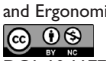
Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2024, Vol. 68(1) 465–472


DOI: 10.1177/10711813241260676
journals.sagepub.com/home/pro


Evaluating Active Learning Strategies for Automated Classification of Patient Safety Event Reports in Hospitals

Shehnaz Islam[1], Myrtede Alfred[1], Dulaney Wilson[2], and Eldan Cohen[1]

Abstract

Patient safety event (PSE) reports, which document incidents that compromise patient safety, are fundamental for improving healthcare quality. Accurate classification of these reports is crucial for analyzing trends, guiding interventions, and supporting organizational learning. However, this process is labor-intensive due to the high volume and complex taxonomy of reports. Previous work has shown that machine learning (ML) can automate PSE report classification; however, its success depends on large manually-labeled datasets. This study leverages Active Learning (AL) strategies with human expertise to streamline PSE-report labeling. We utilize pool-based AL sampling to selectively query reports for human annotation, developing a robust dataset for training ML classifiers. Our experiments demonstrate that AL significantly outperforms random sampling in accuracy across various text representations, reducing the need for labeled samples by 24% to 69%. Based on these findings, we suggest that incorporating AL strategies into PSE-report labeling can effectively reduce manual workload while maintaining high classification accuracy.

Keywords

artificial intelligence, machine learning, automation, healthcare, human factors, natural language processing, classification, incident reporting, patient safety, active learning

Introduction

In healthcare organizations, patient safety events (PSE), referring to incidents that compromise patient safety, such as medical errors, near misses, and adverse events pose a significant threat and can cause avoidable harm, injuries and even death (Wolf & Hughes, 2008). In 2000, the study "To err is human" estimated that clinical errors cause about 44,000 to 98,000 deaths annually in the USA, with later studies raising this estimate to 251,454 deaths, making medical errors the third-leading cause of death in the USA (Donaldson et al., 2000; Makary & Daniel, 2016). The global cost of medication errors alone is estimated at US$ 42 billion annually (World Health Organization, 2021). Following the Global Patient Action Plan by the WHO in 2020 (World Health Organization, 2021), healthcare organizations have widely adopted incident reporting systems to document PSE reports. PSE reporting systems enable healthcare personnel to voluntarily log incidents, including medical errors and unsafe conditions, in the form of structured (event type, harm level, date, and location) and unstructured data (textual event descriptions and outcomes; Albolino et al., 2010; Ong et al., 2010). These reports are then reviewed by hospital staff such as risk managers and patient safety analysts, playing a crucial role in enhancing healthcare quality by providing insights into factors leading to PSEs and facilitating the development of preventive measures (Herzer et al, 2012).

Accurate classification of PSE reports into corresponding event types is crucial for analyzing trends, guiding interventions, and enhancing organizational learning (Brubacher et al., 2011; Fong et al., 2015). However, the high volume of PSEs reported (Koike et al., 2022) and the complexity of the classification taxonomy, which require specialized knowledge, make the classification process labour-intensive, time-consuming, and expensive (Dimitrakakis et al., 2008; Liang et al., 2017). Several studies have demonstrated the efficacy of supervised Machine Learning (ML) models in automating PSE report classification (Chen et al., 2024; Evans et al., 2020; Wang et al., 2017). Improving the reliability of report

[1]Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada
[2]Public Health Sciences, Medical University of South Carolina, Charleston, USA

Corresponding Author:
Shehnaz Islam, Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON, Canada. M5S 1A1.
Email: shehnaz.islam@mail.utoronto.ca

classification can enhance safety. However, supervised ML models require high-quality and voluminous manually-labeled datasets for optimal performance. Manually labeling data is time-consuming, costly, and often impractical due to the large volume of data generated (Goudjil et al., 2018). Active Learning (AL) is a semi-supervised learning technique that involves selectively querying the most informative data points for manual labeling, thereby optimizing the use of human resources (Settles, 2009). Our study aims to utilize human labeling resources efficiently to develop a robust dataset for training ML models, seeking to balance labeling costs and classification accuracy. This study investigates the efficacy of AL strategies in improving the performance of ML models in event type classification of PSE reports. Thus, our goal is to enhance the PSE labeling process by integrating Artificial Intelligence with human expertise. This study uses only free-text incident descriptions, assessing four AL strategies and a random selection baseline across both static text representations and deep learning-based contextual text representations. We apply these strategies to Support Vector Machine (SVM) and Logistic Regression (LR) models, assessing classification accuracy improvements through iterative data sampling. Our goal is to determine if AL enhances performance with fewer labeled samples than random sampling, comparing improvements across both static and contextual text representations.

Methods

Dataset

This study utilized 861 PSE reports from a Southeastern U.S. academic hospital's maternal-care units (Jan 2019–Dec 2020). The study was approved by the hospital's institutional review board (Pro00105892). Among the 25 distinct event types present in the dataset, the ML classifiers were exclusively trained on PSE-types with at least 40 samples—constituting about 73% of all reports—to ensure sufficient learning data (see Table 1). For event type classification, only the free-text incident descriptions were used as model inputs. To abide by the privacy regulations, all PSE reports were anonymized after data extraction.

Data Preprocessing

To prepare the incident descriptions for classifier input, we apply the following preprocessing steps.

Data Cleaning. PSE texts were anonymized and cleaned for classifier input. For static representations, relying on word occurrence and frequency, this involved lowercasing, stemming, removing stop-words and non-alphabetic characters. However, deep-learning representations, which represent sequences of words, used raw text to preserve context and avoid distortions.

Table 1. Distribution of PSE Types.

	Extracted reports ($n = 861$)	
Event type	Count	Percentage (%)
Care coordination or communication	186	21.6
Laboratory test	122	14.2
Medication related	89	10.3
Omission or errors in assessment, diagnosis, and monitoring	67	7.8
Maternal	58	6.7
Equipment or devices	56	6.5
Supplies	49	5.7
Total	**627**	**72.8**

Feature Extraction. We consider two types of static text representations namely binary bag-of-words (CB) and TF-IDF (Manning et al., 2008), and deep-learning contextual representations like RoBERTa (Liu et al., 2019), DeBERTa (P. He et al., 2020). Contextual representations are known to better capture the meaning of words by incorporating their position and context within sentences (Chen et al., 2024; Santiago Gonzalez-Carvajal, 2020; Subakti et al., 2022; Wang, 2020). Additionally, we utilize domain-specific representation, BioMedBERT (Gu et al., 2021) and GatorTron (Yang et al., 2022), trained on biomedical and clinical data, to better capture linguistic patterns.

Data Augmentation. To address the imbalance in our dataset across different event types (see Table 1), which can degrade classifier performance (Hasanin et al., 2019; H. He & Garcia, 2009), we employed the synthetic minority oversampling technique (SMOTE; Chawla et al, 2002). SMOTE was applied to the training dataset after feature extraction at each AL iteration before retraining the classifier, while the validation and test set remained unchanged.

Data Splitting. The dataset was divided into 70% training, 15% validation, and 15% testing sets, using a stratified split. The train set was used for classifier learning, while the validation set facilitated hyperparameter tuning and was used to select the best model across iterations. This approach prevented overfitting and enhanced the stability of AL experiments.

Experimental Setup

In this study, we particularly focus on simulating the pool-based AL sampling approach (Goudjil et al., 2018), by initially introducing a small portion of the labeled dataset to the ML model for training. Then, additional informative samples are selected iteratively by the AL strategy and the model is retrained from scratch to evaluate if the additional samples help improve classification accuracy (see Figure 1).

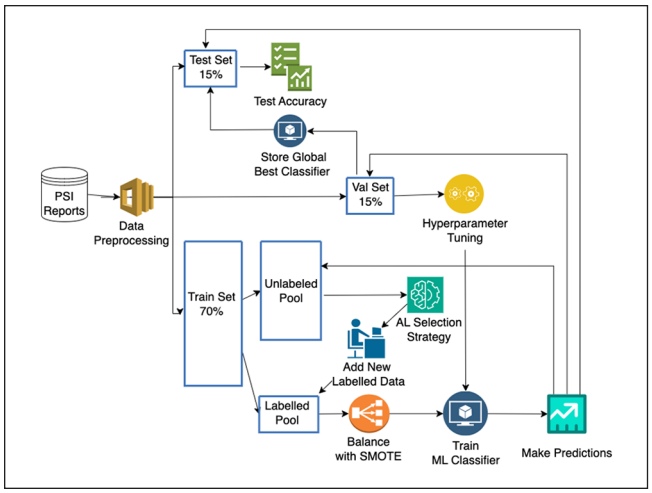


Figure 1. Active learning cycle.

AL Strategies. Several AL selection strategies have been proposed in literature, including uncertainty-based, Query By Committee (QBC), and Information Density (ID) based approaches (Settles & Craven, 2008). Our study explores three uncertainty-based methods: Least Confidence (*LC*), Margin (*M*), and Entropy (*E*) and a cosine similarity-based ID sampling combined with *LC* (*IDCos* × *LC*). Additionally, we implement a random selection (*R*) as a baseline to compare against AL strategies. This method randomly selects instances from the unlabeled set for labeling, in contrast to AL strategies that select instances in a more informed manner. We provide a brief explanation for each AL strategy: (1) *LC* queries the instance with the least confident prediction. (2) *M* measures uncertainty as the difference between the probabilities of the top two class predictions. (3) *E* measures uncertainty using the probability distribution of an event's outcome, with high values indicating multiple equally likely outcomes (Settles, 2011). (4) *IDCos* × *LC* weighs the informativeness of an instance against its similarity to the entire dataset, using measures such as cosine similarity. It is used in conjunction with a base sampling strategy like *LC*, querying instances that are both informative and representative of the input space distribution (Settles, 2011).

All experiments start with a balanced set of 35 samples, with 5 samples per event type, drawn from the training set, allocating the rest to the unlabeled pool. In each AL iteration, 10 samples from the unlabeled pool are selected using the specified AL strategy and, consequently, these samples are added to the labeled set to simulate manual labeling of the selected samples. The ML model was then re-trained on this expanded set, after applying SMOTE oversampling to address potential class-imbalance. This process was repeated until all samples from the unlabeled set were exhausted. We tested the models after each AL iteration with an unseen test set, using them to predict the corresponding PSE type. These predictions were evaluated against the ground truth labels using classification accuracy as the metric. To mitigate bias due to initial selection of labeled samples, each experiment was repeated with 10 random balanced initial labeled sets, and the average test accuracy was reported for each iteration.

AL experiments were conducted using two popular models in text classification tasks: Logistic Regression (LR; Kleinbaum et al., 2002) and Support Vector Machines (SVM; Cristianini & Shawe-Taylor, 2000). Four AL strategies, along with a random sampling baseline, were evaluated across six text representations using both SVM and LR classifiers, totaling 60 experiments.

Results

The performance of ML classifiers, using various AL strategies, was evaluated over 41 iterations by analyzing average test set accuracy until all unlabeled samples were used (see Figures 2 and 3).

To compare the effectiveness of AL strategies against the random baseline *R*, we report the test set accuracy achieved at the 20th iteration, as well as average improvements over *R* at this iteration and between 10 and 30 iterations (see Table 2).

BioMedBERT features yielded the highest accuracy at 20 iterations, with SVM and LR achieving 72% and 71.7%, respectively, surpassing the *R* baseline by 8.4% and 3.9% using both *E* and *IDCos* × *LC* strategies. Following BioMedBERT, RoBERTa, TF-IDF, and GatorTron features reported high accuracies, ranging between 69.2% and 71.6% for SVM and 70.1% to 71.6% for LR classifier. The static TF-IDF representation performed comparably to the clinical domain-specific GatorTron embeddings, indicating that static representations can sometimes match the performance of deep learning models. At the 20th iteration, for SVM, *LC* strategy improves accuracy by 4.93% on average; while for LR, *IDCos* × *LC* strategy increases accuracy by 4.26% on average, over the *R* baseline across all features.

Static CB features reported the lowest accuracy at the 20th iteration, reaching 54.5% with SVM and 61.1% with LR, using *E* and *IDCos* × *LC* strategies, respectively. However, the AL strategies still significantly outperformed the *R* baseline, with improvements of 7.9% and 7.2% by 20 queries. This demonstrates the effectiveness of AL strategies across both static and deep-learning text representations, with highest accuracy gains seen with domain specific embeddings.

Discussion

Many hospitals across the world use incident reporting systems and the data collected on PSEs play a significant role in quality improvement efforts (Hasegawa & Fujita, 2018). Reliable classification of PSE types is crucial for patient safety, and while ML classifiers can accurately classify PSE-reports (Chen et al., 2024; Wang et al., 2017), they typically require large amount of labeled data

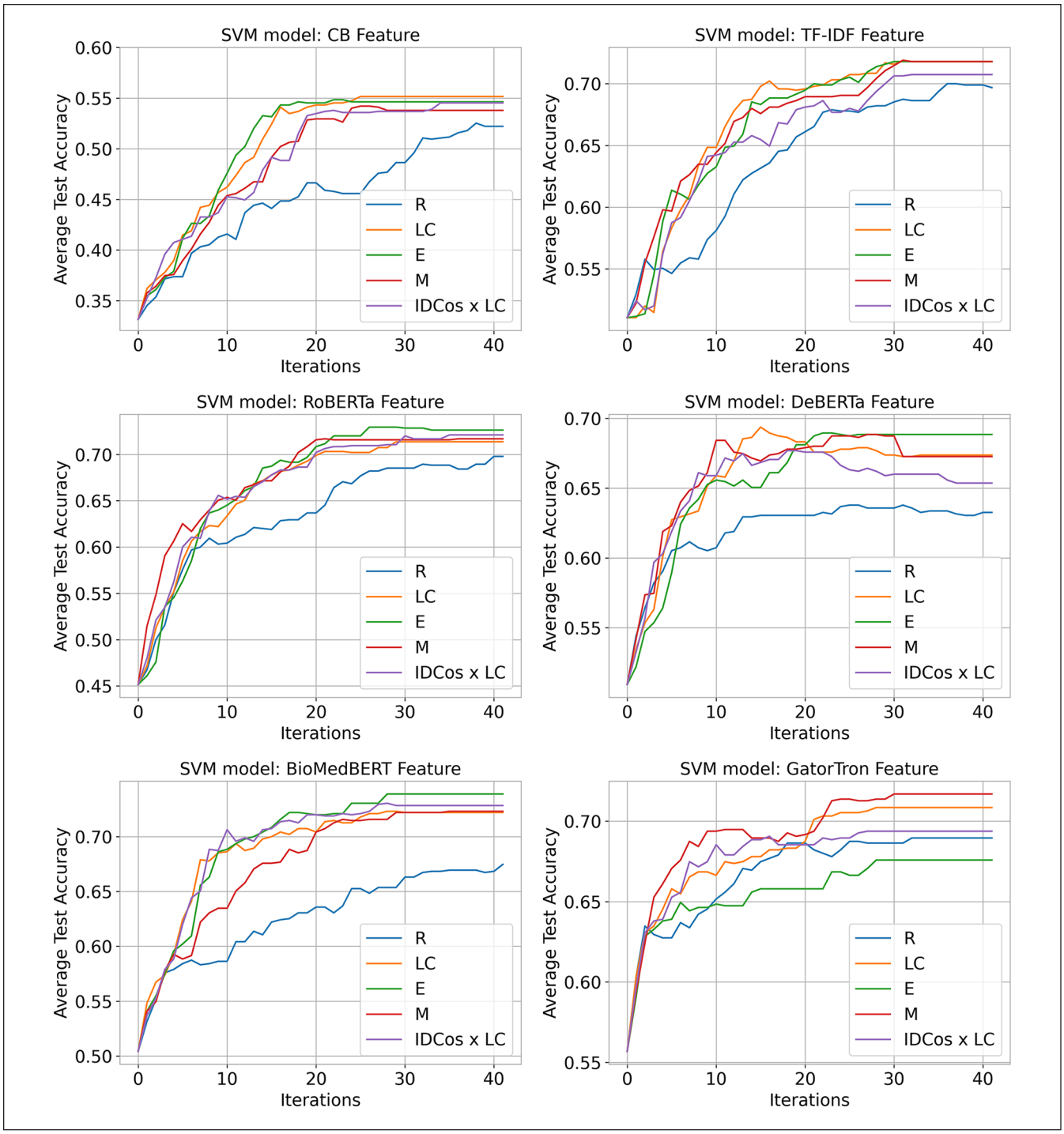


Figure 2. SVM average test accuracy improvements across text representations.

to perform effectively. Our study investigated how AL techniques can minimize labeling efforts by utilizing a smaller subset of informative data points. We evaluated the effectiveness of four AL strategies in improving ML model performance for classifying event types in PSE reports. Additionally, we compared the impact of static versus contextual text representations and generic versus domain-specific embeddings.

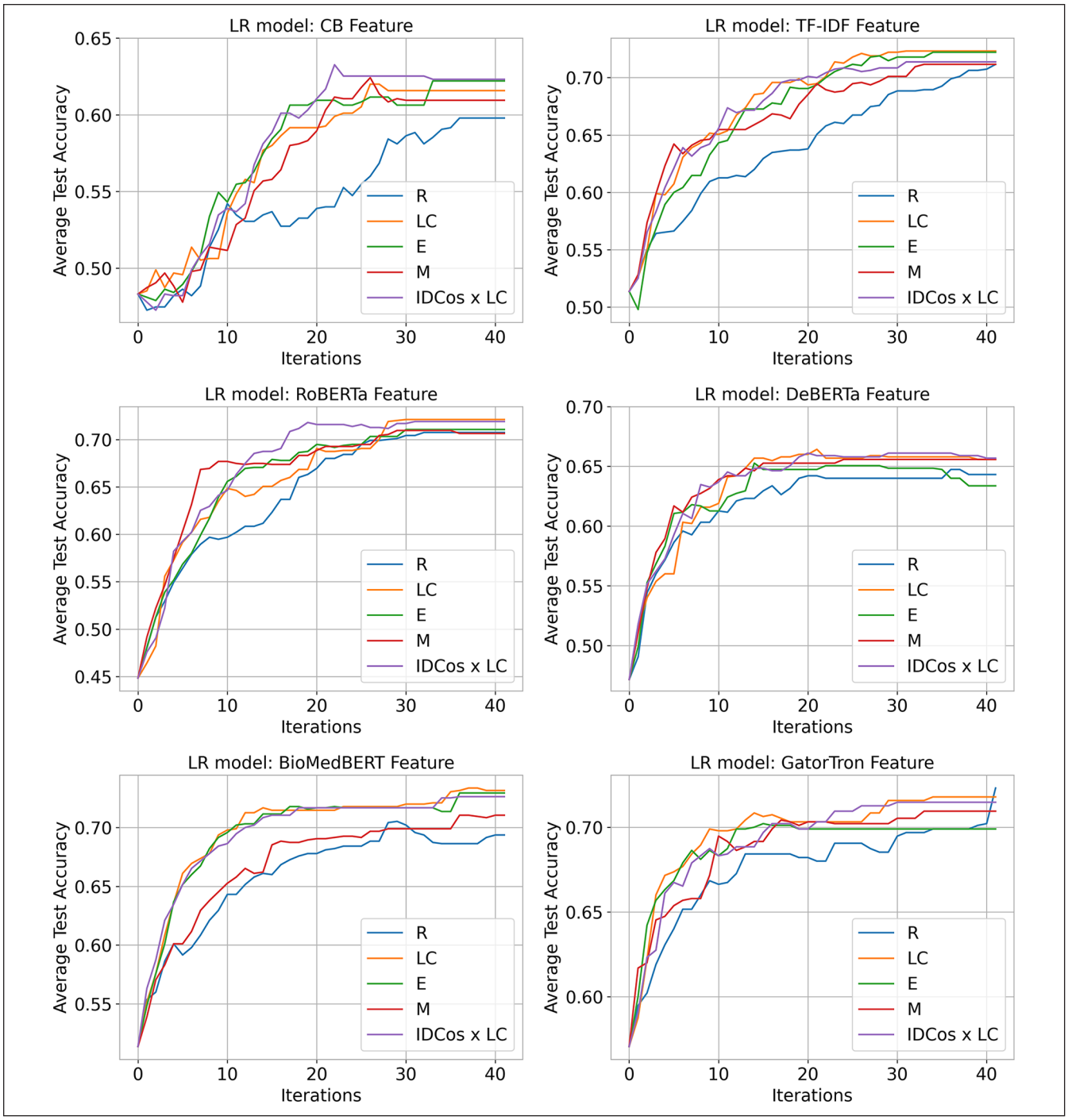


Figure 3. LR average test accuracy improvements across all text representations.

AL Strategies Versus Random Sampling (R)

Our analysis revealed that most AL strategies significantly outperformed random selection (*R*) in terms of accuracy. AL strategies enhance labeling efficiency across all text representations, achieving near peak performance within 10 to 30 iterations (see Figures 2 and 3), reducing the required number of labeled samples by 24% to 69%, considering the total number of samples at the last iteration is 438. These strategies lead to highest accuracy gains particularly for domain-specific embeddings, notably BioMedBERT, improving SVM and LR accuracy to 72% and 71.7% respectively within 20 iterations, using only 54% of PSE-reports. Hence, integrating AL into

Table 2. Comparison of Average Accuracy for SVM and LR Using Active Learning.

	SVM			LR		
AL Strategy	A_{20}	ΔA_{20}	$\Delta A_{10\text{-}30}$	A_{20}	ΔA_{20}	$\Delta A_{10\text{-}30}$
CB						
R	46.6	0	0	53.9	0	0
LC	54.3	7.7	7.6	59.2	5.3	4.4
E	**54.5**	**7.9**	**8.2**	60.9	7.1	4.9
M	52.9	6.3	5.6	58.9	5.1	3.7
IDCos × *LC*	53.5	6.8	5.5	**61.1**	**7.2**	**5.5**
TF-IDF						
R	66.1	0	0	63.8	0	0
LC	**69.6**	**3.5**	**4.4**	69.4	5.6	**5.2**
E	69.5	3.4	3.7	69.1	5.3	4.4
M	68.9	2.8	3.3	68.5	4.7	3.3
IDCos × *LC*	68.1	2	2	**70.1**	**6.3**	4.9
RoBERTa						
R	63.7	0	0	66.9	0	0
LC	69.9	6.2	4.1	69.1	2.1	1.8
E	70.8	7.2	**5.4**	69.5	2.5	2.9
M	**71.6**	**7.9**	5	68.8	1.9	3
IDCos × *LC*	70.2	6.5	4.4	**71.6**	**4.6**	**4.4**
DeBERTa						
R	63.1	0	0	64.2	0	0
LC	**68.3**	**5.3**	4.9	**66**	1.8	**2.1**
E	68.1	5.1	4.3	64.7	0.5	1.1
M	67.9	4.8	**5.1**	65.3	1.1	1.8
IDCos × *LC*	67.6	4.5	3.9	**66.1**	**1.9**	2
BiomedBERT						
R	63.6	0	0	67.8	0	0
LC	70.4	6.8	7.7	71.5	3.7	**4**
E	**72**	**8.4**	**8.8**	**71.7**	**3.9**	3.8
M	70.4	6.8	6.2	69.1	1.3	0.9
IDCos × *LC*	**72**	**8.4**	8.5	**71.7**	**3.9**	3.7
GatorTron						
R	68.6	0	0	68.2	0	0
LC	68.7	0.1	1.3	70.3	2.1	**2.2**
E	65.8	−2.8	−1.8	69.9	1.7	1.5
M	**69.2**	**0.5**	**2.2**	**70.3**	**2.1**	1.6
IDCos × *LC*	68.5	−0.1	1	69.9	1.7	1.8

Note. This table compares various AL strategies against the random baseline (R) for both SVM and LR models, using three metrics: A_{20} (accuracy at the 20th iteration), ΔA_{20} (accuracy improvement over *R* at the 20th iteration), and ΔA_{10-30} (average accuracy improvement over *R* between the 10th and 30th iteration). The representation achieving highest accuracy at 20th iteration is highlighted in yellow.

PSE-report classification can streamline labeling, reduce workload for PSE analysts, and optimize both classifier accuracy and resource allocation. It will enable analysts to focus on annotating fewer incidents for model development, fostering human-AI collaboration.

Best AL Strategy

Among the AL strategies, Entropy (*E*) and the combination of Information Density (ID) Cosine with Least Confidence (*IDCos* × *LC*) were notably effective for SVM and LR, respectively. This demonstrates the efficacy of AL strategies, particularly when integrated with representativeness measures such as ID.

Static Versus Contextual Representations

For both ML models, static CB representations consistently underperformed other features in all experiments. However, static TF-IDF features sometimes matched or even surpassed

the performance of contextual DeBERTa and GatorTron embeddings. Therefore, we cannot definitively state that contextual text representations always outperform static ones. Nonetheless, contextual features can represent text more efficiently in lower-dimensional spaces than TF-IDF enabling faster training, often with comparable or superior accuracy.

Generic Versus Domain-Specific Representations

BioMedBERT domain-specific contextual embedding slightly outperforms generic English RoBERTa embeddings by 20 iterations, which may be attributed to their ability to better capture medical terminology present in the text. With SVM, BioMedBERT achieves the largest overall improvement from the *R* baseline between 10 and 30 iterations, leading to a 7.8% increase in accuracy averaged across all AL strategies. However, with LR, static TF-IDF features yield the highest overall improvement from *R*, increasing accuracy by 4.5%. Thus, we conclude, while domain specific contextual embeddings generally lead to better performance compared to static text representations, they do not necessarily result in larger improvements using AL strategies across all models. Interestingly, when SVM is used with GatorTron features, the random baseline *R* performs relatively well, even outperforming the Entropy strategy and performing comparably with other AL strategies.

Regardless of the text features used, AL strategies prove to be effective tools in helping improve labeling efficiency over random sampling. By 20 iterations of AL sampling, we observe average improvements over the random baseline of 5.03% in accuracy for SVM, and 3.48% for LR, respectively, across all features.

Limitations

The ML models developed in our study were trained using PSE reports from maternal care units of one hospital in the United States, potentially limiting the generalizability of the results to different settings. Additionally, the scope of the analysis was limited by the small quantity of PSE reports available. Only the seven most frequent classes were used for this study. Thus, by not accounting for rarer event types, the performance of our models might be overestimated. We recommended that future research assess the efficacy of ML models integrated with AL strategies over more extensive and disparate datasets.

Conclusion

PSE labeling is particularly challenging as it requires specialized knowledge and human intervention, making full automation difficult. AL aims to support rather than replace manual labeling by filtering instances that are most likely to improve classification accuracy. AL also helps address the imbalanced-data problem typical of PSEs, by prioritizing rare or difficult-to-classify incidents, thus learning from a diverse incident set (Attenberg et al., 2013). PSE reports can change over time due to evolving medical practices and emerging safety issues. AL systems can adapt to these changes more effectively than static models, as they can incrementally learn and re-train from the newly labeled data (Perkonigg et al., 2021). Therefore, AL emerges as an essential tool in the classification of PSE-reports, offering more insightful, efficient, and effective use of this critical data in improving patient safety.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Agency for Healthcare Research and Quality (AHQR) [Grant no. 1R03HS027680].

ORCID iDs

Shehnaz Islam https://orcid.org/0000-0002-4032-1445
Eldan Cohen https://orcid.org/0000-0001-5767-6683

References

Albolino, S., Tartaglia, R., Bellandi, T., Amicosante, A. M. V., Bianchini, E., & Biggeri, A. (2010). Patient safety and incident reporting: Survey of Italian healthcare workers. *BMJ Quality & Safety*, *19*(Suppl 3), i8–i12.

Attenberg, J., & Ertekin, Ş. (2013). Class imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 101–149.

Brubacher, J. R., Hunte, G. S., Hamilton, L., & Taylor, A. (2011). Barriers to and incentives for safety event reporting in emergency departments. *Healthc Q*, *14*(3), 57–65.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, H., Cohen, E., Wilson, D., & Alfred, M. (2024). A machine learning approach with Human-AI collaboration for automated classification of patient safety event reports: Algorithm development and validation study. *JMIR Human Factors*, *11*(1), e53378.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dimitrakakis, C., & Savu-Krohn, C. (2008, February). Cost-minimising strategies for data labelling: Optimal stopping and active learning. In S. Hartmann & G. Kern-Isberner (Eds.), *International Symposium on Foundations of Information and Knowledge Systems* (pp. 96–111). Springer Berlin Heidelberg.

Donaldson, M. S., Corrigan, J. M., & Kohn, L. T. (Eds.). (2000). *To err is human: Building a safer health system*. National Academies Press.

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.

Fong, A., Hettinger, A. Z., & Ratwani, R. M. (2015). Exploring methods for identifying related patient safety events using structured and unstructured data. *Journal of Biomedical Informatics*, *58*, 89–95.

Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2018). A novel active learning method using SVM for text classification. *International Journal of Automation and Computing*, *15*, 290–298.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1–23.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, *6*, 69.

Hasegawa, T., & Fujita, S. (2018). *Patient Safety Policies: Experiences, effects and priorities; lessons from OECD member states*. Ministry of Health, Labour and Welfare.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv Preprint arXiv:2006.03654*.

Herzer, K. R., Mirrer, M., Xie, Y., Steppan, J., Li, M., Jung, C., Cover, R., Doyle, P. A., & Mark, L. J. (2012). Patient safety reporting systems: Sustained quality improvement using a multidisciplinary team and "good catch" awards. *The Joint Commission Journal on Quality and Patient Safety*, *38*(8), 339–347.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression* (p. 536). Springer-Verlag.

Koike, D., Ito, M., Horiguchi, A., Yatsuya, H., & Ota, A. (2022). Implementation strategies for the patient safety reporting system using Consolidated Framework for Implementation Research: A retrospective mixed-method analysis. *BMC Health Services Research*, *22*(1), 409.

Liang, C., & Gong, Y. (2017). Automated classification of multi-labeled patient safety reports: A shift from quantity to quality measure. *Studies in Health Technology and Informatics*, *245*, 1070–1074.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv Preprint arXiv:1907.11692*.

Makary, M. A., & Daniel, M. (2016). Medical error—the third leading cause of death in the US. *BMJ*, *353*, i2139.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Ong, M. S., Magrabi, F., & Coiera, E. (2010). Automated categorisation of clinical incident reports using statistical text classification. *Quality and Safety in Health Care*, *19*(6), e55–e55.

Perkonigg, M., Hofmanninger, J., & Langs, G. (2021, June). *Continual active learning for efficient adaptation of machine learning models to changing image acquisition* [Conference session]. International Conference on Information Processing in Medical Imaging, Springer International Publishing, Cham, 649–660.

Santiago Gonzalez-Carvajal, E. C. M. (2020). Comparing BERT against traditional machine. *Retrieved*, *5*(17), 2020.

Settles, B. (2009). Active learning literature survey. In Computer Sciences Technical Report (No. 1648). University of Wisconsin--Madison. https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf?sequence=1

Settles, B. (2011). *Synthesis lectures on artificial intelligence and machine learning: Active learning*. Morgan & Claypool Publishers. https://link.springer.com/book/10.1007/978-3-031-01560-1

Settles, B., & Craven, M. (2008, October). *An analysis of active learning strategies for sequence labeling tasks* [Conference session]. Proceedings of the 2008 Conference on Empirical methods in Natural Language Processing, 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics. Retrieved from https://aclanthology.org/D08-1112

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data*, *9*(1), 15.

Wang, Y., Coiera, E., Runciman, W., & Magrabi, F. (2017). Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *BMC Medical Informatics and Decision Making*, *17*, 1–12.

Wang, Y., Hou, Y., Che, W., & Liu, T. (2020). From static to dynamic word representations: A survey. *International Journal of Machine Learning and Cybernetics*, *11*, 1611–1630.

Wolf, Z. R., & Hughes, R. G. (2008). Error reporting and disclosure. In R. G. Hughes (Ed.), *Patient safety and quality: An evidence-based handbook for nurses*. Agency for Healthcare Research and Quality. Chapter 35 (pp 1–24). Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK2652/

World Health Organization. (2021). *Global patient safety action plan 2021–2030: Towards eliminating avoidable harm in health care*. Author.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. *NPJ Digital Medicine*, *5*(1), 194.