion coefficient, *ASPECTS* Alberta Stroke Program Early CT Score, *CTP* CT perfusion, *MCA* middle cerebral artery. *Unit of analysis* rows marked *patient-level* contributed to the primary HSROC meta-analysis, *image- or lesion-level* rows are descriptive only (not pooled). Moderators test-level covariates (time-to-CT window, ASPECTS burden, posterior fossa inclusion, RS robustness, validation type, operating-point selection, mode of use) are encoded in the supplementary matrix with codes 0/1/2 as pre-specified. Vendor/platform names appear as reported; custom models are summarized by backbone + task

Results

Study selection

A systematic search of MEDLINE, Embase, and Web of Science (January 2010–May 2025) yielded 1,899 records. After removing 390 duplicates, 1,509 titles/abstracts were screened, of which 1,403 were excluded as irrelevant (e.g., non-diagnostic focus, conventional CAD/non-DL, non-NCCT, lesion-only reports, conference abstracts, or insufficient outcome data). Full texts of 106 articles were assessed; 90 were excluded (common reasons: two-gate case–control designs, absence of patient-level outcomes or a reconstructible 2 × 2 table, reliance on non-robust/inadequately verified reference standards, or index tests without a clear operating point). Sixteen studies met inclusion; 13 provided patient-level 2 × 2 data and were pooled; 3 reported lesion/image-level accuracy only and were summarised narratively (not pooled). PRISMA flow is shown in Fig. 1.

Sixteen studies (15 retrospective, 1 prospective) spanned Asia, Europe, and North America, evenly split between single-centre (8/16) and multicentre (8/16) cohorts. Adults with suspected acute stroke underwent NCCT; several focused on hyperacute windows, LVO/HAS, or ASPECTS-defined severity. Thirteen studies reported patient-level AIS detection and entered the meta-analysis; three were lesion/image-level and were summarised descriptively. Patient-level sample sizes ranged widely (≈ 25–864). Index tests included commercial tools (qER/Qure.ai, Brainomix e-Stroke/e-ASPECTS, RAPID NCCT Stroke) and research CNN/3D-U-Net pipelines. Reference standards were most often MRI-DWI (with clinical follow-up), or expert consensus where MRI was unavailable (Table 1). A test-level covariate matrix is provided in the Supplementary Materials, Table S4.

Risk of bias and reporting quality

QUADAS-2

Across 16 studies overall, risk-of-bias concerns peaked in Patient Selection (High 10/16, 62.5%; Unclear 4/16, 25.0%; Low 2/16, 12.5%) and Index Test (Unclear 8/16, 50.0%; High 4/16, 25.0%; Low 4/16, 25.0%). Reference Standard was mostly Unclear (13/16, 81.2%) with Low 3/16 (18.8%) and no High. Flow/Timing had the highest Low proportion (6/16, 37.5%) yet notable High (5/16, 31.2%) and Unclear (5/16, 31.2%). Applicability was largely acceptable for Index Test (Low 15/16, 93.8%) and Reference Standard (Low 11/16, 68.8%), whereas Patient Selection often raised concerns (High 11/16, 68.8%). (Fig. 2)

AI reporting (STARD-AI/CLAIM items)

Data source was described in 16/16 (100%) studies. Training/test splits were reported in 12/16 (75.0%)

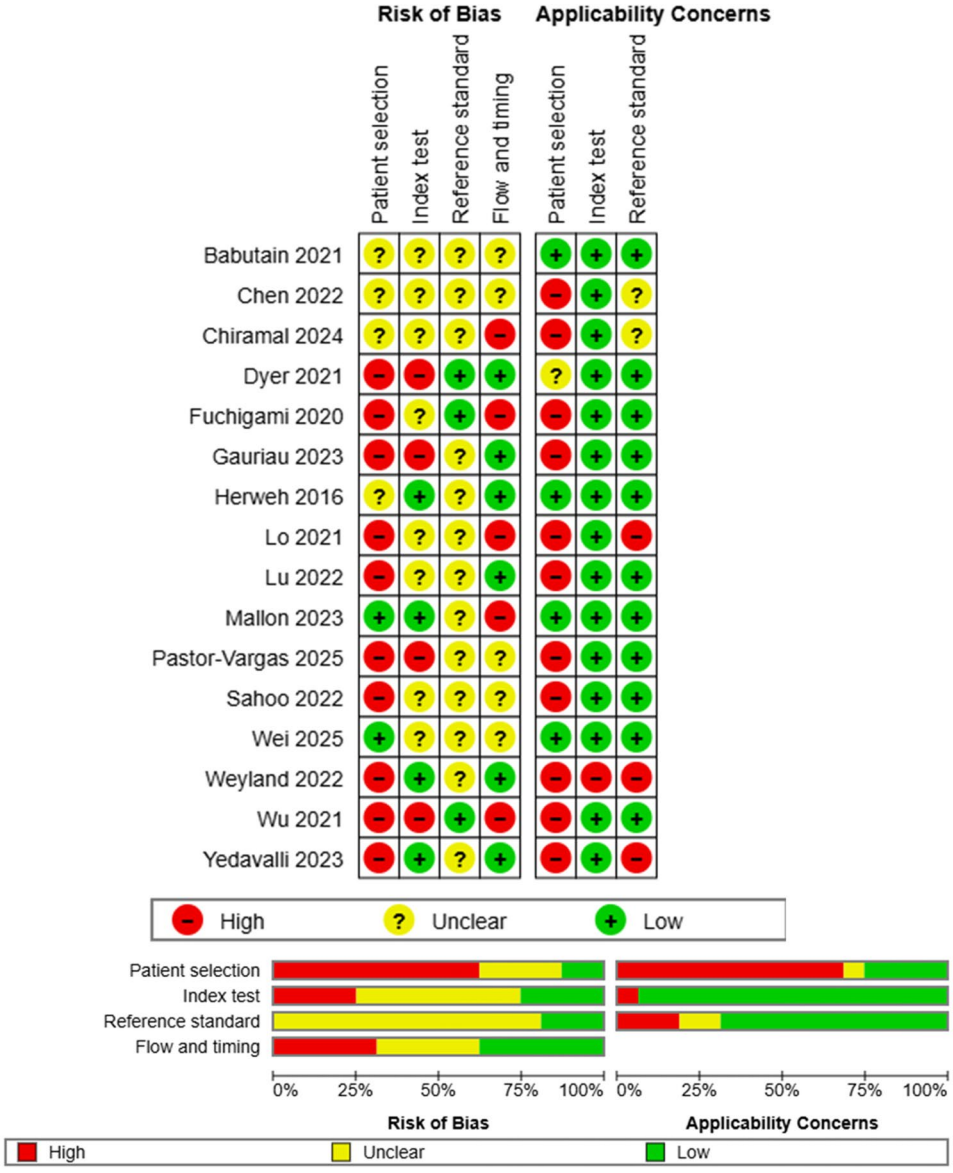

Fig. 2 Visual summary of the QUADAS-2 assessment

(Unclear 1/16, 6.3%; Not reported 3/16, 18.8%). External validation appeared in 3/16 (18.8%) (Unclear 7/16, 43.8%; Not reported 6/16, 37.5%). Explainability was reported in 7/16 (43.8%) (Unclear 1/16, 6.3%; Not reported 8/16, 50.0%). Model availability (code/weights or named commercial system) was reported in 6/16 (37.5%); workflow integration in 10/16 (62.5%) (Unclear 4/16, 25.0%); and regulatory status in 3/16 (18.8%). Overall, 1/16 (6.3%) study met the Good rubric, 11/16 (68.8%) were Moderate, and 4/16 (25.0%) were Poor (Table 2).

Results of individual studies

Across the 13 pooled arms, study-level sensitivity ranged from 58.6% to 100.0% (median 88.6%, IQR 79.5–95.4%) and specificity from 63.8% to 98.0% (median 84.0%, IQR 77.2–90.3%). Individual estimates with 95% CIs are shown in Fig. 3.

Meta-analysis

Base HSROC model

Across the 13 pooled arms, the bivariate/HSROC summary point was sensitivity 0.91 (95% CI 0.81–0.96) and specificity 0.90 (0.85–0.94). Overall discrimination was strong (see DOR and likelihood ratios in Table 3). Heterogeneity was greater for specificity than sensitivity; the positive Se–Sp correlation (0.41; 95% CI – 0.27–0.82) suggests between-study heterogeneity was not dominated by simple threshold shifts; rather, cohorts with clearer signal

Table 2 AI-specific reporting quality assessment

Study ID	Data Source Described	Training/ Test Split Reported	External Validation	Model Explainability	Model Availability	Clinical Workflow Integration	Regulatory Status Discussed	Overall Assessment
Babutain 2021 [16]	Reported	Reported	Not reported	Not reported	Not reported	Unclear	Not reported	Poor
Chen 2022 [17]	Reported	Reported	Not reported	Not reported	Not reported	Reported	Not reported	Poor
Chiramal 2024 [18]	Reported	Not reported	Unclear	Reported	Reported	Reported	Reported	Moderate
Dyer 2021 [19]	Reported	Reported	Unclear	Reported	Not reported	Reported	Not reported	Moderate
Fuchigami 2020 [20]	Reported	Reported	Not reported	Reported	Not reported	Not reported	Not reported	Poor
Gauriau 2023 [10]	Reported	Reported	Not reported	Reported	Not reported	Reported	Not reported	Moderate
Herweh 2016 [21]	Reported	Reported	Unclear	Not reported	Reported	Reported	Reported	Moderate
Lo 2021 [22]	Reported	Reported	Reported	Not reported	Not reported	Unclear	Not reported	Moderate
Lu 2022 [23]	Reported	Reported	Reported	Reported	Reported	Reported	Not reported	Good
Mallon 2023 [24]	Reported	Not reported	Unclear	Unclear	Reported	Reported	Not reported	Moderate
Pastor-Vargas 2025 [25]	Reported	Reported	Not reported	Not reported	Not reported	Unclear	Not reported	Poor
Sahoo 2022 [26]	Reported	Reported	Not reported	Not reported	Not reported	Reported	Not reported	Moderate
Wei 2025 [27]	Reported	Reported	Unclear	Reported	Not reported	Reported	Not reported	Moderate
Weyland 2022 [7]	Reported	Unclear	Unclear	Reported	Reported	Unclear	Not reported	Moderate
Wu 2021 [28]	Reported	Reported	Reported	Not reported	Not reported	Not reported	Not reported	Moderate
Yedavalli 2023 [29]	Reported	Not reported	Unclear	Not reported	Reported	Reported	Reported	Moderate

Seven AI reporting items adapted from STARD-AI/CLAIM/CONSORT-AI were evaluated: (1) data source described; (2) training/test split reported; (3) external validation on an independent dataset; (4) model explainability (e.g., saliency/attribution); (5) model availability (code/weights or named commercial system); (6) clinical workflow integration (mode of use, reader-study/deployment details); and (7) regulatory status discussed. Each item was graded as reported, Unclear, or not reported. Overall categories: Good = all three core items (training/test split, external validation, explainability) reported and ≥ 5/7 items reported (with "Unclear" not counting toward the total and ≤ 1 other item "Not reported"); Moderate = 3–4/7 items reported, or ≥ 5/7 reported with ≥ 1 core item Unclear/Not reported; Poor = ≤ 2/7 items reported, or ≥ 4 items not reported, or both training/test split and external validation not reported

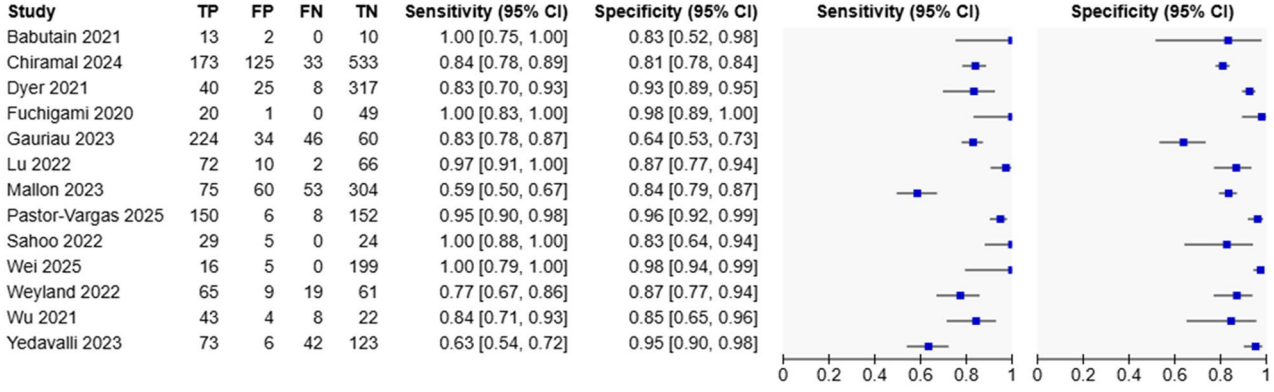

Fig. 3 Coupled forest plots of sensitivity and specificity

tended to improve in both metrics. HSROC parameters were $\lambda = 4.59$ (95% CI 3.41–5.77), $\theta = -0.44$ (−1.24 to 0.36), $\beta = -0.42$ (−1.15 to 0.32; not different from 0), with between-study variances $s^2\alpha = 3.05$ (1.10–8.49) and $s^2\theta = 0.32$ (0.11–0.94) (Fig. 4) Prespecified covariates eligible for both subgroup analysis and univariable meta-regression were: posterior fossa (excluded vs. included), reference-standard robustness (mixed/weaker vs. robust), and validation type (internal vs. external).

Subgroup analyses

- Posterior fossa: Including posterior fossa was associated with higher sensitivity and slightly lower specificity versus anterior-only cohorts; overall discrimination favoured inclusion.
- Reference-standard robustness: Pooled performance was similar under robust versus mixed/weaker standards; no consistent difference emerged.

Table 3 Summary of pooled estimates

Subgroup	Studies (*n*)	Sensitivity (95% CI)	Specificity (95% CI)	DOR (95% CI)	LR+ (95% CI)	LR– (95% CI)
Overall	13	0.91 (0.81–0.96)	0.90 (0.85–0.94)	90.29 (29.32–278.04)	9.27 (5.64–15.24)	0.10 (0.05–0.23)
Posterior fossa Included						
Yes (Included)	7	0.95 (0.83–0.98)	0.88 (0.78–0.93)	124.22 (24.64–626.22)	7.65 (4.05–14.46)	0.06 (0.02–0.21)
Anterior only (Excluded)	4	0.77 (0.53–0.91)	0.93 (0.85–0.97)	41.66 (7.39–234.78)	10.48 (4.04–27.22)	0.25 (0.10–0.60)
Reference standard robustness						
Robust (DWI-MRI for positives; negatives verified by MRI or adequate follow-up)	6	0.90 (0.76–0.97)	0.89 (0.78–0.94)	73.53 (19.82–272.77)	7.90 (4.02–15.52)	0.11 (0.04 – 0.29)
Mixed/Weaker	7	0.92 (0.73–0.98)	0.92 (0.85–0.96)	133.80 (18.38–973.95)	11.46 (5.51–23.83)	0.09 (0.02–0.34)
Validation type						
External	6	0.82 (0.67–0.91)	0.88 (0.83–0.92)	33.37 (14.32–77.78)	6.95 (4.60–10.49)	0.21 (0.11–0.39)
Internal	6	0.95 (0.89–0.98)	0.93 (0.82–0.97)	258.36 (41.20–1620.32)	13.29 (4.87–36.27)	0.05 (0.02–0.13)
Direct AIS construct subset (Sensitivity analyses)						
Direct AIS subset	9	0.92 (0.86–0.96)	0.89 (0.81–0.94)	94.73 (31.09–288.67)	8.24 (4.63–14.69)	0.09 (0.04–0.17)

Pooled sensitivities (Se) and specificities (Sp) with 95% CIs were obtained from random–effects bivariate/HSROC models. DOR and likelihood ratios (LR + and LR–) are derived from the bivariate summary point. Subgroup rows correspond to prespecified moderators (Posterior fossa Included, reference standard robustness, validation type) analyzed in separate bivariate fits. Abbreviations: Se, sensitivity; Sp, specificity; LR+, positive likelihood ratio; LR–, negative likelihood ratio; DOR, diagnostic odds ratio; CI, confidence interval

- Validation type: External validation tended to show lower sensitivity with broadly comparable specificity relative to internal splits, consistent with expected generalisability effects.

Meta-regression (univariable)

- Posterior fossa (excluded vs. included): Excluding posterior fossa was associated with a significant decrease in sensitivity, while specificity did not differ significantly (Table 4).
- Reference-standard robustness (mixed/weaker vs. robust): No significant associations were detected for either sensitivity or specificity.
- Validation type (internal vs. external): Internal validation was associated with higher sensitivity; the specificity difference was not significant.

Additional analyses

- External validation only: Sensitivity was lower while specificity was broadly comparable to the primary model.
- Robust reference standards only: Estimates closely matched the base analysis, indicating robustness to reference-standard choice.
- Posterior fossa included (excluding anterior-only) Sensitivity improved with a small trade-off in specificity, consistent with subgroup/meta-regression patterns.
- Direct AIS construct only (patient-level AIS detection; excluding CTA-defined LVO/HAS and ASPECTS-only tools): Performance was essentially unchanged from the base HSROC, with high sensitivity, high specificity, and strong overall discrimination.

Discussion

Principal findings

Across included studies, DL algorithms applied to NCCT show high sensitivity and high specificity for AIS detection, with performance generally exceeding historical radiologist sensitivity on early CT signs and posterior fossa infarcts and approaching MRI benchmarks in in selected contexts [2, 4, 30]. External validation cohorts consistently showed lower sensitivity than internal tests, underscoring domain shift and generalization gaps that are well-documented in medical imaging AI [31].

Comparison with prior reviews

Our results align with broader reviews of AI in stroke imaging that highlight promising accuracy but limited

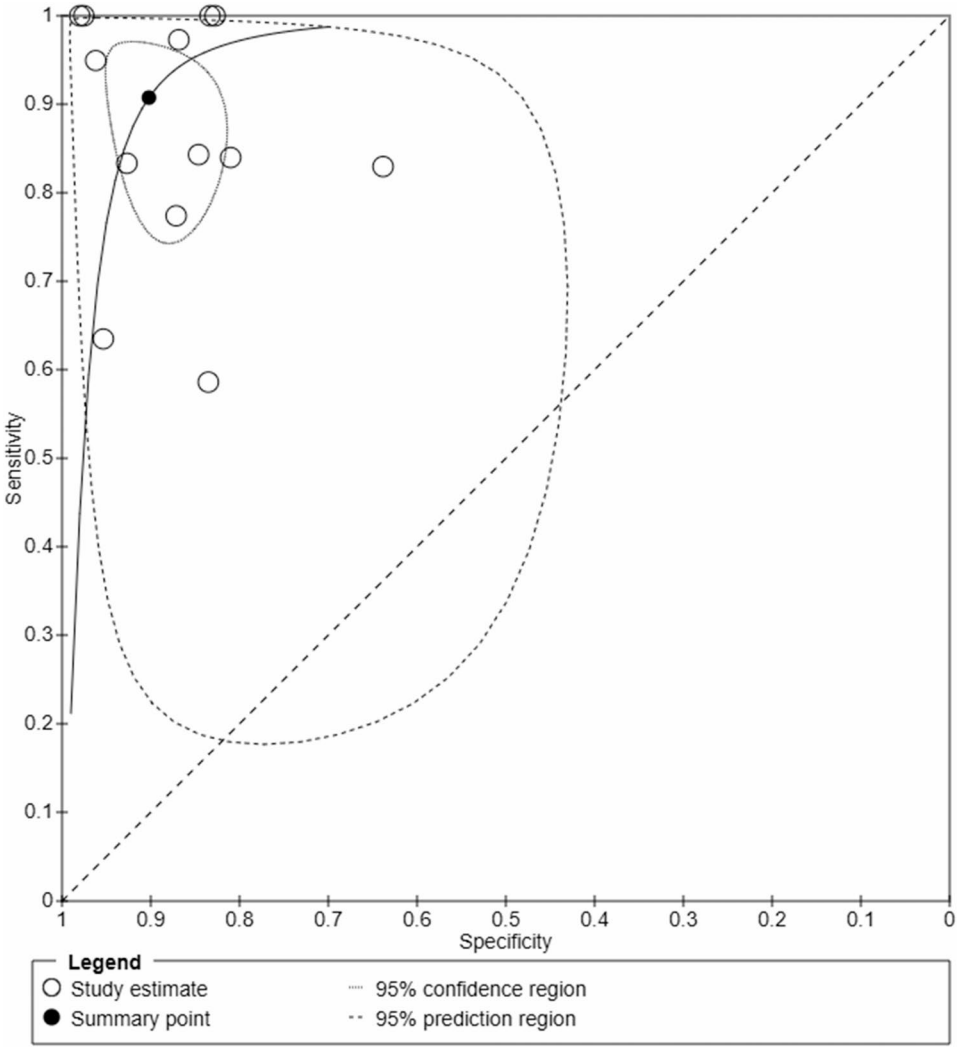

Fig. 4 HSROC Plot

prospective and external validation [11]. Importantly, an NCCT-specific meta-analysis reached similar conclusions: strong internal validation performance contrasted by substantially attenuated sensitivity on external datasets and notable heterogeneity and bias risks [32]. Together, these syntheses and our patient-level meta-analysis converge on the same message: DL for NCCT-AIS is accurate in principle, but robust generalizability remains the rate-limiting step to clinical translation.

Clinical Implications (rule-out vs. rule-in; workflow impact)

Given NCCT's speed and ubiquity, DL may be best positioned as a rule-in adjunct to highlight subtle early ischemic changes (and posterior fossa infarcts) that are often missed by humans, potentially expediting CTA/CTP/MRI and treatment decisions [2, 4, 30]. Caution is warranted for rule-out use, especially in settings with domain shift; thresholding should reflect local disease prevalence and downstream actions. To quantify net clinical value beyond AUC and accuracy, decision-curve analysis (DCA) should be routinely reported; this clarifies when an AI-assisted pathway yields higher net benefit than "scan all" or "scan none" strategies across clinically plausible thresholds [33, 34].

DL can also reshape workflows: (i) pre-read triage and worklist prioritization; (ii) real-time decision support for non-neuroradiologists in EDs without 24/7 subspecialty coverage; and (iii) standardized quantification of early ischemic change to support thrombolysis/thrombectomy selection. Sites adopting AI should pair deployment with local calibration and ongoing performance monitoring.

Reporting gaps and study limitations

Methodological and reporting quality varied. Frequent issues included unclear reference standards, leakage between development and test sets, minimal error analysis, and sparse reporting of model versioning, governance, and human–AI interaction. Adherence to

Table 4 Univariable meta-regression

Moderator	k_Se	tau2_Se	I2%_Se	Se_level0	Se_level1	Se_Δprob (1 – 0)	Se_moderator_logit	Se_interpretation	k_Sp	tau2_Sp	I2%_Sp	Sp_level0	Sp_level1	Sp_Δprob (1 – 0)	Sp_moderator_logit	Sp_interpretation
Posterior fossa included	11	0.348	64.5	Included(0): 0.896 (95% CI 0.819–0.943)	Excluded(1): 0.698 (95% CI 0.537–0.822)	−0.198	−1.316 (95% CI −2.264– −0.368); $p = 0.007$	Excluded(1) shows lower sensitivity than Included(0) (Δ≈−0.198; 0.90→0.70); statistically significant ($p = 0.007$). Heterogeneity: I^2=64.5% (substantial).	11	0.66	88.1	Included(0): 0.861 (95% CI 0.753–0.926)	Excluded(1): 0.924 (95% CI 0.835–0.966)	0.063	0.671 (95% CI −0.449– 1.792); $p = 0.240$	Excluded(1) shows higher specificity than Included(0) (Δ≈+0.063; 0.86→0.92); not statistically significant ($p = 0.240$). Heterogeneity: I^2=88.1% (high).
Reference standard robustness	13	1.007	87.8	Robust(0): 0.882 (95% CI 0.751–0.948)	MixedWeaker(1): 0.855 (95% CI 0.702–0.936)	−0.027	−0.237 (95% CI −1.520– 1.045); $p = 0.717$	MixedWeaker(1) shows lower sensitivity than Robust(0) (Δ≈−0.027; 0.88→0.85); not statistically significant ($p = 0.717$). Heterogeneity: I^2=87.8% (high).	13	0.771	88.8	Robust(0): 0.881 (95% CI 0.774–0.941)	MixedWeaker(1): 0.904 (95% CI 0.818–0.952)	0.023	0.245 (95% CI −0.819– 1.309); $p = 0.652$	MixedWeaker(1) shows higher specificity than Robust(0) (Δ≈+0.023; 0.88→0.90); not statistically significant ($p = 0.652$). Heterogeneity: I^2=88.8% (high).
Validation type	12	0.794	85.1	External(0): 0.807 (95% CI 0.660–0.900)	Internal(1): 0.940 (95% CI 0.851–0.978)	0.133	1.327 (95% CI 0.051–2.603); $p = 0.041$	Internal(1) shows higher sensitivity than External(0) (Δ≈+0.133; 0.81→0.94); statistically significant ($p = 0.041$). Heterogeneity: I^2=85.1% (high).	12	0.87	89.8	External(0): 0.883 (95% CI 0.775–0.943)	Internal(1): 0.909 (95% CI 0.807–0.960)	0.026	0.276 (95% CI −0.897– 1.449); $p = 0.644$	Internal(1) shows higher specificity than External(0) (Δ≈+0.026; 0.88→0.91); not statistically significant ($p = 0.644$). Heterogeneity: I^2=89.8% (high).

Using logit-PLO REML models and restricting moderators to 0/1 (dropping "unclear"), exclusion of the posterior fossa (anterior-only; k=11) was associated with lower sensitivity (Δlogit −1.32, p=0.007; predicted 0.896→0.698) with no significant change in specificity (Δlogit +0.67, p=0.240; 0.861→0.924); heterogeneity was moderate for sensitivity (I^2=64.5%) and high for specificity (I^2=88.1%). Reference-standard robustness (robust vs mixed/weaker; k=13) showed no material effect on sensitivity (Δlogit −0.24, p=0.717; 0.882→0.855) or specificity (Δlogit +0.25, p=0.652; 0.881→0.904), with high heterogeneity (I^2≈88–89%). Validation type (external vs internal; k=12) showed higher sensitivity with internal validation (Δlogit +1.33, p=0.041; 0.807→0.940) and no significant difference in specificity (Δlogit +0.28, p=0.644; 0.883→0.909); heterogeneity remained high (I^2≈85–90%)

CLAIM, CONSORT-AI/SPIRIT-AI, PRISMA-DTA for reviews, and QUADAS-2 for primary diagnostic studies would materially improve transparency and reproducibility [12–15]. Generalizability was the dominant limitation: performance declined on external cohorts, consistent with known site- and scanner-shift effects [31]. Equity risks are non-trivial: AI systems encode latent demographic signals (including race) even when these are not visible to experts, creating potential for disparate performance if training data are geographically or demographically narrow [35]. Many included studies were developed primarily on datasets from high-income, Western settings; broader international, multi-vendor datasets and subgroup reporting are needed.

Strengths of this review

We conducted a comprehensive, protocol-driven synthesis with patient-level meta-analysis where feasible, emphasized external validation and subgroup analyses, and interpreted accuracy alongside potential clinical utility (via DCA concepts) rather than accuracy metrics alone. We also mapped reporting to contemporary guidelines to facilitate reproducibility and appraisal [12–15].

Future Research Directions

Prospective and pragmatic trials should evaluate AI-assisted NCCT pathways on patient-centered outcomes, designed and reported per CONSORT-AI/SPIRIT-AI and early-stage pilot work aligned with DECIDE-AI; future diagnostic-accuracy papers should follow STARD-AI [14, 15, 36, 37].

Explainability and failure analysis: standardized error taxonomies, case-level rationales, and model versioning.

Clinical utility & economics: routine DCA (with uncertainty intervals) and cost-effectiveness analyses tied to local prevalence and downstream imaging/treatment cascades [33, 34].

Generalizability & equity: multi-region, multi-vendor datasets; pre-specified subgroup analyses; fairness audits given evidence that imaging AI can encode protected characteristics [35].

Deployment science: calibration drift monitoring, human–AI teaming studies, and learning-health-system feedback loops.

Conclusion

DL applied to NCCT shows high diagnostic accuracy for AIS and clear integration potential for triage and decision support. The core gaps are rigorous reporting, external validity, equity, and real-world impact evaluation. Addressing these with guideline-concordant prospective studies and utility/economic analyses should enable safe, effective adoption.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12883-025-04528-3.

Supplementary Material 1.

Acknowledgements
The authors would like to thank the librarians and information specialists who provided assistance during the systematic search.

Institutional review board statement
Not applicable.

Informed consent statement
Not applicable.

Conflicts of interest
The authors declare no conflict of interest.

Authors' contributions
Conceptualization, Kalab Yigermal Gete and Asnakew Achaw Ayele; Methodology, Kalab Yigermal Gete and Asnakew Achaw Ayele; Software, Kalab Yigermal Gete; Validation, Kalab Yigermal Gete and Asnakew Achaw Ayele; Formal analysis, Kalab Yigermal Gete; Investigation, Kalab Yigermal Gete and Asnakew Achaw Ayele; Resources, Kalab Yigermal Gete; Data curation, Kalab Yigermal Gete and Asnakew Achaw Ayele; Writing—original draft, Kalab Yigermal Gete; Writing—review & editing, Kalab Yigermal Gete and Asnakew Achaw Ayele; Visualization, Kalab Yigermal Gete; Supervision, Asnakew Achaw Ayele; Project administration, Kalab Yigermal Gete; Funding acquisition, Not applicable. Study selection and data extraction were performed independently and in duplicate by Kalab Yigermal Gete and Asnakew Achaw Ayele; disagreements were resolved by consensus.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability
All data generated or analyzed during this study are included in this published article and its supplementary materials. Additional data extraction sheets and meta-analysis code are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]School of Medicine, College of Medicine and Health Sciences, Bahir Dar University, Bahir Dar, Ethiopia
[2]EPIC Health Systems, Addis Ababa, Ethiopia
[3]School of Health, Faculty of Medicine and Health, University of New England, Armidale 2351, Australia
[4]Department of Clinical Pharmacy, School of Pharmacy, College of Medicine and Health Science, University of Gondar, Gondar, Ethiopia

Received: 19 August 2025 / Accepted: 7 November 2025
Published online: 15 December 2025

References

1. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2019;50(12):e344-418. https://doi.org/10.1161/STR.0000000000000211.

2. Chalela JA, Kidwell CS, Nentwich LM, Luby M, Butman JA, Demchuk AM, et al. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. Lancet. 2007;369(9558):293–8. https://doi.org/10.1016/S0140-6736(07)60151-2.
3. Jazayeri SB, Ghozy S, Hasanzadeh A, Elfil M, Ahmadzade A, Naseh E et al. MR Imaging versus Noncontrast CT for Selecting Patients with Acute Ischemic Stroke of Large Vessel Occlusion for Endovascular Thrombectomy: A Systematic Review and Meta-Analysis. American Journal of Neuroradiology. 2025;46(10):2026. Available from: http://www.ajnr.org/content/46/10/2026.abstract.
4. Hwang DY, Silva GS, Furie KL, Greer DM. Comparative sensitivity of computed tomography vs. magnetic resonance imaging for detecting acute posterior fossa infarct. J Emerg Med. 2012;42(5):559–65. https://doi.org/10.1016/j.jemermed.2011.05.101.
5. Ahmed RA, Dmytriw AA, Regenhardt RW, Leslie-Mazwi TM, Hirsch JA. Posterior circulation cerebral infarction: A review of clinical, imaging features, management, and outcomes. Eur J Radiol Open. 2023;11. Available from: https://doi.org/10.1016/j.ejro.2023.100523.
6. Mair G, Boyd EV, Chappell FM, von Kummer R, Lindley RI, Sandercock P, et al. Sensitivity and specificity of the hyperdense artery sign for arterial obstruction in acute ischemic stroke. Stroke. 2015;46(1):102–7. https://doi.org/10.1161/STROKEAHA.114.007036.
7. Weyland CS, Papanagiotou P, Schmitt N, Joly O, Bellot P, Mokli Y, Ringleb PA, Kastrup A, Möhlenbruch MA, Bendszus M, Nagel S. Hyperdense artery sign in patients with acute ischemic stroke–automated detection with artificial intelligence-driven software. Frontiers in neurology. 2022 Apr 5;13:807145.
8. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. Lancet. 2000;355(9216):1670–4. https://doi.org/10.1016/S0140-6736(00)02237-6.
9. Pexman JHW, Barber PA, Hill MD, Sevick RJ, Demchuk AM, Hudon ME et al. Use of the Alberta Stroke Program Early CT Score (ASPECTS) for Assessing CT Scans in Patients with Acute Stroke. American Journal of Neuroradiology. 2001;22(8):1534. Available from: http://www.ajnr.org/content/22/8/1534.abstract.
10. Gauriau R, Bizzo BC, Comeau DS, Hillis JM, Bridge CP, Chin JK, Pawar J, Pourvaziri A, Sesic I, Sharaf E, Cao J. Head CT deep learning model is highly accurate for early infarct estimation. Scientific reports. 2023 Jan 5;13(1):189.
11. Liu Y, Wen Z, Wang Y, Zhong Y, Wang J, Hu Y et al. Artificial intelligence in ischemic stroke images: current applications and future directions. Front Neurol. 2024;Volume 15-2024. Available from: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1418060.
12. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Group and the PD. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA [Internet]. 2018;319(4):388–96. https://doi.org/10.1001/jama.2017.19163.
13. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(8):529–36. Available from: https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
14. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. SPIRIT-AI, Group CAW, Ashrafian H, Beam AL, Chan AW, Collins GS, editors. BMJ. 2020;370. Available from: https://www.bmj.com/content/370/bmj.m3164.
15. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology: Artificial Intelligence. 2020;2(2):e200029. https://doi.org/10.1148/ryai.2020200029.
16. Babutain K, Hussain M, Aboalsamh H, Al-Hameed M. Deep learning-enabled detection of acute ischemic stroke using brain computed tomography images. Int J Adv Comput Sci Appl. 2021;12(12):386–97.
17. Chen YT, Chen YL, Chen YY, Huang YT, Wong HF, Yan JL et al. Deep learning–Based brain computed tomography image classification with hyperparameter optimization through transfer learning for stroke. Diagnostics. 2022;12(4).
18. Chiramal JA, Johnson J, Webster J, Nag DR, Robert D, Ghosh T et al. Artificial Intelligence-based automated CT brain interpretation to accelerate treatment for acute stroke in rural india: an interrupted time series study. PLOS Global Public Health. 2024;4(7).
19. Dyer T, Chawda S, Alkilani R, Morgan TN, Hughes M, Rasalingham S. Validation of an artificial intelligence solution for acute triage and rule-out normal of non-contrast CT head scans. Neuroradiology. 2022;64(4):735–43.
20. Fuchigami T, Akahori S, Okatani T, Li Y. A hyperacute stroke segmentation method using 3D U-Net integrated with physicians' knowledge for NCCT. In SPIE-Intl Soc Optical Eng. 2020;15.
21. Herweh C, Ringleb PA, Rauch G, Gerry S, Behrens L, Möhlenbruch M, et al. Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. Int J Stroke. 2016;11(4):438–45.
22. Lo CM, Hung PH, Lin DT. Rapid assessment of acute ischemic stroke by computed tomography using deep convolutional neural networks. J Digit Imaging. 2021;34(3):637–46.
23. Lu J, Zhou Y, Lv W, Zhu H, Tian T, Yan S, et al. Identification of early invisible acute ischemic stroke in non-contrast computed tomography using two-stage deep-learning model. Theranostics. 2022;12(13):5564–73.
24. Mallon D, Fallon M, Blana E, McNamara C, Menon A, Ip CL et al. Real-world evaluation of brainomix e-Stroke software. Stroke Vasc Neurol. 2023;svn-2023-002859.
25. Pastor-Vargas R, Antón-Munárriz C, Haut JM, Robles-Gómez A, Paoletti ME, Benítez-Andrades JA. Cerebral ischemia detection using deep learning techniques. Health Information Science and Systems. 2025 May 20;13(1):36.
26. Sahoo PK, Mohapatra S, Wu CY, Huang KL, Chang TY, Lee TH. Automatic identification of early ischemic lesions on non-contrast CT with deep learning approach. Scientific Reports. 2022 Oct 27;12(1):18054.
27. Wei J, Shang K, Wei X, Zhu Y, Yuan Y, Wang M et al. Deep learning-based automatic ASPECTS calculation can improve diagnosis efficiency in patients with acute ischemic stroke: a multicenter study. Eur Radiol. 2024.
28. Wu G, Chen X, Lin J, Wang Y, Yu J. Identification of invisible ischemic stroke in noncontrast CT based on novel two-stage convolutional neural network model. Medical Physics. 2021 Mar;48(3):1262-75.
29. Yedavalli V, Heit JJ, Dehkharghani S, Haerian H, Mcmenamy J, Honce J et al. Performance of RAPID Noncontrast CT stroke platform in large vessel occlusion and intracranial hemorrhage detection. Front Neurol. 2023;14.
30. Wardlaw JM, Mielke O. Early signs of brain infarction at CT: observer reliability and outcome after thrombolytic treatment—systematic review. Radiology. 2005;235(2):444–53. https://doi.org/10.1148/radiol.2352040262.
31. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683-. https://doi.org/10.1371/journal.pmed.1002683.
32. Shen W, Peng J, Lu J. Artificial Intelligence for Ischemic Stroke Detection in Non-contrast CT: A Systematic Review and Meta-analysis. Acad Radiol. 2025; Available from: https://doi.org/10.1016/j.acra.2025.09.010.
33. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006 Nov;26(6):565-74. https://doi.org/10.1177/0272989X06295361.
34. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and prognostic research. 2019 Oct 4;3(1):18. https://doi.org/10.1186/s41512-019-0064-7.
35. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health [Internet]. 2022 Jun 1;4(6):e406–14. Available from: https://doi.org/10.1016/S2589-7500(22)00063-2 https://link.springer.com/article/10.1186/s41512-019-0064-7.
36. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924–33. https://doi.org/10.1038/s41591-022-01772-9.
37. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open [Internet]. 2021;11(6). Available from: https://bmjopen.bmj.com/content/11/6/e047709 https://bmjopen.bmj.com/content/11/6/e047709.

Publisher's Note