CEEM Clinical and Experimental Emergency Medicine

Clin Exp Emerg Med 2023;10(2):132-137
https://doi.org/10.15441/ceem.23.041

Review Article

eISSN: 2383-4625

Current challenges in adopting machine learning to critical care and emergency medicine

Cyra-Yoonsun Kang[1], Joo Heung Yoon[2]

[1]Department of Internal Medicine, John H. Stroger Jr. Hospital of Cook County, Chicago, IL, USA
[2]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Over the past decades, the field of machine learning (ML) has made great strides in medicine. Despite the number of ML-inspired publications in the clinical arena, the results and implications are not readily accepted at the bedside. Although ML is very powerful in deciphering hidden patterns in complex critical care and emergency medicine data, various factors including data, feature generation, model design, performance assessment, and limited implementation could affect the utility of the research. In this short review, a series of current challenges of adopting ML models to clinical research will be discussed.

Keywords Machine learning; Challenges; Artificial intelligence; Critical care

Received: 5 April 2023
Revised: 25 April 2023
Accepted: 28 April 2023

Correspondence to: Joo Heung Yoon
Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh Medical Center, 3459 Fifth Ave, Pittsburgh, PA 15213, USA
Email: yoonjh@upmc.edu

Capsule Summary

What is already known

Machine learning is a powerful tool to handle complex datasets and could serve as a promising research methodology to improve healthcare outcomes in critical care and emergency medicine.

What is new in the current study

Various challenges and pitfalls should be considered in conducting clinical research using machine learning.

How to cite this article:
Kang CY, Yoon JH. Current challenges in adopting machine learning to critical care and emergency medicine. Clin Exp Emerg Med 2023;10(2):132-137. https://doi.org/10.15441/ceem.23.041

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/).

© 2023 The Korean Society of Emergency Medicine

INTRODUCTION

Over the past decades, the field of machine learning (ML) has made great strides in medicine. The greater availability of large datasets—supported by lower data storage fees and the advent of cloud computing—has provided a rich source of information that can be mined for ML algorithms [1]. Furthermore, enhanced computing power has accelerated the development of ML algorithms that can process complex, heterogeneous datasets involving imaging data, electronic health records, and waveforms [2]. The dramatic evolution of ML techniques has inspired researchers to build numerous prototype models for prediction, diagnosis, and prognostication. A number of these models performed equal or better in prediction and diagnosis than existing conventional statistics-based solutions. Various ML models, for example, predicted critical care outcomes—e.g., emergency department (ED) to intensive care unit (ICU) transfer and in-hospital mortality—more accurately than existing screening tools, such as the Modified Early Warning Score, the National Early Warning Score, and the Sequential Organ Failure Assessment [3,4]. In radiology, ML-based radiomics models performed better than radiologists, especially in detecting subtle changes indescribable to the naked eye [5–7].

Nonetheless, several challenges must be overcome before ML algorithms can be adapted to the clinical workflow of the ICU or ED (Fig. 1). In this review, we outline these challenges—both in developing and applying models for critical care medicine—and offer potential solutions.

CRITICAL CARE DATA

The extensive and granular datasets available in critical care medicine are promising resources for developing ML models. It is especially true when the data contain a lot of noise from the environment, such as raw vital sign data acquired from the ED. However, several challenges remain in data standardization and preprocessing.

Building a reliable ML model requires a highly structured, large, and multicenter dataset that allows for proper model training as well as internal and external validation. But obtaining such a dataset is no easy task. Electronic health records contain information collected as part of the workflow and are thus fraught with errors such as mislabeling and omission, and variation in intrahospital and interhospital reporting of clinical data creates additional challenges in data mining.

Thankfully, a number of efforts are underway to standardize data formatting. For example, Fast Healthcare Interoperability Resources (FHIR) is a preformatted healthcare database that analytics platforms can easily access and deconstruct [8]. More recently, the Critical Care Data Exchange Format (CCDEF) was developed to facilitate encoding, storing, and exchanging clinical and physiologic data across institutions globally [9]. CCDEF generates a diverse and well-represented dataset that is ideal for developing robust ML algorithms.

Once data has been gathered, optimizing the dataset through preprocessing is necessary before being employed in an ML model. Preprocessing can involve data cleaning, normalization, feature extraction, and selection to address issues with erroneous, missing, or imprecise data. Proper data preprocessing requires tremendous resources and time due to considerable size of datasets containing physiological and imaging data and directly influences the performance of the ML algorithm [10,11]. Nonetheless, the omission of preprocessing in ML studies appears common and impairs a fair assessment of the model. Even with proper preprocessing, features that may be difficult to capture in data, such as heart rate variability and interhospital differences in ICU resources, can still confound the model performance [12,13].

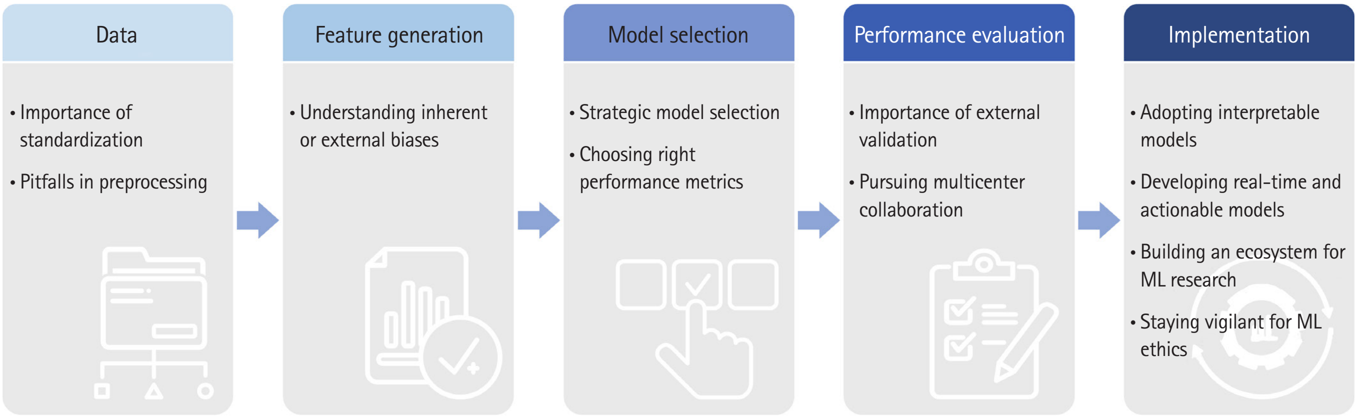
Data
• Importance of standardization
• Pitfalls in preprocessing
Feature generation
• Understanding inherent or external biases
Model selection
• Strategic model selection
• Choosing right performance metrics
Performance evaluation
• Importance of external validation
• Pursuing multicenter collaboration
Implementation
• Adopting interpretable models
• Developing real-time and actionable models
• Building an ecosystem for ML research
• Staying vigilant for ML ethics

Fig. 1. Challenges in adopting machine learning (ML).

FEATURIZATION AND MODEL SELECTION

Featurization involves converting raw variables into numerical vectors that ML algorithms can process. Feature selection and extraction are important steps to identify clinically salient features and improve model predictability [14,15]. Therefore, features need to reflect underlying pathophysiologic mechanisms or core characteristics of data structure. For feature extraction, one needs to understand inherent biases in feature extraction (measurement error, self-reporting, human judgment) which can lead to the problem of fairness in ML. When choosing the models, they need to be built on the same well-understood variables to fairly compare their performance. Likewise, careful contextual consideration needs to be given when choosing the ML model. Another potential problem is using various ML models without consideration of data structure and study objectives, including simultaneous use of supervised and unsupervised learning just to see better performance for publication.

EVALUATION OF ML MODELS

Once the features are decided, models learn from the selected features within the different hyperparameter settings to predict desired outcomes. Choosing the best model requires calibration, which estimates concordance between the predicted probabilities and observed outcomes. Model calibration is a necessary step for measuring the relative performance of models and can assess underfitting or overfitting [16]. To evaluate fair model selection, future studies should use appropriate methods of model calibration, accounting for population size and the type of model [17]. Finally, the evaluation of a model's predictive performance should assess its clinical applicability. Most of the existing studies have used the area under the receiver operating characteristics (AUROC) curve to evaluate model performance. The AUROC curve is plotted by calculating sensitivity and specificity at different thresholds. An AUROC curve provides a single performance value that is easy to interpret and compare [18]. Because the AUROC curve accounts for the true positive rate and the false positive rate, it is useful in a balanced dataset that values both positive and negative outcomes equally. The existing ML studies use the AUROC curve indiscriminately. However, the datasets used to build ML models in medicine tend to have smaller positive classes compared to negative classes. In such imbalanced datasets, the area under the precision recall curve (AUPRC) is more appropriate. The AUPRC represents positive predictive values for each true positive rate and thus focuses on positive values and is not influenced by true negatives [19]. Therefore, the use of AUPRC to evaluate models used for problems such as diagnosis, screening, and predicting mortality will lead to better estimation of the models' performance in real clinical settings.

MODEL VALIDATION

Although ML models are being rapidly developed for potential use in critical care medicine, their clinical utility is still unclear with a lack of generalizability. Small datasets, especially those produced from a single institution could lead to overfitting in a similar environment, but often not performing in other datasets. To properly implement ML algorithms in the clinical workflow of the ICU or ED, the algorithms must be externally validated. However, a recent study assessing the clinical readiness of existing ML models revealed that only 5% of the models have been externally validated [20]. Ongoing data-standardization initiatives, such as CCDEF, will hopefully integrate large datasets across multiple centers, which in turn can be employed for model validation. Further, successful model performance on prospectively collected data can demonstrate the value of ML support in clinical settings and assure clinicians of its safety.

Lack of external validation and small datasets can lead to overfitting and reduce generalizability. While curating multicenter databases in a centralized center can resolve such issues, it invites other challenges, especially in international configurations due to concerns over privacy, technical process, and data ownership. Federated learning (FL) provides a more efficient solution by allowing multiple collaborators to train models in parallel and send the updates to a central server to be integrated into a consensus model [21]. Recently, 20 centers across the globe collaborated to build a comprehensive FL model for predicting outcomes from COVID-19 infection [22]. Trained on electronic medical records and chest x-ray images, the FL model performed 16% better than a locally-trained model in predicting 24-hour oxygen treatment, with improved generalizability of 25.3%. As above, data sharing in a federated environment could overcome the limitation of external validation where data governance and privacy become obstacles.

MODEL IMPLEMENTATION

Despite a great deal of evolution in ML, several challenges still remain before their deployment. When the models are deployed at the bedside to alert physicians of impending crises, for example, their overt sensitivity can cause unintended harm. Excessive alarms that do not require clinicians' immediate awareness can lead to missed real events. The ML-based alerting tools should be designed in a judicious manner to maximize the accuracy of the alarms.

Although not required, model interpretability could have a paramount impact on successful implementation. To implement a model in a clinical setting, its decisions must be verified by clinicians before use. An earlier study tested ML algorithms to build a model that could triage patients with pneumonia and predict mortality. Among the algorithms evaluated in the study, multitask neural networks were deemed to be the most accurate [23]. However, later analysis revealed a pattern in the algorithm that linked asthma to lower mortality—explained by the fact that patients with asthma received more attentive care and close monitoring, thus leading to better outcomes.

Since the study was published, efforts have been made to improve the interpretability of ML models. Prescience, a complex ML algorithm, accurately predicts the near-term risk of hypoxemia during surgery and displays the specific risk factors that informed its prediction [24]. The model is built using a gradient boosting machine based on both static features—such as body mass index, age, and sex—and dynamic parametric values, such as tidal volume and vital signs. The impact of each feature is assigned Sharply values, which makes the predictions more interpretable through the concept used in the game theory.

The successful implementation of ML models relies on clinicians' confidence in the models, which depends on how well users can explain the models' decision-making process [25]. For an ML model to play a supportive role to physicians, it is paramount to focus on features that are available real time and actionable in clinical settings. Therefore, researchers and developers should involve clinicians in an early phase of design to facilitate smooth integration into clinical workflow [26,27].

As seen from the above examples, the ML model implementation in ICU or ED population still could be far-fetched from the practice pattern of clinicians. To address that aspect from the end-user standpoint, the US Food and Drug Administration (FDA) under the Department of Health and Human Services has published an action plan for the use of artificial intelligence and ML, specifically in the form of "Software as a Medical Device (SaMD)." In the white paper, the FDA argued the need for the "Predetermined Change Control Plan" to assure the quality of usable ML models for patient care [28]. Similar efforts could be seen in the Bridge2AI, a US National Institutes of Health funding initiative to promote the ML modeling environment at a multicenter level [29]. Filling the gap between the developer's machine to the real world remains to be a huge challenge for healthcare researchers.

Lastly, For the ML model to be successfully implemented at the bedside and performed in the realm of current clinical practice, not only the ML researchers but also clinicians (end-users) need to understand the ethical aspects of adopting it. First, all data and system-related biases should be minimized with vigilance. Bias could include a thorough examination of the environment where the model was initially developed, identification of inadequate perpetuation of systematic errors abundant in different types of healthcare practices, and so on [30]. Secondly, both researchers and clinicians need to recognize that patients and colleagues might not accept or adhere to the results of the ML model. In a survey, around 60% of Americans feel uncomfortable using artificial intelligence-driven healthcare and are also suspicious that the ML model could improve their outcomes [31]. More studies are required to understand the underlying characteristics of barriers to accepting ML in practice. Thirdly, clinicians still should strive for excellence in patient care in their traditional ways. This due diligence is mainly to avoid the moral hazards smoldering when high-performing ML models become functional at the bedside.

CONCLUSION

Research in the application of ML to critical care or emergency medicine has seen tremendous growth owing to the increasing availability of highly granular, large critical care databases. Nevertheless, for the ML-based models to serve as reliable decision support tools, the steps involved in model building to final implementation must be carefully examined. As the validity of the models is entirely dependent on the datasets, standardization of the data-gathering process and proper preprocessing of the datasets are imperative. A large portion of published studies lack a description of the preprocessing and featurization bringing into question the clinical saliency of the features involved in model performance. Moreover, selecting the right model should involve model calibration to objectively compare the accuracy of the model prediction. Even after choosing the well-calibrated model, many of the earlier studies failed to have the models externally validated which may result in the models overfitting the testing dataset reducing generalizability. In addition, the decision-making process of the algorithms should be explainable. Lastly, the endeavor to create a sustainable and scalable ecosystem should be pursued across different healthcare systems where ethical datasets could be collected and shared for fair ML research.

ETHICS STATEMENTS

Not applicable.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

This work was supported by a US National National Institutes of Health grant (No. K23GM138984) acquired by Joo Heung Yoon.

AUTHOR CONTRIBUTIONS

Conceptualization: all authors; Formal analysis: all authors; Funding acquisition: JHY; Methodology: all authors; Writing–original draft: all authors; Writing–review & editing: JHY. All authors read and approved the final manuscript.

ORCID

Cyra-Yoonsun Kang	Not available
Joo Heung Yoon	https://orcid.org/0000-0002-0127-8384

REFERENCES

1. Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. Crit Care 2022;26:75.
2. Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater 2019;5:83.
3. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Ann Emerg Med 2019;73:334–44.
4. Yoon JH, Jeanselme V, Dubrawski A, Hravnak M, Pinsky MR, Clermont G. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. Crit Care 2020;24:661.
5. Frix AN, Cousin F, Refaee T, et al. Radiomics in lung diseases imaging: state-of-the-art for clinicians. J Pers Med 2021;11:602.
6. Guiot J, Vaidyanathan A, Deprez L, et al. A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev 2022;42:426–40.
7. Wu YJ, Liu YC, Liao CY, Tang EK, Wu FZ. A comparative study to evaluate CT-based semantic and radiomic features in preoperative diagnosis of invasive pulmonary adenocarcinomas manifesting as subsolid nodules. Sci Rep 2021;11:66.
8. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9:e21929.
9. Laird P, Wertz A, Welter G, et al. The critical care data exchange format: a proposed flexible data standard for combining clinical and high-frequency physiologic data in critical care. Physiol Meas 2021;42:065002.
10. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min 2017;10:35.
11. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 2016;104:444–66.
12. Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020;24:478.
13. Wu TT, Zheng RF, Lin ZZ, Gong HR, Li H. A machine learning model to predict critical care outcomes in patient with chest pain visiting the emergency department. BMC Emerg Med 2021;21:112.
14. Abromavicius V, Plonis D, Tarasevicius D, Serackis A. Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models. Electronics 2020;9:1133.
15. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PLoS One 2019;14:e0218760.
16. Rajaraman S, Ganesan P, Antani S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One 2022;17:e0262838.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
18. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
19. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Cohen WW, Moore A, editors. Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25–29; Pittsburgh, PA, USA; Association for Computing Machinery; 2006. p. 233–40.
20. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PW; Right Data Right Now Collaborators. Machine learning in intensive care medicine: ready for take-off? Intensive Care Med 2020;46:1486–8.
21. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in

medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020;10:12598.

22. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 2021;27:1735–43.
23. Samek W, Wiegand T, Muller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv [Preprint] 2017 Aug 28. https://doi.org/10.48550/arXiv.1708.08296
24. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60.
25. van der Meijden SL, de Hond AA, Thoral PJ, et al. Intensive care unit physicians' perspectives on artificial intelligence-based clinical decision support tools: preimplementation survey study. JMIR Hum Factors 2023;10:e39114.
26. de Hond AA, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5:2.
27. Thoral PJ, Fornasa M, de Bruin DP, et al. Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: uniting intensivists and data scientists. Crit Care Explor 2021;3:e0529.
28. US Food and Drug Administration (FDA). Artificial intelligence and machine learning in software as a medical device [Internet]. FDA; 2021 [cited 2023 Mar 31]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
29. US National Institutes of Health (NIH) Common Fund. Bridge to artificial intelligence (Bridge2AI) [Internet]. NIH; 2023 [cited 2023 Mar 31]. Available from: https://commonfund.nih.gov/bridge2ai
30. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21); 2021 Oct 5–9; Virtual; Association for Computing Machinery; 2021. Article 17.
31. Tyson A, Pasquini G, Spencer A, Funk C. 60% of Americans would be uncomfortable with provider relying on AI in their own health care. Pew Research Center; 2023.