RADAR (Real-time AI Data Assessment and Reporting): A Solution for Automated Monitoring of Commercial AI Model Performance Built with “Off-the-Shelf” Components

Adam Flanders, MD, CIIP, FSIIM, Vice-Chair Imaging Informatics, Radiology, Thomas Jefferson University
Paras Lakhani, MD; Prahlad Menon, PhD; Robyn Ball, PhD; Luciano Prevedello, MD; George Shih, MD; Avi Sharma, MD; Ryan Lee, MD

Background/Problem Being Solved

Most commercial AI customers have no independent ability to measure model performance or drift and must rely upon the vendor for this critical task.. Presented herein is a prototype of a semi-autonomous application that continuously measures model performance for a triage brain CT hemorrhage model built with “off-the-shelf” components.

Intervention(s)

The tool consists of four components: (1) an AI result receiver, (2) a result database, (3) a report interpreter and (4) a statistical engine. All results processed by CT hemorrhage inference engine were sent simultaneously to a MIRTH HL7 receiver for processing. The database was updated with the final radiology report when it was made available. The report impression was parsed and processed by an ensemble of five LLMs(llama3.2:1b, llama3.2:3b, codellama:7b, llama3.1:8b, granite3-dense:2b) running in the Ollama framework. A “consensus” was reached when three or more LLMs agreed. Using the consensus as reference, a confusion matrix was created to generate AI model performance metrics. Fleiss’ and Cohen’s kappas were calculated to check agreement between the LLMs. Throughout, the administrator interacted with a web dashboard that provided the updated performance of the model and provided a means to inspect discordant results.

Barriers/Challenges

Automating report review requires a consensus of an ensemble of LLMs and an iterative approach using a combination of human review and prompt engineering as a means to minimize human evaluation. The challenge was to find a balance where only periodic human review of the automated report validation was necessary.

Outcome

The database monitored results from over ~16,000 BRAIN CT exams derived from eighteen hospitals and 35 scanners collected from nine months of continuous use. Due to heterogeneous inter-model agreement an ensemble of LLMs was chosen as consensus to confer more consistent results. An iterative process removed of low performing LLMs to boost performance. Finally, the chosen consensus was scored against expert evaluation of a subset of reports. Fleiss kappa for these LLMs: 0.73 and Cohen’s kappa ranged from 0.14 to 0.78.

Conclusion/Statement of Impact/Lessons Learned

An ensemble of LLMs was employed as a first pass to verify radiology report imaging findings and can be used to automate independent quality control of a triage AI application in the clinical setting.

Figure(s)

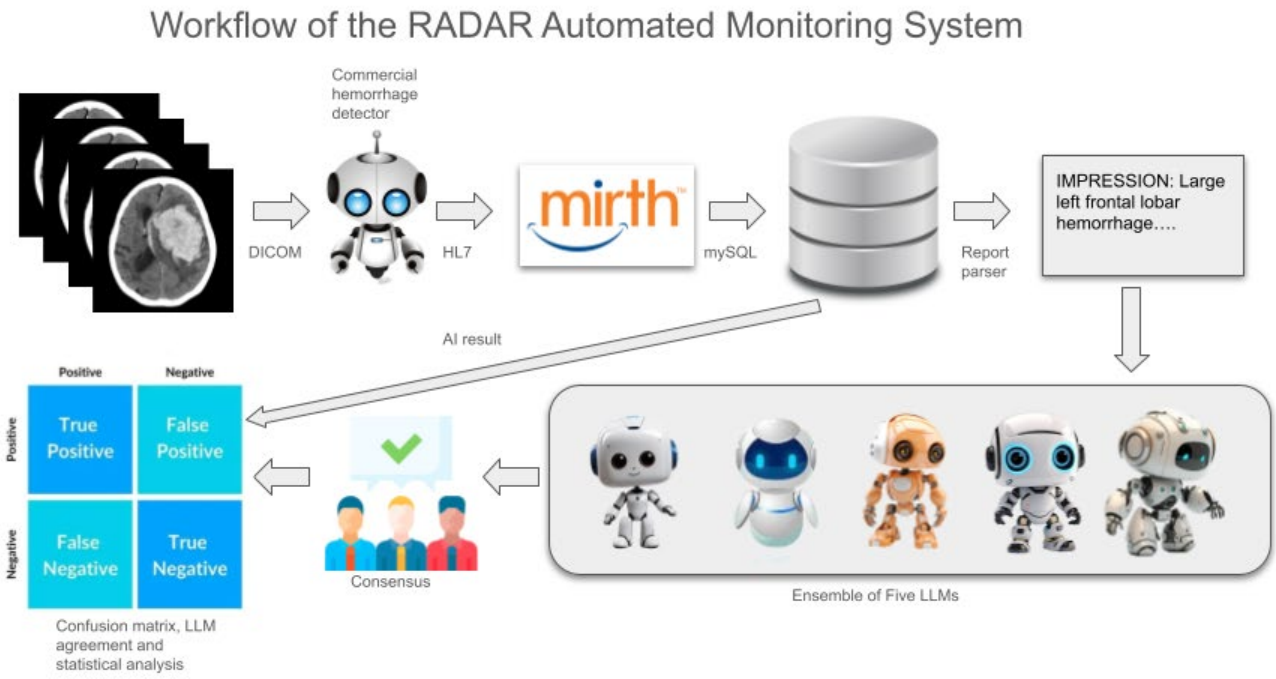

Figure 1. Workflow of the RADAR Automated Monitoring System

RADAR AI Metrics Dashboard

Today's Date: Wednesday, 11-Dec-2024 10:13:28 Eastern Standard Time

This page provides insights into the performance of the AI hemorrhage model against the consensus of five LLMs (>3) assessing the impression section from reports used as ground truth.

LLM models used: "llama3.2:1b" : 1, "llama3.2:3b" : 2, "codellama:latest" : 3, "llama3.1:8b" : 4, "granite3-dense:latest" : 5

Performance Metrics

Performance

Processing Date: 2024-12-11 10:12:49.762016

Metric	Value
Accuracy	0.9432
Precision (PPV)	0.7436
Recall (Sensitivity)	0.4624
Specificity	0.9859
F1-Score	0.5702
Negative Predictive Value (NPV)	0.9539

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	609	708
Actual Negative	210	14645

Kappa Metrics

Fleiss' Kappa: 0.3241

Pairwise Cohen's Kappa:

- nlpheme1 vs nlpheme2: 0.1087
- nlpheme1 vs nlpheme3: 0.0844
- nlpheme1 vs nlpheme4: 0.1235
- nlpheme1 vs nlpheme5: 0.1092
- nlpheme2 vs nlpheme3: 0.7021
- nlpheme2 vs nlpheme4: 0.7671
- nlpheme2 vs nlpheme5: 0.7433
- nlpheme3 vs nlpheme4: 0.6071
- nlpheme3 vs nlpheme5: 0.5931
- nlpheme4 vs nlpheme5: 0.7875

Figure 2.

Keywords

Administration & Operations; Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Quality Improvement & Quality Assurance