Novel machine learning model to improve performance of an early warning system in hospitalized patients: a retrospective multisite cross-validation study

Hojjat Salehinejad,[a,b,*] Anne M. Meehan,[c] Parvez A. Rahman,[a] Marcia A. Core,[d] Bijan J. Borah,[a,f] and Pedro J. Caraballo[c,e,f]

[a]Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA
[b]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
[c]Department of Medicine, Mayo Clinic, Rochester, MN, USA
[d]Department of Information Technology, Mayo Clinic, Rochester, MN, USA
[e]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA

eClinicalMedicine
2023;66: 102312
Published Online 16 November 2023
https://doi.org/10.1016/j.eclinm.2023.102312

Summary

Background Threshold-based early warning systems (EWS) are used to predict adverse events (Aes). Machine learning (ML) algorithms that incorporate all EWS scores prior to an event may perform better in hospitalized patients.

Methods The deterioration index (DI) is a proprietary EWS. A threshold of DI >60 is used to predict a composite AE: all-cause mortality, cardiac arrest, transfer to intensive care, and evaluation by the rapid response team in practice. The DI scores were collected for adult patients (≥18 y-o) hospitalized on medical or surgical services during 8-23-2021 to 3-31-2022 from four different Mayo Clinic sites in the United States. A novel ML model was developed and trained on a retrospective cohort of hospital encounters. DI scores were represented in a high-dimensional space using random convolution kernels to facilitate training of a classifier and the area under the receiver operator characteristics curve (AUC) was calculated. Multiple time intervals prior to an AE were analyzed. A leave-one-out cross-validation protocol was used to evaluate performance across separate clinic sites.

Findings Three different classifiers were trained on 59,617 encounter-derived DI scores in high-dimensional feature space and the AUCs were compared to two threshold models. All three tested classifiers improved the AUC over the threshold approaches from 0.56 and 0.57 to 0.76, 0.85 and 0.94. Time interval analysis of the top performing classifier showed best accuracy in the hour before an event occurred (AUC 0.91), but prediction held up even in the 12 h before an AE (AUC 0.80 at minus 12 h, 0.81 at minus 9 h, 0.85 at minus 6 h, and 0.88 at minus 3 h before an AE). Multisite cross-validation using leave-one-out approach on data from four different clinical sites showed broad generalization performance of the top performing ML model with AUC of 0.91, 0.91, 0.95, and 0.91.

Interpretation A novel ML model that incorporates all the longitudinal DI scores prior to an AE in a hospitalized patient performs better at outcome prediction than the currently used threshold model. The use of clinical data, a generalized ML technique, and successful multisite cross-validation demonstrate the feasibility of our model in clinical implementation.

Funding No funding to report.

Keywords: Deterioration index; Early warning system; Machine learning; Random convolution kernels

Introduction

Early warning systems (EWS) in hospitals began with a focus on enhancing patient care and reducing adverse events, evolving from manual assessments by healthcare professionals to technologically advanced systems. Initially, bedside monitors allowed for continuous vital sign monitoring, albeit with isolated systems and no standardized response protocols. The introduction of Modified Early Warning Score (MEWS) improved risk identification.

The electronic medical health record (EHR) transformed clinical practice. This deluge of data, albeit

*Corresponding author.
E-mail address: salehinejad.hojjat@mayo.edu (H. Salehinejad).
[f]Both BJB and PJC contributed equally as senior authors.

Research in context

Evidence before this study

We searched PubMed for papers published from database inception to September 28, 2023, using the search terms ("deep learning" OR "artificial intelligence" OR "AI" OR "machine learning") AND ("deterioration index", "early warning system"), without any date or language restrictions. This search returned about 454 results. The deterioration index (DI) score was used during the COVID-19 pandemic as a decision-making support tool. None of the articles proposes a machine learning (ML) model for early prediction of adverse events (AEs) using DI score. This shows the novelty of the proposed model as a decision-making tool for early intervention in hospitalized patients with potential risk of experiencing an AE.

Added value of this study

Our study uses a sample size of encounters (n = 59,617) that yielded DI scores from 52,471 unique patients. Our findings show there is a relationship between retrospective DI scores of a patient and the patient outcome (adverse, no adverse). This relationship was captured by training a novel ML approach using high-dimensional representation of the DI scores, for the first time in literature. The leave-one-out multisite validation showed high generalization performance of the model on different geographically located sites.

Implications of all the available evidence

Our study, with the state-of-the-art multisite cross-validation performance, shows the usability of DI score in predicting adverse events among hospitalized patients. Future work should focus on implementing the model in clinical settings for prospective evaluation of the model for its clinical and cost effectiveness in a controlled set up. This work provides a baseline for further investigation in building novel models to predict AEs from DI score.

important can contribute to cognitive overload and provider burnout.[1,2] Less time for face-to-face patient care increases reliance on electronic monitoring and highlights the need for reliable electronic alerts. The introduction of Rapid Response Systems (RRS) enabled timely intervention for deteriorating patients. Standardized systems like the National Early Warning Score (NEWS) in the UK ensured consistency in monitoring and response protocols across healthcare institutions.

Epic's Deterioration Index (EDI) is a proprietary algorithm and one of the most widely used EWS deployed in hundreds of hospitals across the US.[3,4] This system generates a patient risk score after hospital admission, and it is then regularly calculated based on most recent available data at 15-min intervals until discharge. A DI score value ranges between 0 and 100, defining low (<30 green), intermediate (30–60 orange), or high risk (>60 red) of a composite AE: all-cause mortality, cardiac arrest, transfer to intensive care, and evaluation by the rapid response team.[4–6] Each generated risk score is determined based on routinely recorded physiological, clinical, and laboratory parameters within Epic's EHR to support medical decision-making. The risk score is determined based on age, neurological assessment, cardiac rhythm, oxygen requirement, Glasgow Coma Scale, vital sign measurements (temperature, systolic blood pressure, pulse rate, oxygen saturation, respiratory rate), and laboratory values (hematocrit, white blood cell count, blood urea nitrogen, potassium, sodium, blood pH, platelet count).

Accurate proactive identification of patient deterioration is essential to prevent morbidity and mortality in the hospital setting. Clinical and laboratory parameters are used in EWS to try to prevent adverse events (AEs), including cardiac arrests, and death. Calculated risk scores are used at a predetermined threshold to alert the clinical staff.[7–10] Multiple organizations agree regarding the potential benefits of EWS in combination with Rapid Response Teams (RRT) to save lives in the hospital setting.[10,11] Nonetheless, hesitancy and skepticism persist due to methodological weakness, diversity of outcomes, and lack of convincing evidence of post-implementation impact.[7,12]

Machine learning (ML) can encompass big structured clinical data and discern patterns not obvious to humans. While there has been some movement towards use of ML in prediction of AEs,[8,9] adoption of these methods in healthcare has lagged behind other industries for many reasons.[3,13] Alarms based on single threshold scores may not achieve the accuracy needed to engender trust in the system. Trends rather than a single score or the most recent score may be more useful and informative.[14] ML may detect deterioration associated signals before they are clinically obvious, hopefully at a time point where risk mitigation is feasible.[15]

The DI score is part of the EPIC EHR at our institution. Based on feedback from clinical users we posited that factors other than the threshold score might correlate better with outcome. Thus, the objective of this study was to propose a novel ML model for automated early prediction of AEs based on the retrospective trajectory of all DI scores and compare its performance with the currently deployed threshold model (i.e., DI >60). The proposed model can learn from the pattern of DI scores over time from various patients to predict adverse events in future.

Methods

Clinical setting and study design

Mayo Clinic is an academic institution providing healthcare at different geographical locations including

Arizona, Florida, Minnesota, and Wisconsin in the United States. It has an integrated EHR across all the locations (Epic, Verona, USA).

A ML model for automated early prediction of AEs based on the retrospective DI scores was developed. Given the time series nature of the DI scores, a set of random convolution filters were initialized to represent the retrospective DI scores in a high-dimensional feature space (n = 9996). The extracted features were used to train a classifier. The model was cross-validated on the study cohort for predicting AEs. It was also trained with retrospective DI scores of occurrences at different time intervals prior to the AE. The generalization performance of the model was evaluated based on a leave-one-out cross-validation method in patient populations from four different Mayo Clinic sites.

Ethics

We used retrospective de-identified clinical data abstracted from EHR and the need for informed consent was therefore waived. This study was reviewed and approved by Mayo Clinic Institutional Review Board.

Datasets

The DI scores were collected for adult patients (≥18 y-o) hospitalized on medical or surgical services during 8-23-2021 to 3-31-2022. Patients in the intensive care unit (ICU), emergency department, obstetrics wards and hospice were excluded. The DI score was collected every 15 min for a total of 59,617 encounters (contributed by 52,471 unique patients) in four different locations, Rochester, Minnesota (RST) 25,127 with 2802 AEs, Mayo Clinic Health System (Minnesota and Wisconsin) (MCHS) 16,330 with 779 AEs, Jacksonville, Florida (FLA) 9695 with 825 AEs, and Scottsdale, Arizona (ARZ) 8465 with 1567 AEs. The first DI score was collected 3 h after admission, ensuring enough time to document all the clinical variables needed to calculate the score. The model requires eligible subjects to have a minimum of 10 DI scores collected every 15 min. Therefore, model prediction begins only after 5.5 h after the patient is admitted (3 h for the first DI score and 2.5 h to capture the next 9 DI scores).

Model development and training

The most common application of the DI score classifies patients with a DI score of >60 as high risk, indicating prompt intervention to mitigate the risk of an AE. Two threshold methods were evaluated based on this (DI >60) hypothesis. The first method is prediction of AEs if at least one DI >60 exists in the retrospective DI scores. The second method is prediction of AEs only if the last DI score before an event is >60 (or before dismissal for those without an AE). The proposed model is described in Fig. 1. Three different and independent classifiers (Ridge regression classifier,[16] support vector machine (SVM),[17] extreme gradient boosting (XGBoost)[18]) were trained and evaluated using the high-dimensional representation of the retrospective DI scores using random convolution kernels.[19]

The AE prediction problem using DI scores was modeled as a binary classification problem (class 0: Not AE; class 1: AE). A bank of K 1-dimenstional fixed-length random convolution kernels[19,20] were generated. The weights of each kernel were selected randomly from {−1, 2}. A set of dilation factors per kernel controls the spread of the kernel over an input DI score series. A set of bias terms is then calculated based on the quantiles of the convolution on the input DI score series to generate features in a high-dimensional space (n = 9996). The extracted features are based on the proportion of the positive values after applying the convolution kernels with various bias terms. Typical time series representation and classification models require all input time series to be of the same length. Padding the time series is a one of the major preprocessing methods to prepare the data for such models. Given the input DI score series are of various length, this requirement limits implementation of these models in practice. One of the main advantages of the proposed approach is feature representation of any-length time series without padding. This major feature facilitates implementation

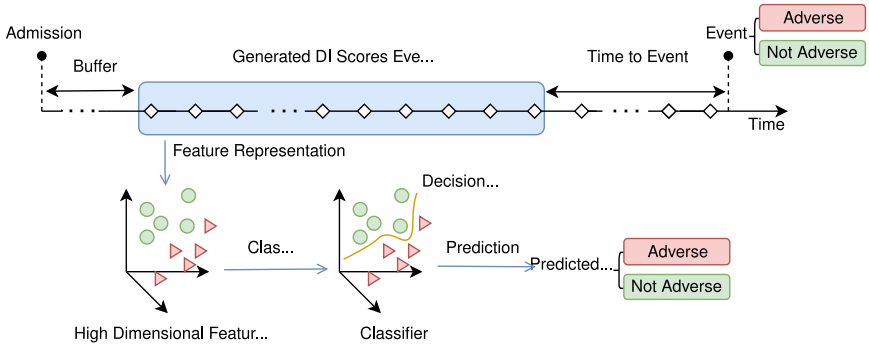

Fig. 1: Proposed machine learning model for adverse event prediction using the longitudinal deterioration index (DI) scores.

Classifier	Hyperparameters
Ridge	Regularization strengths cross-validation: {1e-3, 1e-2, 1e-1, 1}
SVM	Regularization parameter: 0.1; Kernel: Radial basis function
XGBoost	Step size shrinkage: 0.3; Maximum depth of a tree: 10; Subsample ratio of the training instances: 0.5; Number of parallel trees: 10; Number of estimators: 300

Table 1: **Hyperparameters of the classifiers.**

of the proposed model in clinical settings. Once the input DI score series is represented in a high-dimensional feature space, the extracted features are used to train a classifier.

Evaluation and cross-validation

A 10-fold cross-validation was applied and the average and standard deviation of performance metrics were collected. In each independent run, balanced splits of dataset were used for training and testing. Since the dataset was naturally imbalanced, the occurrences from the not adverse outcome data class were randomly down sampled to the number of occurrences in the adverse outcome data class (n = 5973 occurrences per data class). The test dataset was sampled from the balanced dataset with 20% contribution from each data class (totally n = 2390 occurrences) and the rest of the balanced dataset was used as the training dataset (n = 7166 occurrences). For the multisite cross-validation, a leave-one-out cross-validation was performed with respect to the four Mayo Clinic sites across the USA. For each validation site, the best performing model (XGBoost) was trained on the data from other sites and the validation results were collected after a 10-fold cross-validation of the model. The hyperparameters used to train the classifiers are summarized in Table 1.

Statistical analysis

Classification performance of each model was assessed using the area under the receiver operating curve or AUC values. Bootstrap sampling with replacement with 10-fold cross-validation was used to obtain 95% CIs for all performance measures. The models were developed and evaluated using Python programming language, version 3.9 (Python Software Foundation). The threshold for statistical significance was 2-tailed with P = 0.05.

Role of the funding source

No funding to report.

Results

There were 59,617 (51.51% female) encounters, 10.02% (41.69% female) with AEs. The overall average age was 62.07 ± 18.53 y-o, with an average age of 63.15 ± 15.92 y-o and 61.95 ± 18.80 y-o with and without AEs, respectively.

Adverse event prediction performance

The XGBoost classifier trained with the high-dimensional represented features had the best 10-fold cross-validated accuracy with Mean ± STD of 0.88 ± 0.01, F1-score 0.88 ± 0.01, sensitivity 0.85 ± 0.01, and specificity 0.91 ± 0.01 (Table 2). All the evaluated ML approaches have a significantly higher prediction performance than the thresholding approaches. The distribution of 10-fold cross-validated models with respect to the accuracy metric shows no significant outlier and limited variance (Fig. 2A). Accuracy evaluation of the models revealed AUC 0.57 (95% CI, 0.57–0.58) for the thresholding with any DI score >60, AUC 0.56 (95% CI, 0.56–0.56) for the thresholding with the last DI score >60, AUC 0.85 (95% CI, 0.77–0.78) for Ridge, AUC 0.76 (95% CI, 0.64–0.65) for SVM, and AUC 0.94 (95% CI, 0.88–0.88) for XGBoost (Fig. 2B). The XGBoost classifier yielded the best performance (referred to hereafter as best model).

Adverse vs. not adverse probabilistic analysis of classifiers

Distribution analysis of the computed probability values of AE occurrences over 10-fold cross-validation revealed a Mean ± STD of 0.25 ± 0.01 (95% CI, 0.25–0.26) for the thresholding with any DI >60, 0.13 ± 0.01 (95% CI, 0.12–0.14) for the thresholding with the last DI >60,

Model	Sensitivity	Specificity	Accuracy	F1-score	AUC
Thresholding (any)	0.25 ± 0.01	0.89 ± 0.01	0.57 ± 0.01	0.53 ± 0.01	0.57 ± 0.01
Thresholding (last)	0.13 ± 0.01	0.99 ± 0.00	0.56 ± 0.01	0.46 ± 0.01	0.56 ± 0.01
Ridge	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.85 ± 0.01
SVM	0.42 ± 0.02	0.87 ± 0.01	0.65 ± 0.01	0.63 ± 0.01	0.76 ± 0.01
XGBoost	0.85 ± 0.01	0.91 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.94 ± 0.01

Table 2: **Performance (Mean ± STD) of models in predicting adverse events (AEs) from deterioration index (DI) scores after 10-fold cross-validation.**

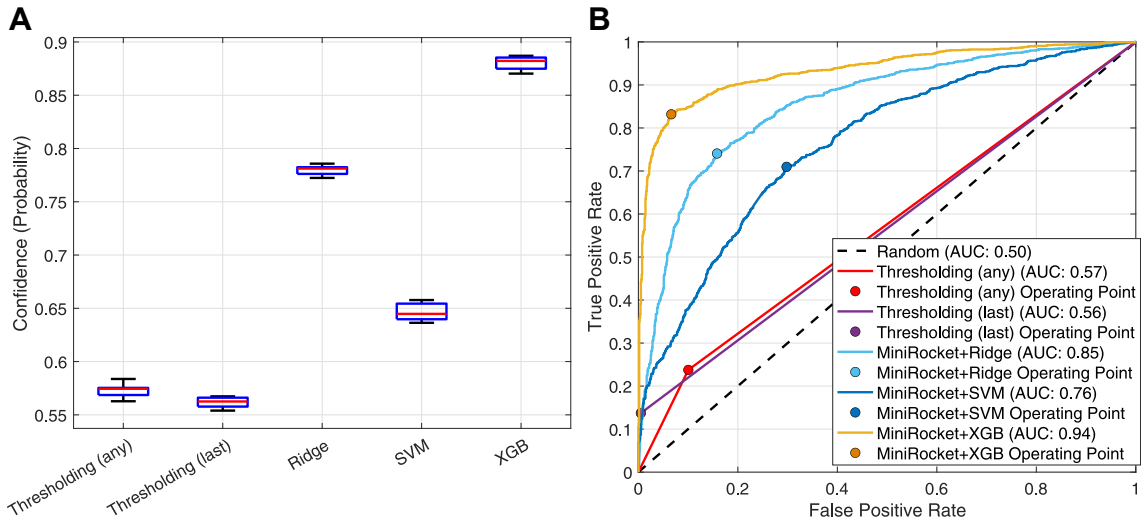

Fig. 2: Results after 10-fold cross-validation. (A) Distribution of the collected accuracy values in predicting events. (B) Receiver operator characteristics (ROC) curve of the classifiers trained with extracted features using random convolution kernels.

0.62 ± 0.01 (95% CI, 0.61–0.62) for Ridge, 0.59 ± 0.01 (95% CI, 0.58–0.60) for SVM, and 0.85 ± 0.01 (95% CI, 0.84–0.85) for the best model. Similarly, distribution analysis of the generated probability values of not AE occurrences over 10-fold cross-validation revealed a Mean ± STD of 0.89 ± 0.01 (95% CI, 0.89–0.90) for the thresholding with any DI score >60, 0.99 ± 0.01 (95% CI, 0.99–0.99) for the thresholding with the last DI score >60, 0.62 ± 0.01 (95% CI, 0.61–0.62) for Ridge, 0.58 ± 0.01 (95% CI, 0.58–0.58) for SVM, and 0.90 ± 0.01 (95% CI, 0.89–0.90) for the best model.

Normalized count of AE and not AE occurrences over average distribution of predicted probability values (Appendix) generated by the best model over 10-fold cross-validation shows the greatest number of occurrences with correct prediction of event type fall in a confidence range of 0.9–1.0 in prediction of not AE and AE with 0.82 and 0.78, respectively. The normalized count values for the probability range 0.0–0.1 is 0.05 and 0.11 for not AE and AE, respectively (Fig. 3).

Advanced prediction of adverse events

Accurate advance warning of an impending adverse event should provide an opportunity for caregivers to intervene. A time course analysis of the best model (XGboost) in the hours before AEs was performed and the AUC values calculated (Fig. 4). The Mean ± STD accuracy was best minus 1 h prior to an AE, with AUC 0.84 ± 0.01 (95% CI, 0.83–0.84). The best model performed well across a minus 12 h interval prior to the event with minus 3 h AUC 0.81 ± 0.01 (95% CI, 0.80–0.81), minus 6 h AUC 0.77 ± 0.02 (95% CI, 0.75–0.78), minus 9 h AUC 0.73 ± 0.02 (95% CI, 0.72–0.74), and minus 12 h AUC 0.72 ± 0.02 (95% CI, 0.71–0.73) (Fig. 4A). The AUC curves are in Fig. 4B.

The Mean ± STD PPV of the best model in predicting an AE minus 1 h prior to the event is 0.80 ± 0.01 (95% CI, 0.80–0.81), minus 3 h prior to the event is 0.78 ± 0.01 (95% CI, 0.78–0.79), minus 6 h prior to the event is 0.76 ± 0.02 (95% CI, 0.74–0.77), minus 9 h prior to the event is 0.73 ± 0.01 (95% CI, 0.72–0.73), and minus 12 h prior to the event is 0.71 ± 0.02 (95% CI, 0.70–0.73). The Mean ± STD NPV of the best model in predicting an AE minus 1 h prior to the event is 0.88 ± 0.01 (95% CI, 0.87–0.88), minus 3 h prior to the event is 0.84 ± 0.01 (95% CI, 0.83–0.84), minus 6 h prior to the event is 0.78 ± 0.02 (95% CI, 0.77–0.80), minus 9 h prior to the

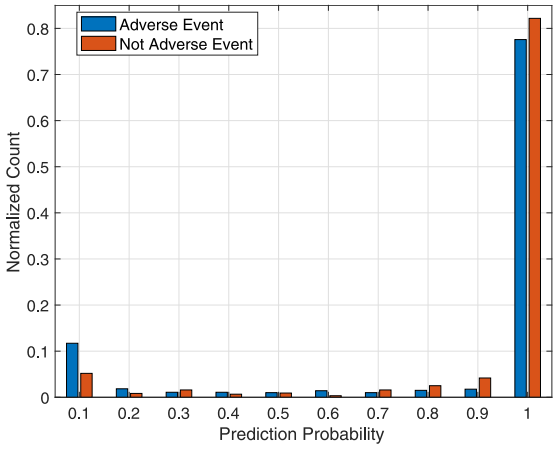

Fig. 3: Normalized count of adverse event (AE) and not AE occurrences over average distribution of predicted probability values generated by the best model after averaging over 10-fold cross-validation.

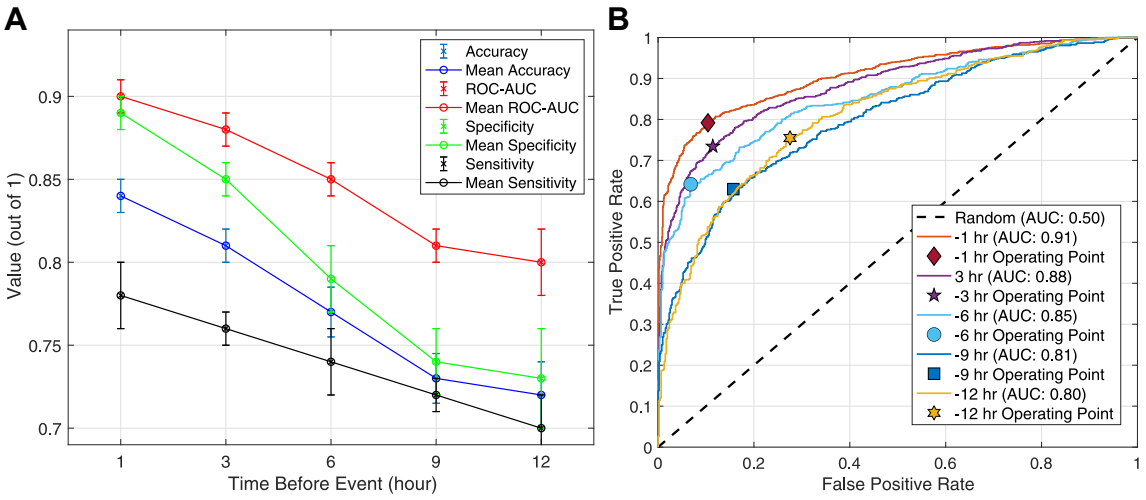

Fig. 4: (A) Average accuracy, area under the receiver operator characteristics curve (AUC), specificity, sensitivity, and corresponding standard deviation of the best model in prediction of the correct event using retrospective deterioration index (DI) scores for various hours before the event. Results are reported after averaging over 10-fold cross-validation. (B) Receiver operator characteristics (ROC) curve of the best model in prediction of the correct event using retrospective DI scores for various hours before the event. Results are reported after averaging over 10-fold cross-validation.

event is 0.74 ± 0.01 (95% CI, 0.73–0.75), and minus 12 h prior to the event is 0.73 ± 0.02 (95% CI, 0.71–0.75).

Multisite cross-validation

We performed a one-leave-out multisite validation of the best model. The validation dataset of each site reflects its natural distribution with respect to the AE and not AE classes and was imbalanced. A 10-fold cross-validation was conducted per site validation. These four sites are geographically distinct, representing patient populations with heterogenous socio-demographic characteristics, including rural vs. urban populations. For each validation, results revealed a Mean ± STD balanced accuracy of 0.82 ± 0.01 for RST, 0.78 ± 0.01 for MCHS, 0.88 ± 0.01 for ARZ, and 0.81 ± 0.01 for FLA (Table 3). The AUC curves are in Fig. 5. The results showed generalization performance of the best model validated on different clinic sites.

The Mean ± STD PPV of the best model in predicting AEs validated on the retrospective DI scores of the RST site is 0.96 ± 0.01 (95% CI, 0.96–0.96), MCHS site is 0.98 ± 0.01 (95% CI, 0.98–0.98), ARZ site is 0.95 ± 0.01 (95% CI, 0.95–0.95), and FLA site is 0.97 ± 0.01 (95% CI, 0.96–0.97). The NPV of the best model in predicting AEs validated on the retrospective DI scores of the RST site is 0.75 ± 0.01 (95% CI, 0.75–0.76), MCHS site is 0.79 ± 0.02 (95% CI, 0.78–0.80), ARZ site is 0.97 ± 0.01 (95% CI, 0.96–0.97), and FLA site is 0.81 ± 0.01 (95% CI, 0.81–0.82).

Discussion

Our study shows improved performance of an EWS, the DI, over the established threshold danger score of 60 or higher by application of a novel ML model to use all scores generated for a given patient during the hospital encounter prior to an event. DI scores represented in a high dimensional feature space using random convolution kernels were used to train a classifier, and then tested on a comparable patient pool. Time interval analysis of the new method showed acceptable performance over a 12-h prediction horizon. Multisite cross

Validation Site	Sensitivity	Specificity	Accuracy	Balanced accuracy	F1-score	AUC
RST	0.66 ± 0.01	0.97 ± 0.01	0.94 ± 0.01	0.82 ± 0.01	0.84 ± 0.01	0.91 ± 0.01
MCHS	0.56 ± 0.01	0.99 ± 0.01	0.97 ± 0.01	0.78 ± 0.01	0.82 ± 0.01	0.91 ± 0.01
ARZ	0.76 ± 0.01	0.99 ± 0.01	0.95 ± 0.01	0.88 ± 0.01	0.91 ± 0.01	0.95 ± 0.01
FLA	0.62 ± 0.02	0.99 ± 0.01	0.96 ± 0.01	0.81 ± 0.01	0.84 ± 0.01	0.91 ± 0.01

For each validation site, the model is trained on the DI scores from the other sites.

Table 3: **Multisite cross-validation performance (Mean ± STD) of the best model in predicting adverse events (AEs) from deterioration index (DI) scores after 10-fold cross-validation.**

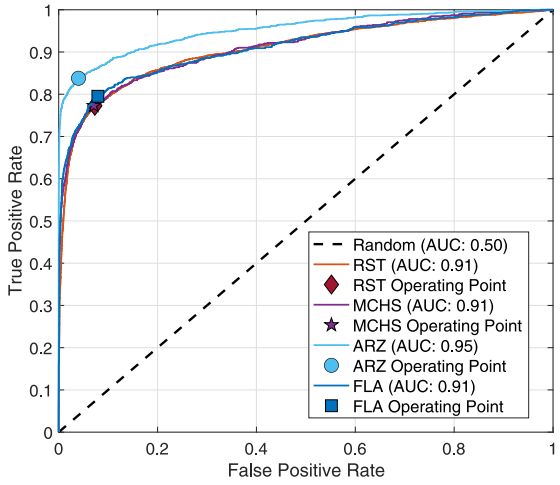

Fig. 5: Receiver operator characteristics (ROC) curve of the best model after leave-one-out multisite validation and 10-fold cross-validation per site.

validation using a leave-one-out approach showed broad applicability across four geographically distinct clinical sites with heterogenous health care populations.

Publications on this EWS are sparse, at least in part due to the proprietary nature of the algorithm. Several studies[5,21] highlight moderate performances and several drawbacks of using DI in predicting adverse events. The DI has been found to have fair performance, improve patient outcomes and reduce ICU admissions.[4] However, health systems utilize DI in conflicting ways and with substantially disparate thresholds.[5]

We find that the threshold approach is useful when negative (i.e., when patients are deemed unlikely to deteriorate) but lacks sensitivity at identifying those who will deteriorate. Winslow et al.[22] performed an analysis using logistic regression on 27 discrete time variables to identify patients at risk of adverse events. Similarly, Escobar et al.[23] used a logistic-regression model to generate hourly Advance Alert Monitor (AAM) scores based on a threshold system, where an AAM score of 5 (alert threshold) indicated a 12-h risk of clinical deterioration of 8% or more. The predictors laboratory tests, individual vital signs, neurologic status, severity of illness and longitudinal indexes of coexisting conditions, care directives, and health services indicators (e.g., length of stay) were used in this study and the principal dependent variable was mortality within 30 days after an AAM alert. The innovation that our model offers is to use the entire series of DI scores instead of a single DI score as in threshold approach, which improves the predictive ability of the model significantly. Our results show that projection of DI scores to a high-dimensional space using random convolution kernels and training an ML model can help predict adverse events in hospitalized patients using retrospective DI scores.

The performance of our novel model compared favorably with the performance of other EWS. Liu and others evaluated the performance of five commonly used EWS: National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), Between the Flags (BTF), Quick Sequential Sepsis-Related Organ Failure Assessment (qSOFA), and Systemic Inflammatory Response Syndrome (SIRS).[24] Direct comparison was not done but our model performs well based on published literature; AUC of 0.87 (95% CI, 0.87–0.87), for the NEWS compared with our model AUC 0.94 (95% CI, 0.88–0.88). Implementation of the NEWs has been associated with a reduction of AE including cardiopulmonary resuscitation and transfers to intensive care.[25] A fair comparison between different models is not easy due to the significant differences in the care setting, clinical data used as predictors and outcome definitions. None of the studies identified in two recent systematic reviews used an ML model like the one described in our study.[8,9]

Feature representation of time series with random convolution kernels is a state-of-the-art method for classification of time series.[19,20] It has achieved a better performance than deep learning and other classification tools with almost deterministic outcomes on standard time series benchmark datasets.[19,20,26] Particularly, this technique works very well on limited-imbalanced data.[27] From a computational complexity perspective, it is much faster than deep learning models which is a major factor in regular prediction of AEs. It has also been used for various applications such as functional near infrared spectroscopy signals classification,[28] human activity recognition,[27,29] driver's distraction detection using electroencephalogram (EEG) signals,[30] and transcription factor binding site prediction for DNA sequences.[31]

The high computational complexity of existing state-of-the-art methods for time series classification makes these methods slow, even for smaller datasets, and intractable for large datasets. However, high-dimensional representations can capture complex relationships in the data, which is a fundamental concept in ML and data analysis. By leveraging these rich representations, high-dimensional feature representations can significantly improve the discriminative power of ML models. In many ML tasks, including time series classification, more features can enhance predictive modeling accuracy. This trade-off between computational complexity and the potential for improved model performance underscores the importance of carefully selecting and engineering features, especially when dealing with high-dimensional data.

The represented features in a high-dimensional space can be classified based on the target data classes using classifiers like Ridge, SVM, and XGBoost. While Ridge regression is a straightforward and interpretable linear classifier, its performance may be suboptimal when faced with intricate non-linear feature patterns. In

contrast, SVM can effectively handle both linear and non-linear challenges. However, its performance relies on the choice of kernel, parameter tuning, and the complexity of the features. The computational demands of SVM are rooted in intricate mathematical calculations for identifying an optimal margin of separation between data classes, and this complexity scales with the data size, especially when non-linear kernels are employed. On the flip side, XGBoost, through the amalgamation of gradient boosting, regularization techniques, efficient parallel processing, and other attributes, emerges as a potent and versatile algorithm. It frequently delivers high classification performance across a broad spectrum of ML tasks.

The limitations of this study are those traditionally attributed to the use of ML to develop predictive models in healthcare.[32] Mistrust of a score that has been calculated by a computer and ML models cannot be underestimated when promoting clinical adoption.[33] The algorithm used to derive the EPIC DI score is proprietary and not available for scrutiny. Intrinsic bias in the algorithm before the score is calculated cannot be addressed.[21] There is always the risk of inaccurate clinical data generated by human observation, bias, overfitting, lack of transparency and interpretability. However, we attempted to overcome these limitations by providing supporting analysis across disparate clinic settings. In addition, the DI score values are generated by the Epic electronic health record (EHR) software, which might not be adopted by other health systems. The other limitation of this method is a 5-h delay in starting the prediction process. This is due to the admission process and waiting for collecting the clinical data (i.e., labs, etc.) as well as the nature of time series prediction models, which require time series (enough number of samples along time) to make a prediction.

Regarding generalization of the data, the data used to train and validate the models was collected from four different Mayo Clinic sites across the US. However, the model should be validated in other institutions, particularly outside the US for extended generalization performance evaluation. The one-site-out cross-validation of the model targets generalization performance evaluation of the models using the natural imbalanced dataset of each validation site.

This study presents a novel ML algorithm for early prediction of adverse events in hospitalized patients based on the Epic DI score. This model utilizes all the retrospective scores generated for a patient throughout their hospital stay in advance of an adverse event to make predictions. The study's innovation lies in using the entire series of DI scores, rather than a single score, which enhances predictive ability. Additionally, the model's performance is compared with other ML-based approaches, emphasizing its potential for clinical use. However, challenges such as mistrust of computer-generated scores and data availability in different health systems, along with a 5-h prediction delay, are acknowledged as limitations.

Contributors

HS, AMM, BJB, and PJC participated in designing the study and reviewing the literature. BJB and PJC supervised the study. PAR collected the data. HS contributed to code writing, model training, and data analysis. HS prepared the code repository. HS contributed to expert review and data interpretation. HS designed the figures. HS drafted the manuscript. HS, AMM, BJB, and PJC revised the manuscript. All authors had access to the result data presented in the final manuscript and all authors read and approved the final manuscript. Verification and access to raw data was available to all authors. The decision to submit was made by all authors. BJB and PJC contributed equally to this work and share senior authorship.

Data sharing statement

The material and dataset are not publicly available due to restrictions by privacy laws. Model output and ground truth labels are available upon reasonable request by contacting the corresponding author. The source code used in this project can be made available on reasonable request by contacting the first author.

Declaration of interests

BJB is a consultant to Boehringer-Ingelheim and Exact Sciences on unrelated projects.

Other authors declare no competing interests.

Acknowledgements

The authors thank Mayo Clinic Kern Center for the Science of Health Care Delivery for their support.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102312.

References

1 Harry E, Pierce RG, Kneeland P, Huang G, Stein J, Sweller J. Cognitive load and its implications for health care. *NEJM Catal Carryover*. 2018;14(2):4.
2 Collins R. Clinician cognitive overload and its implications for nurse leaders. *Nurse Lead*. 2020;18(1):44–47.
3 Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF, Intelligence AAOTFoA. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. *Transl Vis Sci Technol*. 2020;9(2):45.
4 Mu E, Jabbour S, Dalca AV, Guttag J, Wiens J, Sjoding MW. Augmenting existing deterioration indices with chest radiographs to predict clinical deterioration. *PLoS One*. 2022;17(2):e0263922.
5 Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc*. 2021;18(7):1129–1137.
6 Cummings BC, Blackmer JM, Motyka JR, et al. External validation and comparison of a general ward deterioration index between diversely different health systems. *Crit Care Med*. 2023;51(6):775–786.
7 Gerry S, Bonnici T, Birks J, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ*. 2020;369:m1501.
8 Jahandideh S, Ozavci G, Sahle BW, Kouzani AZ, Magrabi F, Bucknall T. Evaluation of machine learning-based models for prediction of clinical deterioration: a systematic literature review. *Int J Med Inform*. 2023;175:105084.
9 Muralitharan S, Nelson W, Di S, et al. Machine learning–based early warning systems for clinical deterioration: systematic scoping review. *J Med Internet Res*. 2021;23(2):e25187.
10 Smith ME, Chiovaro JC, O'Neil M, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc*. 2014;11(9):1454–1465.
11 Institute for Health Improvement. *Early warning systems: scorecards that save lives*; 2009. Retrieved from: www.ihi.org/resources/pages/improvementstories/earlywarningsystemsscorecardsthatsavelives.aspx. Accessed August 4, 2023.

12 McGaughey J, Fergusson DA, Van Bogaert P, Rose L. Early warning systems and rapid response systems for the prevention of patient deterioration on acute adult hospital wards. *Cochrane Database Syst Rev*. 2021;11(11):CD005529.
13 Tachkov K, Zemplenyi A, Kamusheva M, et al. Barriers to use artificial intelligence methodologies in health technology assessment in central and East European countries. *Front Public Health*. 2022;10:921226.
14 Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*. 2016;102:1–5.
15 Brankovic A, Hassanzadeh H, Good N, et al. Explainable machine learning for real-time deterioration alert prediction to guide preemptive treatment. *Sci Rep*. 2022;12(1):11734.
16 Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67.
17 Steinwart I, Christmann A. *Support vector machines*. Springer; 2008.
18 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. San Francisco, California, USA: Association for Computing Machinery; 2016:785–794.
19 Dempster A, Petitjean F, Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min Knowl Discov*. 2020;34(5):1454–1495.
20 Dempster A. MiniRocket. A very fast (Almost) deterministic transform for time series classification. In: Schmidt DF, ed. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining; 7/14/2021*. Singapore; 2021:248–257.
21 Byrd TF, Southwell B, Ravishankar A, et al. Validation of a proprietary deterioration index model and performance in hospitalized adults. *JAMA Netw Open*. 2023;6(7):e2324176.
22 Winslow CJ, Edelson DP, Churpek MM, et al. The impact of a machine learning early warning score on hospital mortality: a multicenter clinical intervention trial. *Crit Care Med*. 2022;50(9):1339–1347.
23 Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med*. 2020;383(20):1951–1960.
24 Liu VX, Lu Y, Carey KA, et al. Comparison of early warning scoring systems for hospitalized patients with and without infection at risk for in-hospital mortality and transfer to the intensive care unit. *JAMA Netw Open*. 2020;3(5):e205191.
25 Wu CL, Kuo CT, Shih SJ, et al. Implementation of an electronic national early warning system to decrease clinical deterioration in hospitalized patients at a tertiary medical center. *Int J Environ Res Public Health*. 2021;18(9):4550.
26 Salehinejad H, Wang Y, Yu Y, Jin T, Valaee S. S-Rocket: selective random convolution kernels for time series classification. *arXiv [preprint]*. 2022. arXiv:2203.03445.
27 Salehinejad H, Valaee S. LiteHAR: lightweight human activity recognition from WIFI signals with random convolution kernels. In: *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2022:4068–4072.
28 Andreu-Perez AR, Kiani M, Andreu-Perez J, et al. Single-trial recognition of video gamer's expertise from brain haemodynamic and facial emotion responses. *Brain Sci*. 2021;11(1):106.
29 Salehinejad H, Hasanzadeh N, Djogo R, Valaee S. Joint human orientation-activity recognition using WiFi signals for human-machine interaction. In: *ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2023:1–5.
30 Tan CW, Salehi M, Mackellar G. *Detecting driver's distraction using long-term recurrent convolutional network*. 2020.
31 Morrow AK, Shankar V, Petersohn D, Joseph AD, Recht B, Yosef N. Convolutional kitchen sinks for transcription factor binding site prediction. *arXiv Genomics*. 2017. https://doi.org/10.48550/arXiv.1706.00125.
32 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
33 Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. *J Med Ethics*. 2022;48(10):764–768.