



Real-World Performance Evaluation of a Commercial Deep Learning Model for Intracranial Hemorrhage Detection: Anatomical and Demographic Analysis

Mohammadreza Chavoshi, MD, Emory University School of Medicine; Aawez Mansuri, MS; Judy Gichoya, MD, MS, FSIIM; Theo Dapamede, MD; Dan I.G. Cohen-Addad, MD; Ninad V. Salastekar, MD; Hanzhou Li, MD; Wasif Bala, MD; John T. Moon, MD; Bardia Khosravi, MD, MPH, MHPE; Frank Li, PhD; Chad Robichaux, MPH; Beatrice Brown-Mulry, MS; Rohan Satya Isaac, MS; Hari Trivedi, MD

Introduction/Background

To externally evaluate the performance of a commercial deep learning model for intracranial hemorrhage (ICH) detection, with a focus on demographic subgroups and ICH subtype/location analysis.

Methods/Intervention

We retrospectively analyzed 52,719 non-contrast head CT reports processed by a commercial ICH detection model over one year. Three residents manually annotated 500 exams for ICH presence and 36 subgroup labels (type, location, acuity, size), with 100 cases adjudicated by board-certified radiologists. Using prompt-engineered GPT-4o (optimized on 400 reports, tested on 100 attending-labeled cases), we extracted labels from the remaining reports. Model predictions were compared to extracted labels.

Results/Outcome

Optimized GPT-4o achieved overall high accuracy (Cohen's Kappa: 0.92; sensitivity: 0.96; specificity: 0.97) for extracting ICH and subtypes from reports. After extracting labels, we excluded uncertain cases (as defined by the radiologist) and post-operative ICHs, yielding 48,208 cases. The AI model showed 81.5% sensitivity and 97.4% specificity across 48,208 exams (3,136 ICH cases). Sensitivity was lower for chronic (66.8%) and subacute (69.2%) hemorrhages compared to acute ones (86.8%). Single-compartment ICHs (77.8%) were less accurately detected than multi-compartment hemorrhages (93.6%), and small hemorrhages (75.2%) were less well identified than large ones (>10mm) (95.9%). Among intra-axial bleeds, occipital lobe intraparenchymal hemorrhages had the lowest sensitivity (78%), while intraventricular bleeds were detected with higher sensitivity (96.0%), especially when involving multiple ventricles (100%). For extra-axial bleeds, isolated convexity hemorrhages were less well detected (77%) than those involving both the convexity and tentorium/falx (93%). Convexity-limited subarachnoid hemorrhages had lower sensitivity (86.8%) compared to diffuse ones (94.6%).

Conclusion

The model demonstrated high specificity but moderate sensitivity for ICH detection in real-world settings. Performance varied by hemorrhage type, size, location, and patient demographics, with lower sensitivity for small, chronic, and single-compartment bleeds. Notably, model performance was worse than the FDA clearance metrics of sensitivity 93.6% and specificity 92.3%.

Statement of Impact

This large-scale evaluation highlights performance gaps in commercial AI models that may impact patient care. Lower detection in subtle presentations underscores the need for cautious clinical integration and vigilant

radiologist oversight.

Category	Sensitivity	Specificity	Count
All Patients	81.54%	97.44%	48208
Demographic			
Patient Class			
Emergency	80.35%	98.09%	35388
Inpatient	82.96%	95.28%	9353
Outpatient	74.90%	95.44%	3375
Race			
Asian	83.66%	96.09%	1890
Black	78.71%	97.52%	27102
Other	83.50%	98.18%	1422
Unknown	85.08%	96.98%	2069
White	83.92%	97.46%	15703
Gender			
Female	82.22%	97.82%	27162
Male	80.97%	96.94%	21039
Age Group			
30-44	77.82%	98.32%	7410
45-59	81.65%	97.41%	9440
60-74	82.47%	96.75%	13650
75+	81.96%	97.32%	11627
<30	74.68%	98.80%	3653
Acuity			
Acute	86.81%		1675
Chronic	66.77%		334
Subacute	69.18%		279
Acute on Chronic	90.16%		122
Anatomical Location			
Extra-axial			
Epidural/Subdural	82.91%		1603
Subarachnoid	88.58%		963
Intra-axial			
Intraventricular	96.03%		479
Intraparenchymal	87.06%		1453

Table1. ICH model performance across demographic and clinical subgroups.

Keywords

Intracranial Hemorrhage; Diagnostic Performance; Deep learning; Large Language Models; Computed Tomography