

MIND THE GAP:

Assessing the Generalizability of a publicly available AI Model for Intracranial Hemorrhage Detection on Non-Contrast CT



Rabia Shaukat¹, Ali Mansoor², Fatima Iqbal³, Muhammad Arslan³, Muneeb Ahmad², Nighat Haroon Khan²



1 Eastern Medical Technology Services.Lahore, Pakistan

2 Lahore General Hospital/Post Graduate Medical Institute/ Ameer ud Din Medical College, Pakistan

3 Jinnah Hospital/Allama Iqbal Medical College, Lahore, Pakistan

Background

Spontaneous intracranial haemorrhage (ICH) carries high mortality (30–50%) and significant long-term disability (1,2). In low- and middle-income countries such as Pakistan, delayed diagnosis is worsened by limited access to expert imaging interpretation (3). Early detection on non-contrast CT is therefore critical. Artificial intelligence (AI), particularly deep learning models, has shown strong performance in ICH detection, with reported AUCs of 0.85–0.96 (4). However, approved AI tools remain limited in LMIC settings, and real-world comparative data against clinical readers in Pakistan are scarce.

Objective

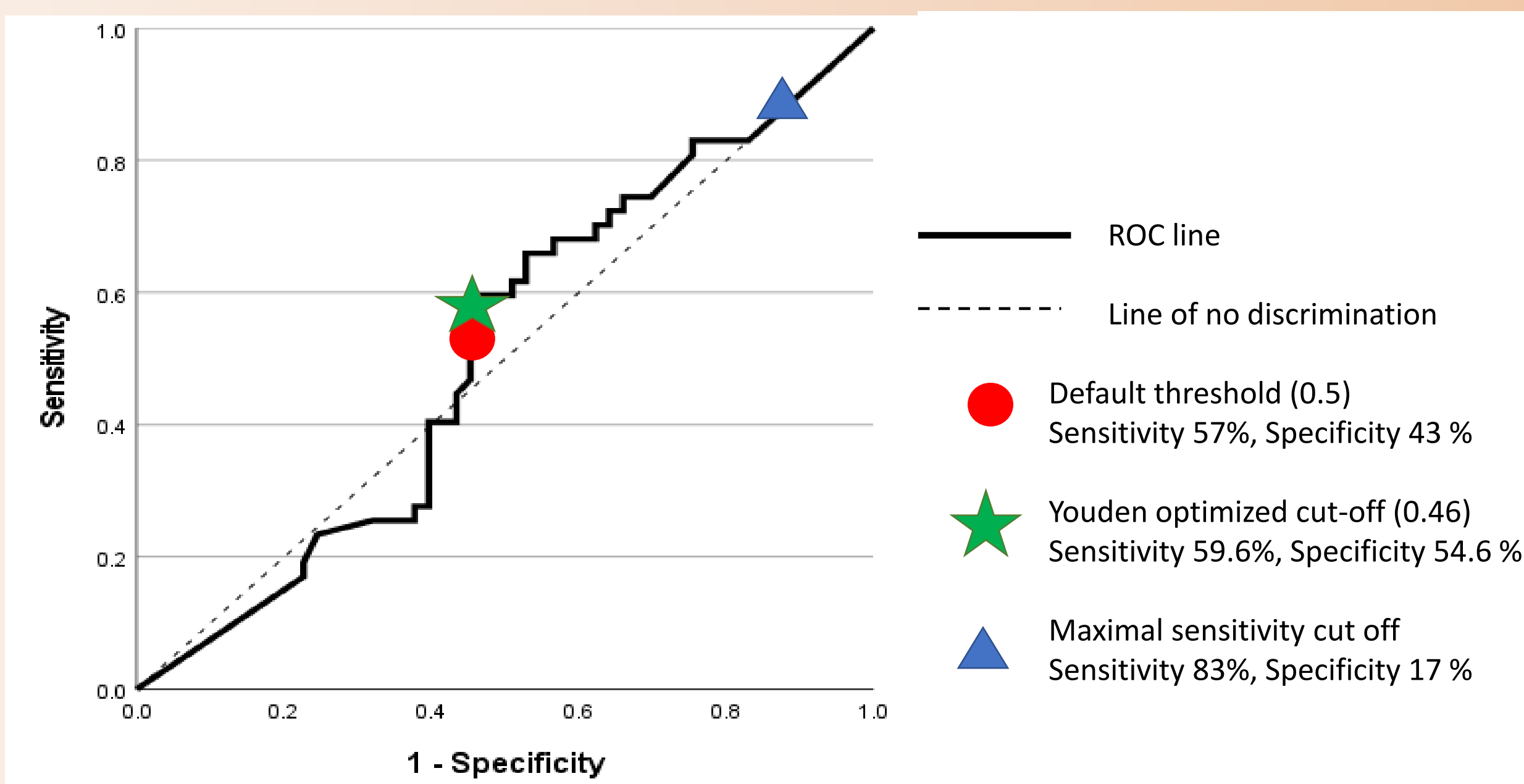
To compare the diagnostic accuracy of an AI algorithm with radiologists of varying experience levels in detecting acute ICH on non-contrast CT scans.

Materials and Methods

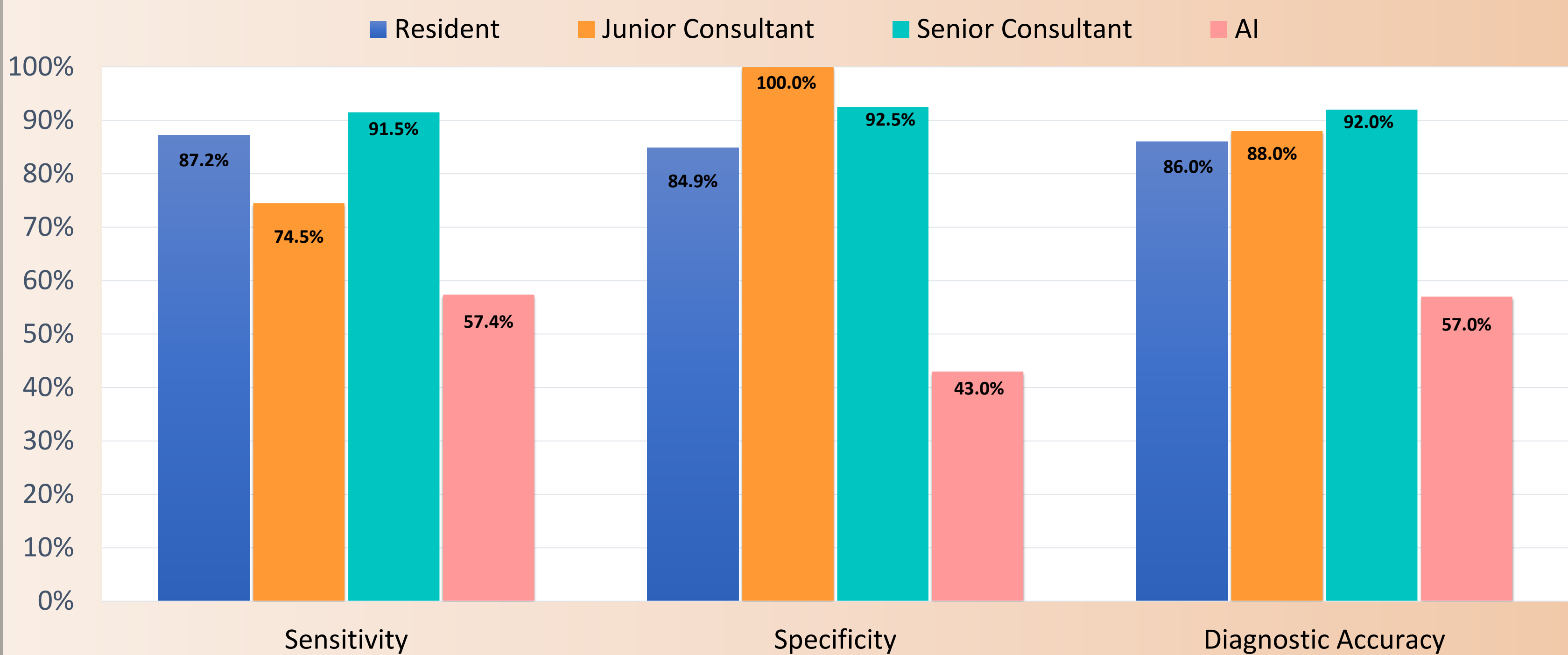
Following Ethical Review Board approval, this pilot retrospective study included 100 non-contrast CT brain scans (47 ICH-positive, 53 negative) from a tertiary trauma referral centre in Lahore, Pakistan. A radiology resident, junior consultant, and senior consultant independently reviewed anonymized scans for ICH presence and subtype, blinded to reports and each other's assessments. Ground truth was established by consensus of two senior consultant radiologists.

For AI analysis, anonymized DICOM images were processed using a deep learning pipeline in Google Colab based on a 2D EfficientNet-B3 convolutional neural network, pretrained on ImageNet and adapted from the RSNA Intracranial Hemorrhage Detection Challenge. Diagnostic performance was assessed using sensitivity, specificity, PPV, NPV, and inter-observer agreement, with AI evaluated at the default 0.5 threshold and optimized cut-offs using the Youden Index and a maximal-sensitivity approach.

Results

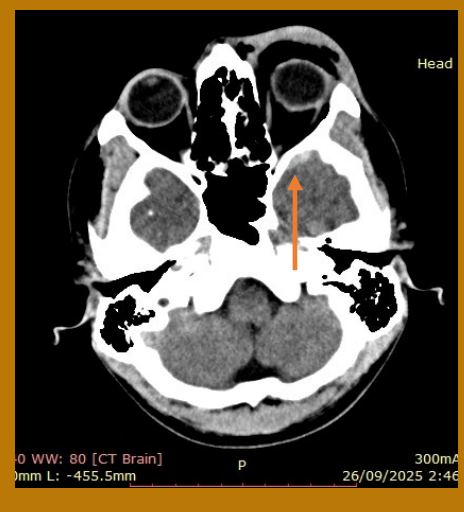
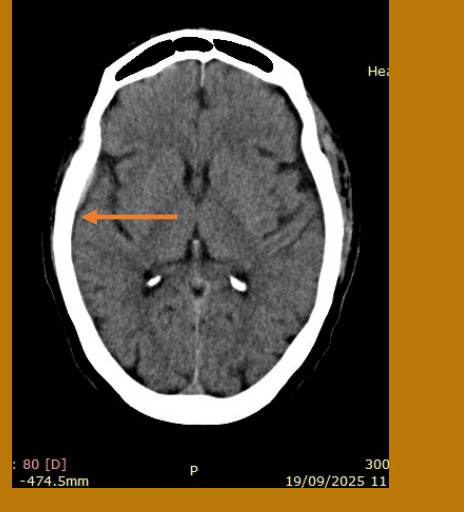
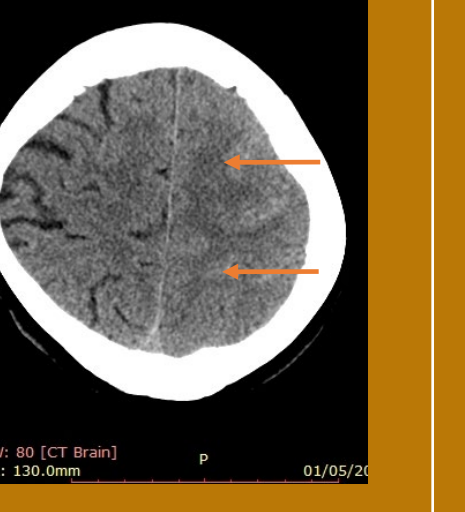
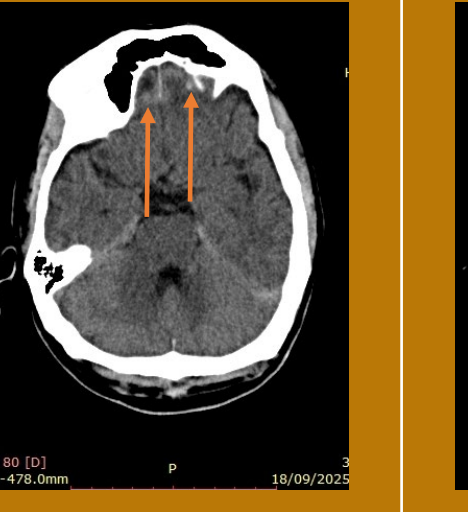
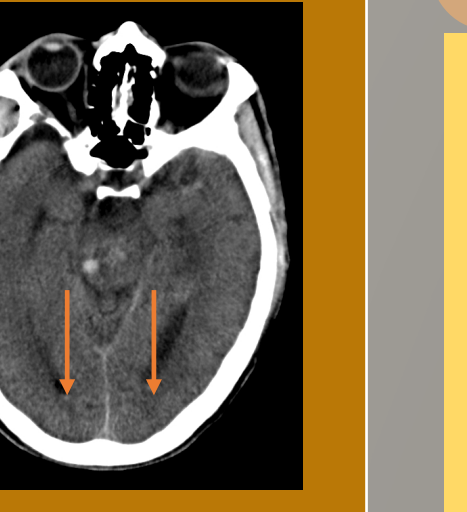


Receiver Operating Characteristic (ROC) Curve for AI Model in detecting ICH



Sensitivity, specificity & diagnostic accuracy of human readers and AI for detection of ICH

Table showing number of false negatives for each subtype of hemorrhage for each reporter

Type of Hemorrhage					
	Epidural (14)	Subdural (9)	Subarachnoid (17)	Intra Parenchymal (31)	Intra Ventricular (9)
Resident	5	6	2	12	1
Junior Consultant	7	7	10	16	5
Senior Consultant	1	3	3	4	1
AI	<u>10</u>	6	<u>11</u>	<u>21</u>	<u>6</u>

Key Results

- At default threshold of **0.5**, AI 57% sensitivity, 43% specificity, and 50% accuracy.
- Youden-optimized cut-off of **0.46** – sensitivity ↑ to 59.6% and specificity ↑ to 54.7%.
- Maximal-sensitivity strategy -- sensitivity ↑ to 83% but specificity ↓ to 17%.
- McNemar tests confirmed that all human readers significantly outperformed AI (**p < 0.001**)
- Comparisons among human readers revealed no significant differences (**p > 0.05**)
- AI had:
 - Highest **false-negative** rate (71.4%) for **epidural** hemorrhage
 - **Intraparenchymal** hemorrhage - **most misses** numerically (21)
 - **Overcalls** most common for **intraventricular** hemorrhage (26.4%)

Discussion

- Literature on AI performance in ICH detection is mixed : some studies report high diagnostic performance (5) with a comparative study concluding that AI outperforms radiologists (6) while others have found that radiologists are still better - including a prospective multicenter study (7); others report that residents outperform AI (8) even under time pressure (9)
- Consistent with our findings of low specificity of AI, even studies that report high sensitivity report lack of specificity in comparison to human readers (10)
- A meta-analyses identified epidural hemorrhage as particularly challenging to detect with AI, consistent with our highest false-negative rate (71.4%) in this category (11).
- Prior studies also demonstrate reduced detection of subdural and subarachnoid hemorrhages, supporting the variability in AI performance seen in our error patterns (12).
- Reported improvements in AI performance with larger hemorrhage size and higher acuity (13) and multiplicity (8) may explain missed subtle or small lesions in our cohort.
- The lower sensitivity of AI for ICH detection in our study, highlights the limitations of current algorithms as standalone diagnostic tools. Similar concerns in the literature emphasise the need for robust external validation and real-world clinical trials before routine implementation (14) and about generalizability of these tools with indeterminate failure modes (15).

Conclusion

This study demonstrates a clear generalizability gap in AI performance on local CT brain imaging data, with reduced accuracy compared to expected performance in published settings. Improved sensitivity after threshold adjustment was offset by poor specificity. Local validation and adaptation are essential before clinical use.

References



Acknowledgement

- Dr. Suneela Shaukat**, East Lancashire NHS Trust, UK
- Dr. Amna Ajmal**, Health Services Academy, Islamabad, Pakistan