Examining the Usability of a Machine Learning Enhanced Patient Safety Event Reporting System

Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

Deenar Virani, Victoria Yeung, Dr. Myrtede Alfred
Department of Mechanical and Industrial Engineering, University of Toronto

SEDlab
SAFETY, EQUITY, & DESIGN

Background

- Patient safety events (PSEs) describe **instances of avoidable harm** in healthcare [1]
- 1 in 17 hospital stays results in a harmful event [2]
- Most hospitals have implemented PSE reporting [2]
- **Issues**: 50-96% of PSEs go **underreported** [3], **misclassification** & errors are common [4,5,6], **barriers** due to time constraints & usability [7,8]
- Results in **delays, burdens** of reclassification, and hindered learning [9]
- **Machine learning (ML)** can be used to automate classification of event types, & has been successful with increased accuracy [10,11,12]
- **Human-AI collaboration (HAIC)**, which describes humans and AI complementing each other to enhance decision making, can enhance PSE reporting [10,13]
- **Explainability** techniques can be integrated to build trust & transparency in ML [14]

Goal: Evaluate the usability of a PSE interface with an integrated ML classifier for event types & LIME explainability

Methods

System Development

- Used 861 obstetric PSE reports (2019-2020) [10]
- SVM Roberta-base model (75.4%) accuracy [10]
- Integrated LIME explainability [14] to show highlighted words influencing ML classification
- Interface with 4 sections, built using Gradio [15]

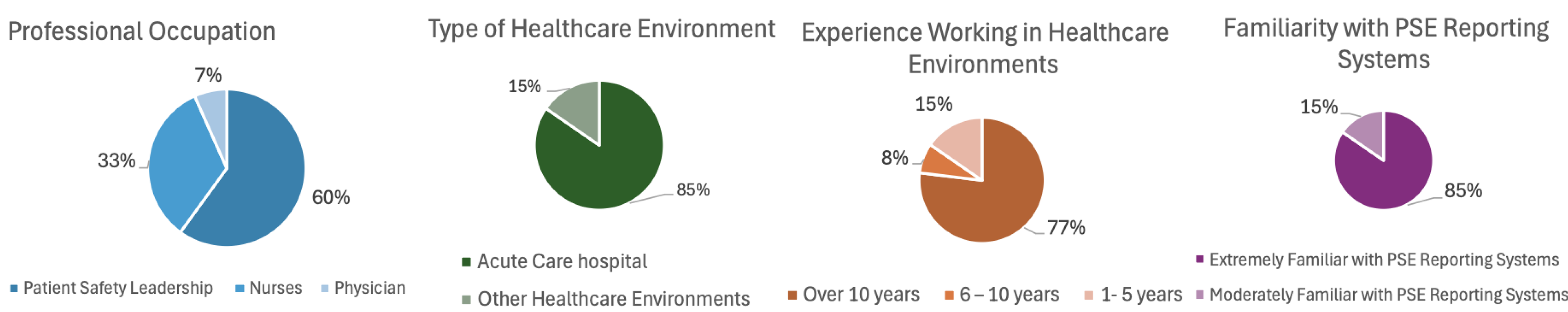
Usability Testing (3 parts, 2 scenarios each)

- Part 1 (P1): Full PSE report with ML
- Part 2 (P2): Classification with ML (50% reliability)
- Part 3 (P3) : Manual classification

Measures

- Reporting & classification time (min), classification accuracy (%), agreement with predictions & recommendations selected (%), SUS scores [16], qualitative debriefing interview feedback

Results & Key Points



Measure	With ML (P1, P2)	Without ML (P3)
Completion Time (min)	5.66 min (SD = 1.80)	N/A
Classification Time (min)	1.96 min (SD = 0.90)	1.19 min (SD = 0.59)
Classification Accuracy (%)	P1: 92.3%, 92.3% P2: 23.1%, 46.2%	31%, 62%

- **Mean SUS Score**: 87.9
- **High confidence**: mean 4.54/5 on SUS question about confidence
- **Agreement with ML predictions**: 76.9–100%
- **Explainability**: relevant in 7.7–38.5% of cases in P1, 0% in P2
- **Speed-accuracy tradeoff**: increased classification time was manageable, support of ML predictions worth minor time cost [17]
- **Calibration of trust issue**: some users agreed with incorrect predictions, however others responded to the shift in reliability accordingly
- **Explainability gaps**: LIME is often unstable [18]
- **HAIC improved classification accuracy and user confidence**

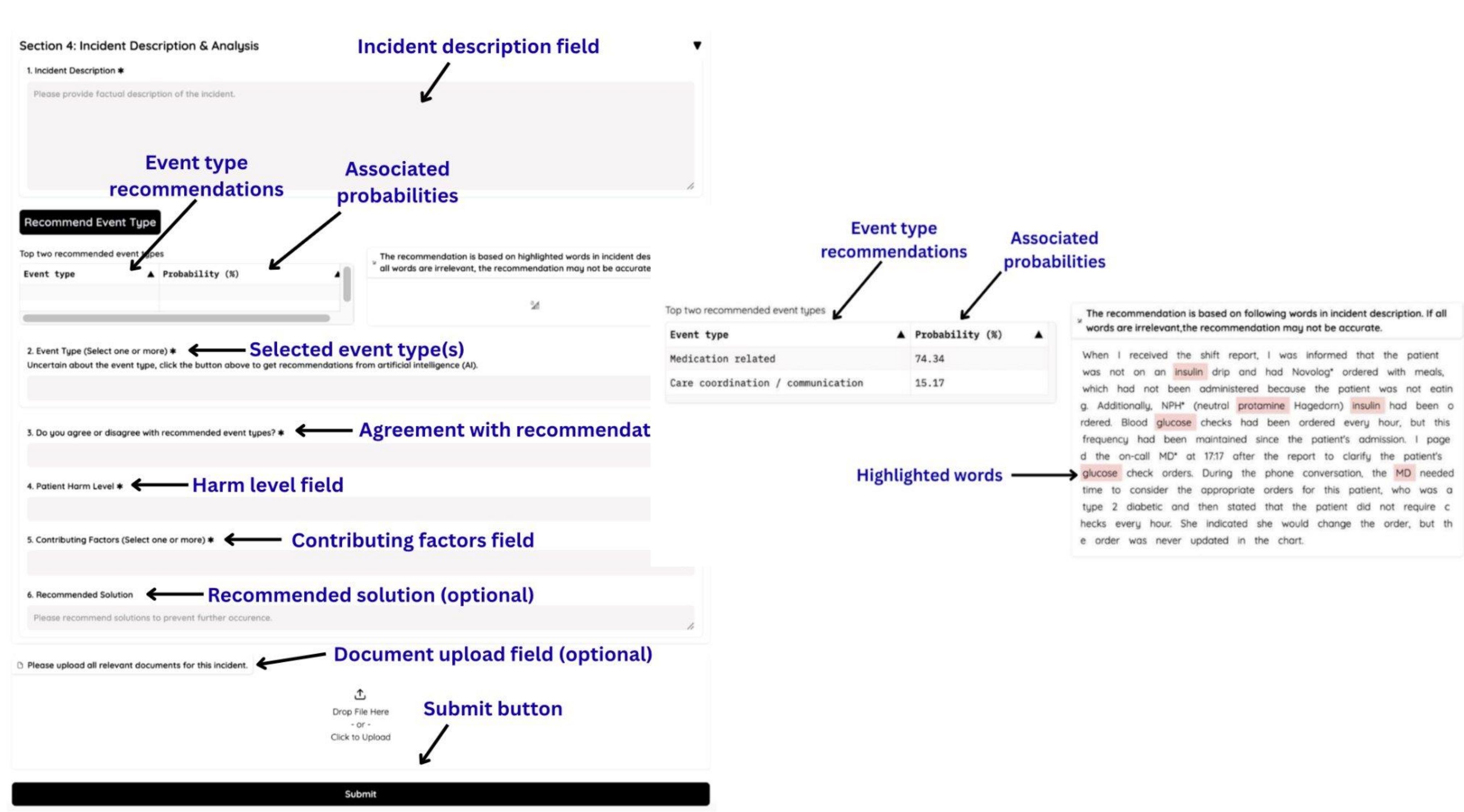


Figure 1. Event description & analysis section of PSE interface (Section 4)

Limitations & Next Steps

- Trained with a small dataset
- LIME feature not robust
- Small participant group
- Some scenarios possibly easier to classify

Future directions:

- Expand ML integration other PSE reporting categories & text generation
- Test alternative explainability methods
- Conduct additional usability testing

Conclusion

- Integration of ML into PSE reporting systems is a relatively new area of research that has potential to optimize & streamline the reporting process through HAIC
- This study demonstrates the integration of an ML classifier in PSE reporting systems shows potential to mitigate challenges related to the completion & quality of reporting

References

[1] Skelly CL, Cassagnol M, Munakomi S. Adverse Events. [Updated 2023 Aug 13]. In: *StatPearls* [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK558963/
[2]Canadian Institute for Health Information. Patient harm in Canadian hospitals? It does happen.. Available: https://www.cihi.ca/en/patient-harm-in-canadian-hospitals-it-does-happen
[3] Gong, Y., Song, HY., Wu, X. et al. Identifying barriers and benefits of patient safety event reporting toward user-centered design. *Saf Health* 1, 7 (2015). https://doi.org/10.1186/2056-5917-1-7
[4] Brubacher, J. R., Hunte, G. S., Hamilton, L., & Taylor, A. (2011). Barriers to and incentives for safety event reporting in emergency departments. *Healthcare quarterly (Toronto, Ont.)*, 14(3), 57–65. https://doi.org/10.12927/hcq.2011.22491
[5] Chen, H., Cohen, E., Wilson, D., & Alfted, M. (2023). Improving Patient Safety Event Report Classification with Machine Learning and Contextual Text Representation. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual meeting*, 67(1), 1063–1069. https://doi.org/10.1177/21695067231193645
[6] Puthumana, J. S., Fong, A., Blumenthal, J., & Ratwani, R. M. (2021). Making Patient Safety Event Data Actionable: Understanding Patient Safety Analyst Needs. *Journal of patient safety*, 17(6), e509–e514. https://doi.org/10.1097/PTS.0000000000000400
[7] Kousgaard, M. B., Joensen, A. S., & Thorsen, T. (2012). Reasons for not reporting patient safety incidents in general practice: a qualitative study. *Scandinavian journal of primary health care*, 30(4), 199–205. https://doi.org/10.3109/02813432.2012.732469
[8] Wiele, Paul, "Healthcare Incident Reporting: The Impacts of Usability of Input Interfaces, Usability of Resulting Data, and Attitudes Towards Reporting" (2016). Thesis. Rochester Institute of Technology. Accessed from https://repository.rit.edu/cgi/viewcontent.cgi?article=10174&context=theses
[9] Gong, Y. Data Consistency in a Voluntary Medical Incident Reporting System. *J Med Syst* 2011 Aug 1;35(4):609–615. doi: 10.1007/s10916-009-9398-y
[10] Chen H, Cohen E, Wilson D, Alfred M A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study JMIR Hum Factors 2024;11:e53378 doi: 10.2196/53378PMID: 38271086PMCID: 10853856
[11]Fong, A., Behzad, S., Pruitt, Z., & Ratwani, R. M. (2021). A machine learning approach to reclassifying miscellaneous patient safety event reports. *Journal of Patient Safety*, *17*(8), e829-e833. https://pubmed.ncbi.nlm.nih.gov/32555052/
[12] Evans, H. P., Anastasiou, A., Edwards, A., Hibbert, P., Makeham, M., Luz, S., Sheikh, A., Donaldson, L., & Carson-Stevens, A. (2020). Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. *Health informatics journal*, *26*(4), 3123–3139. https://doi.org/10.1177/1460458219833102
[13]Fragiadakis G., Diou C., Kousiouris G., Nikolaidou M. 2025. “Evaluating Human-AI Collaboration: A Review and Methodological Framework”. https://doi.org/10.48550/arXiv.2407.19098
[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[15] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, Gradio: Hassle-free sharing and testing of ML models in the wild. (Jun. 2019). Python. doi: 10.48550/arXiv.1906.02569.
[16] J. Brooke, “SUS: A quick and dirty usability scale,” Usability Eval. Ind., vol. 189, Nov. 1995.
[17] Heitz R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. Frontiers in neuroscience, 8, 150. https://doi.org/10.3389/fnins.2014.00150
[18] Alvarez-Melis, David, and Tommi S. Jaakkola. 2018. “On the Robustness of Interpretability Methods.” arXiv. https://doi.org/10.48550/arXiv.1806.08049.