AMA
AMERICAN MEDICAL
ASSOCIATION

Abstract Title

Evaluating GPT-4o Performance for Hemorrhage Detection and Subtype Classification in Head CT Scans

Background

Accurate detection and classification of intracranial hemorrhages (ICH) from computed tomography (CT) images are critical for timely clinical intervention. Large language models (LLMs) like GPT have shown potential in medical imaging tasks, yet their effectiveness in complex classification scenarios remains unclear. This study assesses GPT-4o's capability to identify hemorrhage presence and categorize hemorrhage subtypes using publicly available CT imaging data.

Methods

CT scans from the PhysioNet Intracranial Hemorrhage CT Dataset were preprocessed into axial slices, normalized to 8-bit grayscale, and combined into composite images to comply with GPT-4o’s token constraints. The evaluation consisted of two tasks:

1. Binary classification (hemorrhage presence vs. absence) involving 75 scans (36 positive, 39 negative).
2. Multi-class classification of hemorrhage subtypes (intraventricular, intraparenchymal, subarachnoid, epidural, subdural) among the 36 positive cases.

Model outputs were assessed using accuracy, precision, recall, F1-score, exact match accuracy, and Hamming scores.

Results

In the binary classification task, GPT-4o achieved an overall accuracy of 0.52, with precision of 0.50, recall of 0.14, and F1-score of 0.22 for detecting hemorrhages. Non-hemorrhagic cases were identified with higher recall (0.87) but similar precision (0.52), yielding an F1-score of 0.65.

For subtype classification, precision was highest for epidural hemorrhages (1.0) but accompanied by low recall (0.14) and an F1-score of 0.25. Intraparenchymal hemorrhage showed balanced precision (0.46) and recall (0.75), achieving an F1-score of 0.57. Exact match accuracy was low (0.06), indicating GPT-4o rarely identified all subtypes correctly in a single prediction. The average Hamming score was moderate at 0.54, highlighting partial accuracy across subtypes.

Conclusion

GPT-4o demonstrated modest performance in hemorrhage detection, particularly limited by low recall in positive cases. While specificity was comparatively better, the overall diagnostic capability remained suboptimal. Subtype classification posed additional challenges, reflected by poor exact match accuracy despite acceptable partial label accuracy. These findings suggest the necessity of refined training approaches, targeted model tuning, and potential integration with hybrid analytical methods to improve clinical reliability.