MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)
© 2024 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI231036

Automating the Identification of Safety Events Involving Machine Learning-Enabled Medical Devices

Ying WANG[a,1], David LYELL[a], Enrico COIERA[a] and Farah MAGRABI[a]
[a] *Australian Institute of Health Innovation, Macquarie University, Australia*
ORCiD ID: Ying Wang https://orcid.org/0000-0001-8537-3954, David Lyell https://orcid.org/0000-0002-2695-0368, Enrico Coiera https://orcid.org/0000-0002-6444-6584, Farah Magrabi https://orcid.org/0000-0002-8426-5588

Abstract. With growing use of machine learning (ML)-enabled medical devices by clinicians and consumers safety events involving these systems are emerging. Current analysis of safety events heavily relies on retrospective review by experts, which is time consuming and cost ineffective. This study develops automated text classifiers and evaluates their potential to identify rare ML safety events from the US FDA's MAUDE. Four stratified classifiers were evaluated using a real-world data distribution with different feature sets: report text; text and device brand name; text and generic device type; and all information combined. We found that stratified classifiers using the generic type of devices were the most effective technique when tested on both stratified (F1-score=85%) and external datasets (precision=100%). All true positives on the external dataset were consistently identified by the three stratified classifiers, indicating the ensemble results from them can be used directly to monitor ML events reported to MAUDE.

Keywords. Machine learning, medical device, text classifier, safety event, rare class classification, unbalanced dataset

1. Introduction

Machine learning (ML) enabled systems have the potential to improve healthcare delivery [1]. However, there are inherent risks to patient safety when technology is poorly designed, implemented or used [2,3]. With use in clinical settings and by consumers safety events involving ML-enabled medical devices are emerging [4,5]. For example, a patient suffering a heart attack delayed seeking care after receiving an automatic interpretation of "normal sinus rhythm" from an over-the-counter consumer ECG device [5]. One source of information about safety events involving ML devices is the US FDA Manufacturer and User Facility Device Experience (MAUDE) database [6]. However, the FDA does not report whether devices utilise ML therefore current analysis of these safety events relies on manual review which is time consuming and expensive. Considering over 2 million events were reported to MAUDE in 2021, manual review on an ongoing basis is not feasible.

[1] Corresponding Author: Ying Wang, email: ying.wang@mq.edu.au.

Text classifiers driven by ML methods have been shown to be feasible for identifying events about a variety of patient safety problems such as falls and medication errors [7-10]. We have previously demonstrated the feasibility of using text classification to identify safety events involving health IT from MAUDE [11]. In this study we sought to identify safety events involving ML from MAUDE, a rare class among other large-scale safety problems with medical devices. By manually reviewing MAUDE we identified 275 events (0.005%) involving ML devices approved since 2015 [5]. To better model their skewed distribution, stratified classifiers were trained, validated and tested using datasets with the real-world distribution. In addition to free text reports, structured device information, including generic types and brand names, were examined to improve classifier performance. We also evaluated the generalizability of classifiers by testing them on new events reported to the FDA.

2. Methods

We performed a classic ML training and testing process involving the following steps:

2.1. Data collection, annotation and pre-processing

For gold-standard labels, we used 275 reports about ML safety events that were identified in a previous study by manually searching MAUDE [3]. The majority of events (n=258) involved devices approved after 2018. We therefore collected general safety reports between 1 January 2018 and 31 October 2021 (n=5,393,062) via the OpenFDA API. This combined dataset (n=5,393,337) was divided into training (75%) and testing (25%) sets (Table 1). The training set was further split into training and validation subsets for 10-fold cross-validation to optimize classifier parameters and identify the most effective classifier for testing. Generalizability of the method was assessed by testing the best-performing classifiers on an external dataset i.e. new events reported to MAUDE from 1 November 2021 to 31 March 2022 (N=895,627). As events often included multiple reports (incl. updates or follow-up investigations), narratives were combined in descending order of reporting date. Empty reports were excluded, and text was changed into lower case. To provide informative features for classification, word tokenization, removal of stop words, stemming and lemmatization were applied [8].

Table 1. Composition of stratified datasets for training and testing text classifiers.

Datasets	Event type	Stratified dataset (n=)
Training set	ML safety events	207
	General safety events	4,044,796
	Total	**4,045,003**
Testing set	ML safety events	68
	General safety events	1,348,266
	Total	**1,348,334**
External testing set		895,627

2.2. Feature representation

Features were extracted using the bag-of-words model, commonly used in document classification [9]. A bag of unique words from training samples was extracted, then transformed into a numeric representation using term frequency–inverse document frequency (TF-IDF) which is an effective feature representation for SVM classifiers [9,10]. Analysis of labelled reports showed that 80% involved imaging systems and nearly 20% used signal data (e.g., ECG signals) [4,5]. As the type of device indicates useful information, we examined the generic device type and brand name as additional features. These were combined to extend the feature space.

2.3. Training and testing classifiers

We developed binary discriminative classifiers of SVM with radial-basis function kernel as they perform better when training on small samples with a large feature space [8-10]. A 10-fold subsampling cross-validation method was applied to optimize the classifier parameters and the best performing classifiers (achieving the highest F-score) were adopted for testing and examining generalizability. In total, four stratified classifiers were trained and validated using each of the four feature sets: i/ report classifier; ii/ brand name classifier; iii/ generic type classifier; and iv/ combined classifier (integrating report text, brand name and generic type together). Precision, recall, and F1 score metrics were used to evaluate performance.

2.4. Error analysis and verification of classifiers results

Incorrectly classified events were analysed for performance improvement. Classifier identified positives on the external testing set were manually verified by two of the investigators (YW and DL). Disagreements were resolved by discussion.

3. Results

We found the combined classifier performed better than the other three individual classifiers on the stratified testing set, achieving the highest F score of 97% (Table 2). However, it generalized poorly on the external testing set. The three classifiers with different feature sets including report alone, generic type and brand name achieved comparable performance on both stratified and external testing sets. Additional device information did not significantly improve classification performance.

3.1. Generalizability and error analysis on external testing set

Fifty-two events were identified by the three classifiers (Table 2). Of these, 56% (n=29) were identified by all three involving known ML devices. Nine events were identified by two classifiers including eight events by the brand name and generic type classifiers and one event by report and generic name classifiers. Thirteen events were identified by the report classifier alone and one by the brand name classifier.

From error analysis, 77% of 52 events (n=40) were manually verified as true positives (TP). Five false positives (10%) did not involve ML devices but were associated with device input problems, such as poor data acquisition. The last seven false positives (13%) were neither HIT problems nor ML devices. The ratio of identified events (0.0045%) on external dataset is very close to the manually reviewed dataset (0.005%). All 29 events overlapped by the three classifiers were verified as TPs, indicating huge potential of the ensemble decisions from the three classifiers to monitor ML events automatically. Out of 29 events, 90% (n=26) involved data acquisition problems from imaging systems. Two TPs (6.9%) involved an ECG device designed for detecting normal sinus rhythm and several arrythmias. Here events involved contraindicated use, where consumers received a device interpretation of a normal sinus rhythm while suffering a heart attack. In the new events, one user reported that devices suggested atrial fibrillation while experiencing monomorphic ventricular tachycardia, and the other experienced an atrial fibrillation, but device indicated a normal rhythm. Both events were associated with algorithm errors, which is among the most critical safety risks of ML devices.

Table 2. Performance of report classifier, generic type classifier, brand name classifier and combined classifier on testing and external datasets.

Dataset		Report classifier	Generic type classifier	Brand name classifier	Combined classifier
Testing set	False negative (n)	27	22	23	4
	False positive (n)	4	2	5	0
	Total (n)	**31**	**24**	**28**	**4**
	Precision (%)	94.44	97.14	93.15	100
	Recall (%)	71.58	75.56	74.73	93.94
	F1 score (%)	81.44	**85.00**	82.93	**96.88**
External testing set	Classifier identified MLSEs (n)	43	38	38	7,707
Verified as	MLSEs (n)	31	38	38	/
	Data input issues attributed to non-ML	5	0	0	/
	False positives (n)	7	0	0	/

4. Discussion

Our results indicate that stratified text classifiers efficiently identified rare ML events from large collections e the FDA's MAUDE. A direct consequence for modelling such highly imbalanced dataset is that rare-class events cannot be well modelled, and classifiers are not generalizable, failing to identify ML events exclusively [9,11]. Error analysis from the external dataset showed a slightly lower ratio of ML events compared to the gold-standard set. We also found that structured device information did not improve generalizability and the combined classifier failed to predict ML events reliably on the external set. It might because more information increases overfitting risk when the number of rare classes is insufficient and limits generalizability.

Device input problems were the main contributors to events in training set where data acquisition occurred with data errors or contamination, mostly related to artefacts in imaging devices [4,5]. It is not a surprise that new events on external set were highly related to data acquisition issues from imaging devices. Although five false positive events were not associated with ML devices, report narratives commonly described actions to solve data collection-related problems, such as calibrating device to expected operation, or checking device settings or replacement of worn components. The terminology was quite similar to the description of representative events. This misclassification implied that words-based features may not be sufficient to capture the semantic language structure, such as event contributors [8]. Error analysis showed that only the report classifier found new devices. Although the other two classifiers with device information did not improve generalizability, their positive events overlapped with report classifier can be relabeled as ML safety events directly and added to training set for retraining the classifier without human intervention. In the future, an automated process flow of classifying and retraining will be adopted to improve the active learning of ML events in a timely manner. We will also examine transfer learning which has been commonly applied to improve imbalanced classification by taking advantage of auxiliary data from similar domain [12].

5. Conclusions

This study contributes an efficient way to identify safety events involving ML devices. Although report classifiers generalized better than generic type and brand name classifiers, the latter two secure an automated fashion of annotating ML events using their ensemble results. Semantic feature representation and transfer learning techniques maybe necessary to enhance classification performance with this unbalanced dataset.

References

[1] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019 Jun;6(2):94-8, doi: 10.7861/futurehosp.6-2-94.

[2] Donaldson MS, Corrigan JM, Kohn LT, editors. To err is human: building a safer health system. 2000.

[3] Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019 Sep;25(9):1337-40, doi: 10.1038/s41591-019-0548-6.

[4] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ Health Care Inform. 2021 Apr;28(1):e100301, doi: 10.1136/bmjhci-2020-100301.

[5] D. Lyell, Y. Wang, E. Coiera, and F. Magrabi, More than algorithms: an analysis of safety events involving ML enabled medical devices reported to the FDA, in: Under Review.

[6] Gurtcheff SE. Introduction to the MAUDE database. Clin Obstet Gynecol. 2008 Mar;51(1):120-3, doi: 10.1097/GRF.0b013e318161e657.

[7] Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. J Am Med Inform Assoc. 2019 Dec;26(12):1600-8, doi: 10.1093/jamia/ocz146.

[8] Wang Y, Coiera E, Magrabi F. Can Unified Medical Language System-based semantic representation improve automated identification of patient safety incident reports by type and severity? J Am Med Inform Assoc. 2020 Oct;27(10):1502-9, doi: 10.1093/jamia/ocaa082.

[9] Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak. 2017 Jun;17(1):84, doi: 10.1186/s12911-017-0483-8.

[10] Wang Y, Coiera E, Runciman W, Magrabi F. Automating the Identification of Patient Safety Incident Reports Using Multi-Label Classification. Stud Health Technol Inform. 2017;245:609-13.

[11] Chai KE, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):980-5, doi: 10.1136/amiajnl-2012-001409.

[12] Azunre P. Transfer learning for natural language processing. Simon and Schuster; 2021 Aug.