SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Withholding Sensitive Radiology Reports: Real-World Evaluation of an LLM-Based Classifier

Aaron T. Chin, MD, University of California, Los Angeles; Kelly Bartlett; Eric M. Cheng, MD, MS; Mojtaba Heidarysafa, PhD; Kamran Kowsari, PhD; Paul J. Lukac, MD, MBA, MS; James W. Sayre, PhD; Lucy Chow, MD

Introduction/Background

According to California law, imaging results that suggest a new or recurrent malignancy should not be immediately released to the portal. Our health system radiologists manually tag imaging results, but this is difficult to do consistently. We developed an LLM-based model that could supplement the actions taken by radiologists.

Methods/Intervention

The model combines a rule-based system (detecting suspicious malignancy Reporting and Data System standardized classifications or a predefined key phrase) with a deep neural network that scores radiology impressions from 0.001-1; those ≥0.5 are withheld from immediate release (Figure 1). Three-hundred reports were randomly sampled and reviewed by clinical informatics experts (AC, KB, LC, EC, PL): 100 with the minimum model score of 0.001 and 200 with scores >0.001. Performance metrics were calculated against consensus review.

Results/Outcome

Of 300 sampled reports, 298 reached consensus review, with 64 (21.5%) determined to be withheld. The model flagged 65 reports (21.8%) and radiologists flagged 39 reports (13.0%). Agreement between the model, radiologist, and reviewers was measured by intraclass correlation coefficient (ICC), and was highest between the model and reviewers (ICC 0.791), followed by radiologists and reviewers (0.738), and AI and radiologists (0.536). In the >0.001 score tier (n=198), model performance included sensitivity 75%, specificity 87%, and ROC AUC 0.88 (Table 1). In the lowest-scored 0.001 tier group, the model appropriately released 99 of 100 cases, with one report flagged by reviewers that the model did not capture. Among the 17 false negatives, the median model score was 0.08 with an interquartile range of 0.04–0.36 (Figure 2).

Conclusion

Model performance was robust, with the ROC AUC likely underestimated due to the exclusion of the lowest-score reports. False negatives are the most critical errors, as it represents risk for noncompliance for the health system. These errors occurred across a wide range of scores without a discernible pattern, prompting an ongoing qualitative review to identify potential biases in the model. We also plan to analyze cases of model-radiologist disagreement, using these cases both as feedback to improve radiologist consistency and to guide model refinement.

Statement of Impact

An LLM-based model can support consistent delayed release of malignancy-related impressions.

	Metric Value (95% CI)
Reviewer Withhold Count	63 / 198
Sensitivity	0.75 (0.62, 0.84)
Specificity	0.87 (0.79, 0.92)
Positive Predictive Value	0.72 (0.59, 0.82)
Negative Predictive Value	0.880 (0.81, 0.93)
Accuracy	0.82
F1 Score	0.73
ROC AUC	0.88 (0.82, 0.93)

Table 1. Performance of Model in >0.001 Score Tier (n = 198). CI: Confidence Interval

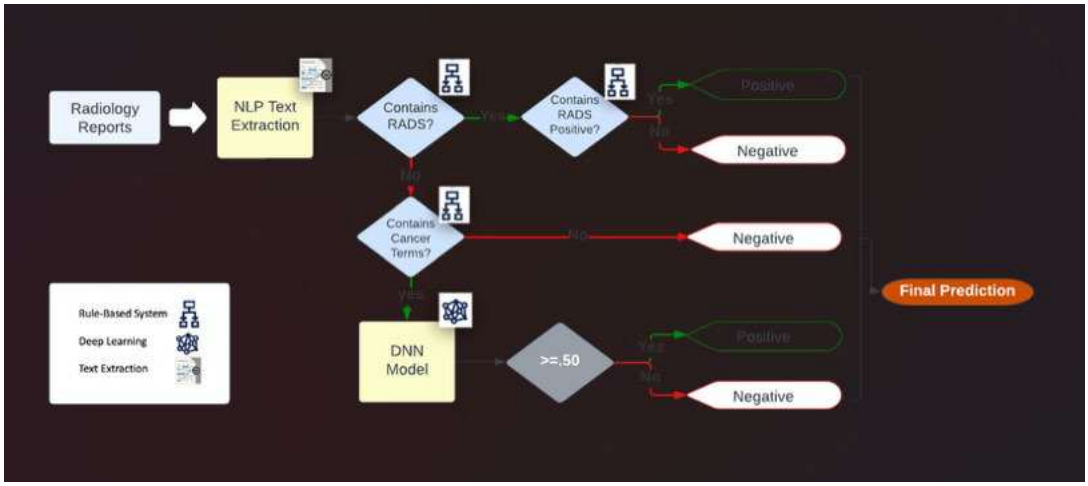

Figure 1. Model architecture combining rule-based and deep neural network model

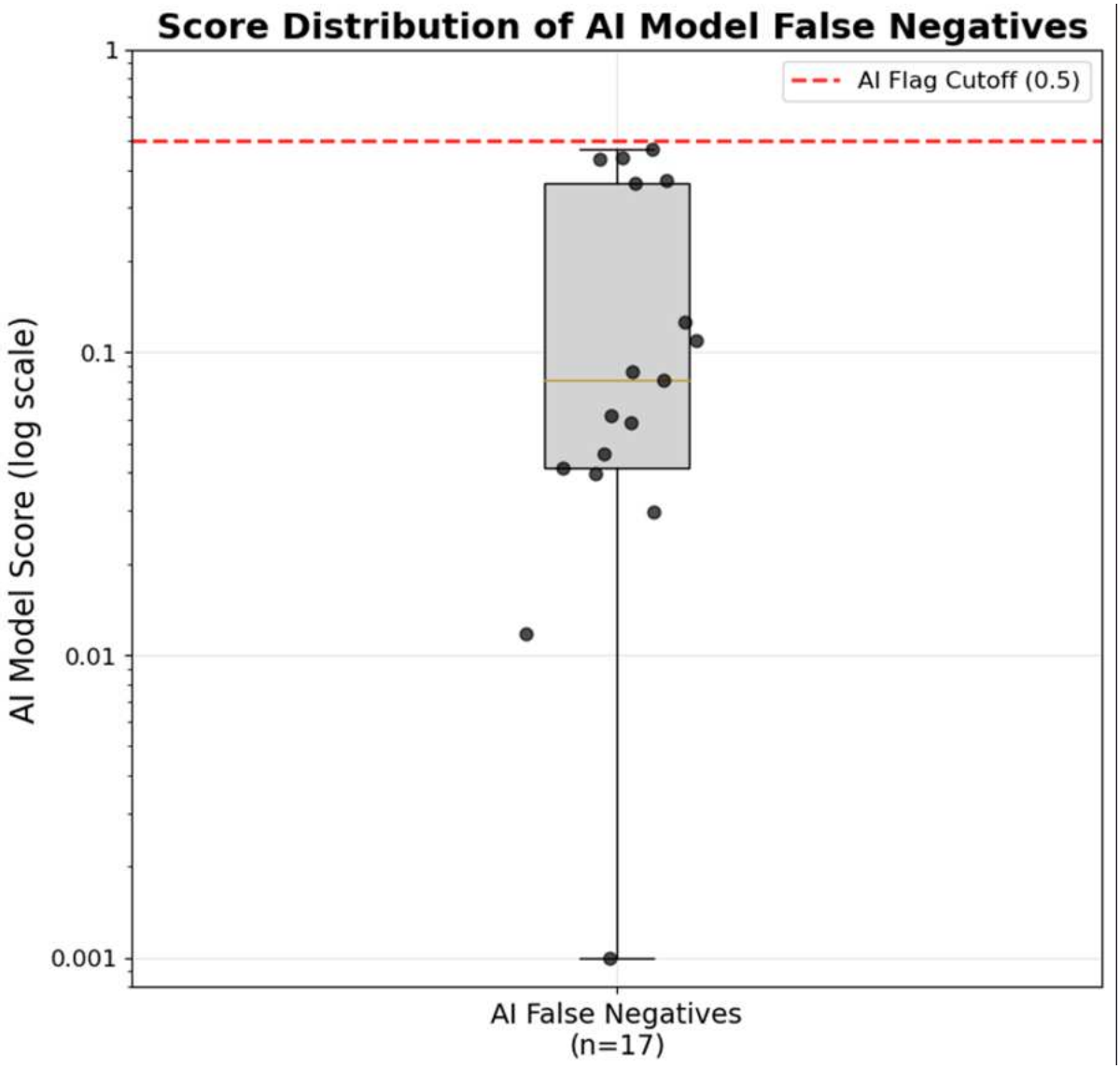

Figure 2. Score distribution of model false negatives

Keywords

Large Language Models; Natural Language Processing; Clinical Decision Support; Release of Information; Regulatory Compliance