Texas Society of Neuroradiology (TSNR)
Scientific Abstract
2026 Annual Meeting – Dallas, TX
February 21–22, 2026

Evaluation of Conventional Deep Learning and Slice-Aware Models For High Sensitivity Intracranial Hemorrhage Coverage

Souradeep Bhattacharya[1,2], Steve H Fung[2], Jason Handwerker[3]
[1]Texas A&M University, School of Engineering Medicine
[2]Houston Methodist Hospital
[3]UC San Diego

Purpose

Radiologists need reliable tools that can safely identify head CTs with no hemorrhage and reduce review workload without missing true bleeds. While receiver-operating-characteristic (ROC) analysis and area-under-the-curve (AUC) scores summarize overall accuracy, they do not answer the operational question: How many studies can an algorithm mark as low risk while maintaining near-perfect sensitivity?

We compared two neural network models for intracranial hemorrhage (ICH) detection: a conventional 2D model that analyzes each slice individually, and a sequential "slice-aware" model that also considers relationships between slices in a scan. We then tested whether these models could maintain 97%/99% sensitivity when applied to both internal and external datasets. "Coverage" was defined as the percentage of head CT exams that could be confidently identified as low risk at a given sensitivity threshold.

Materials and Methods

Models were trained on the RSNA 2019 ICH dataset [1] (Conventional: ResNet; Sequential: SE-ResNeXt + BiGRU). Probabilities were temperature-calibrated [2]. Thresholds guaranteeing 97% or 99% sensitivity were chosen on a held-out set using a split-conformal approach [3] and then applied to the internal RSNA test set and the external CQ500 cohort [4] without retuning. Heat maps were generated for visual review [5].

Results

On the internal RSNA test set, AUCs were 0.9648 (conventional) and 0.9727 (sequential). Although this AUC gap is small, the practical difference at a fixed sensitivity is larger:

- Sensitivity 99%: the sequential model marked 33.5% of exams as low-risk vs 26.8% for the conventional model, and produced fewer false positives among flagged studies (44.9% vs 55.8%).
- Sensitivity 97%: coverage rose to 52.9% (sequential) vs 50.0% (conventional), with lower false-positive rates (14.1% vs 18.8%).

External testing revealed distribution shift. With RSNA-tuned thresholds applied unchanged to CQ500:

- Sensitivity 99%: the conventional model's sensitivity fell to 63.4% at 46.9% coverage. The sequential model maintained 98.9% sensitivity at 6.6% coverage and generated significant false positives (86.3%).

- Sensitivity 97%: the sequential achieved 88.1% sensitivity at 23.0% coverage (false-positive 62.4%); the conventional model reached 57.5% sensitivity at 53.5% coverage (false-positive 32.2%).

Calibration effects: Threshold tuning improved the model's Expected Calibration Error (ECE), meaning the predicted probabilities better matched actual outcomes. However, this calibration did not improve external coverage. When thresholds were refit to preserve 97% and 99% sensitivity, coverage decreased even further. This finding suggests that calibration alone cannot overcome distribution shift between datasets. Most missed cases involved very small, low-contrast hemorrhages. Lesion size labels were unavailable, which we note as a limitation. Heat maps correctly highlighted hemorrhage regions when present.

Conclusion

Both conventional and sequential models achieved high AUC scores, but they resulted in poor coverage especially when applied to an external dataset. Calibration improved sensitivity but did not restore coverage when applied to the external dataset.

In practical terms, a high AUC does not guarantee coverage across institutions. Calibration and threshold tuning are not enough; local validation is strongly advised before deployment in a particular clinical workflow.

References

1. Flanders, A.E., et al., Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell, 2020. 2(3): p. e190211.
2. Chilamkurthy, S., et al., Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. The Lancet, 2018. 392(10162): p. 2388–2396.
3. Guo, C., et al., On Calibration of Modern Neural Networks, in Proceedings of the 34th International Conference on Machine Learning, P. Doina and T. Yee Whye, Editors. 2017, PMLR: Proceedings of Machine Learning Research. p. 1321––1330.
4. Tibshirani, R.J., et al., Conformal Prediction Under Covariate Shift, H. Wallach, et al., Editors. 2019.
5. Selvaraju, R.R., et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017. p. 618–626.

Texas Society of Neuroradiology (TSNR)
Scientific Abstract
2026 Annual Meeting – Dallas, TX
February 21–22, 2026

Figures

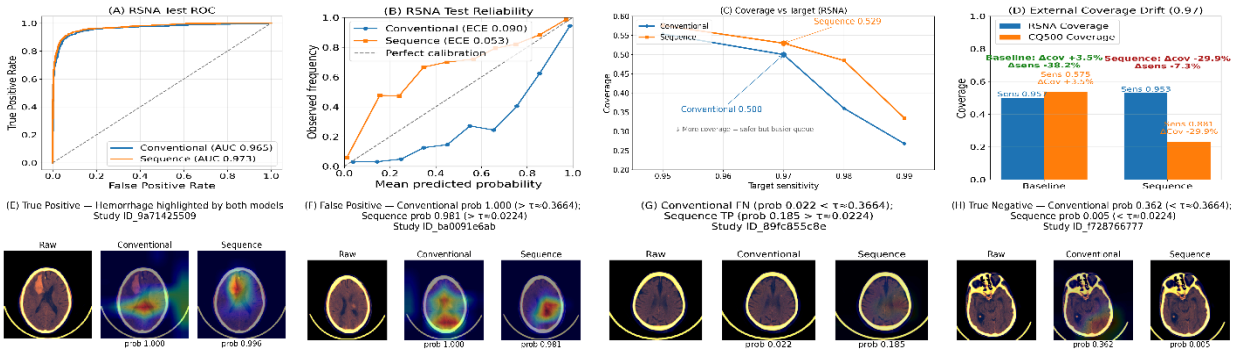

Figure 1: (A) ROC curve on internal test set for ICH detection. (B) Reliability diagram of predicted vs true ICH probability after calibration (ideal = diagonal). (C) Model coverage (fraction of exams not flagged) vs required sensitivity. (D) Impact of external shift on performance (CQ500 vs internal). (E) True-positive heatmap overlay highlighting a hemorrhage. (F) False-positive heatmap overlay highlighting a calcification misinterpreted as hemorrhage. (G) False-negative case (missed small hemorrhage with minimal activation). (H) True-negative exam with no spurious activation.