Original research article

DIGITAL HEALTH

Hybrid triaging assistance algorithm for continuous patient monitoring

DIGITAL HEALTH
Volume 11: 1–11
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076251388141
journals.sagepub.com/home/dhj
Sage

Zongrun Li[1], Julie Lockington[2], Samya Torres[3], Nooshin Jafari[3], Michael Lim[3], Dragan Andjelic[4], Edmond Cretu[4], Kendall Ho[3] and Bhushan Gopaluni[1]

Abstract

Objective: This study aims to develop and evaluate a transformer-based neural network model that leverages both vital signs and chief complaints to predict patient acuity more accurately and automatically. This study is a part of a major project, which envisions continuous monitoring in the emergency department, while this model provides a machine learning based tool to risk stratify patients.

Methods: This study utilized the public MIMIC-IV-ED dataset, containing patients' vital signs, chief complaints, and triage acuity levels. We developed multiple machine learning models, including a baseline model using only vital signs and a hybrid model that contains the transformer architecture and feed-forward neural networks, incorporating numerical and textual data types. A secondary analysis was performed after filtering inconsistent data points to test the model in an idealized scenario.

Results: Models incorporating chief complaints achieved significantly higher accuracy (over 70%) compared to baseline models that relied solely on vital signs (around 60%). After filtering out inconsistent data, the hybrid model's accuracy improved to over 90%.

Conclusion: Integrating chief complaints is critical for improving the accuracy of AI-driven triage models. These findings highlight the potential for hybrid systems to enhance patient monitoring and prevent deterioration in ED waiting rooms. Future work should focus on incorporating more diverse datasets and time series data to further validate and improve model performance.

Keywords

Emergency medicine, triaging, machine learning, natural language processing, hybrid model

Received: 21 July 2025; accepted: 29 September 2025

Introduction

Patients present to emergency department (ED) every day with a wide variety of illnesses in various degrees of severity. The vital function of the triage unit of every ED is to rapidly assess and risk stratify these patients upon their arrivals to discern their level of acuity, prioritize those that are most critically ill to receive immediate medical attention, and place others in descending order to criticality for later assessment and management. As a result, patients deemed less ill are streamed to the ED waiting room until medical personnel can assess them.

In addition to the patients' chief complaints and their background history, triage also relies on assessing patients'

[1]Chemical Engineering, University of British Columbia, Vancouver, Canada
[2]Vancouver Coastal Health Research Institute, Vancouver, Canada
[3]Emergency Medicine, University of British Columbia, Vancouver, Canada
[4]Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada

Corresponding authors:
Zongrun Li, Chemical Engineering, University of British Columbia, Vancouver, V6T 1Z3, British Columbia, Canada.
Email: gregli@student.ubc.ca

Bhushan Gopaluni, Chemical Engineering, University of British Columbia, Vancouver, V6T 1Z3, British Columbia, Canada.
Email: bhushan.gopaluni@ubc.ca

vital signs to assess severity. For those patients in the waiting room, ongoing monitoring would be ideal to prevent deterioration, including repeated vital signs monitoring as one strategy. Some examples of this warning system, such as the MEWS or NEWS,[1,2] monitor patient vital signs repeatedly to detect deterioration. However, such systems currently have shown either low sensitivity or high false alarm rates, with the investigators arguing that these systems have not incorporated ED patients' chief complaints into the monitoring. Ideally, if this system, together with the patients' presenting problems and their health background, can be deployed in the ED waiting room, we could monitor patients' clinical status and avoid undetected deterioration of patients while waiting for medical attention.

While traditional telemedicine has long relied on nurses interpreting verbal chief complaints,[3] recent advancements have focused on AI-powered symptom checkers.[4] This study builds upon this proposes a hybrid transformer-based neural network model that can provide predictions of patient acuity based on both vital signs and patient chief complaints, possibly utilizing chief complaints that are provided through telemedicine practices. By leveraging this combined dataset, the model aims to provide more accurate predictions of patient acuity compared to systems reliant solely on vital signs. Subsequently, by developing this model, this study argues that chief complaints are crucial to the triaging process and should be incorporated in early warning systems.

This study is a part of a major project, which envisions continuous monitoring in the ED. The model pairs extremely well with continuous monitoring and telemedicine, allowing accurate risk stratification on multiple patients at once.

For the remaining sections of the study, we will explain in detail the issues of overcrowding and the lack of adequate monitoring in ED in backgrounds section; briefly review and explain related studies in the research area in related works section; explain the data source, architecture of our baselines and the proposed model in methods section; compare the performances between the baselines, our proposed model and the model trained on filtered data in the results section; discuss the performance differences, and issues with the current protocols in discussion section; and finally, highlight the limitations and conclude the study.

Background

Currently, in Canada, when a patient arrives at the ED, a triage nurse inquires about their presenting complaint, past medical history, records the patient's vital signs, including heart rate, blood pressure, body temperature, oxygen saturation, and respiratory rate, and conducts a rapid physical assessment that would usually take several minutes. Then, the nurse assigns to this patient a score ranging from 1 to 5 based on the Canadian Triage and Acuity Scale (CTAS).[5] Patients with CTAS scores of 1 and 2 are highest risk and therefore prioritized for more immediate assessment and treatment from ED teams. However, patients with lower risk chief complaints and presentations, reflected by lower CTAS scores ranging from 3 to 5, typically wait longer periods of time until health professionals are available to see them.

A critical concern in this process is that patients with initially assigned moderate risk scoring can deteriorate, occasionally even rapidly, while waiting without substantial monitoring. This deterioration can escape detection by health professionals,[5] making early detection a matter of life and death. Also, with current ED overcrowding and staffing shortages, this results in further limitations to closely monitor patients in the ED waiting rooms.

To monitor patients in EDs, the NEWS is one such system that can potentially be used. NEWS analyzes patient vital signs, including heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), and oxygen saturation (SPO2). Each vital sign is scored based on its deviation from the norm, and these scores are aggregated to calculate the overall NEWS score. A high NEWS score indicates a greater likelihood of health deterioration and the need for medical attention.[2]

Despite their utility, these Early Warning Systems have limitations. They lack the ability to capture the correlation between parameters, as each parameter's score is evaluated independently and then simply summed.[6] Moreover, as reflected in the CTAS guide, vital signs are not the only indicators of a patient's health condition. Symptoms are also crucial and may affect the importance of each vital sign.[5] For instance, an elevated respiratory rate could be lower risk for a patient experiencing anxiety, but it could indicate a serious condition if the patient is in respiratory distress.

Related work

In medicine, machine learning has shown promise in predicting specific diseases, such as cardiac arrest,[1,7] and improving triaging processes.[8,9] Noteworthy studies include J. Kwon et al.'s proposal of a 3-layer Recurrent Neural Network with Long Short-Term Memory units, which generates a Deep-learning Early Warning Score (DEWS). Impressively, DEWS outperformed conventional MEWS in all metrics, exhibiting higher sensitivity and fewer alarms.[1] Similarly, M. Ong et al. employed an undisclosed 'black box' algorithm that surpassed MEWS in accuracy.[6] However, none of these studies uses chief complaint in their features.

On the front with chief complaint, telemedicine has traditionally relied on nurses interpreting chief complaints over telephone calls.[3] In much later period, Semigran et al. assessed the accuracy of various mobile symptom checkers, which aims to provide triage advice based on patient-entered symptoms. However, they found that these symptom checkers had deficits in both triage and diagnosis.[4] It is within

Table 1. Data inclusion criteria.

Data features	Range	Unit
Heart rate	0–300	bpm
Blood oxygen saturation	0–100	%
Systolic blood pressure (SBP)	0–300, higher than DBP	mmHg
Diastolic blood pressure (DBP)	0–300, lower than SBP	mmHg
Respiratory rate	0–100	bpm
Body temperature	20–50	°C
Numeric pain score	0–10	
Chief complaints	All available chief complaints in the dataset	

recent years that we see the bloom in natural language processing technology, with the advent of chatbots. R. Luo et al. presented the "BioGPT," which exhibit significant potential for generating biomedical text, as it reached 78.2% accuracy on PubMedQA question answering task.[10] A more recent study by Pasli et al. investigates the effectiveness of ChatGPT in performing real-time patient triage using voice commands, and surprisingly outperformed human triage personnel and shows great promise as a clinical decision-support tool.[11]

To our knowledge, there are no similar studies that have leveraged a combination of chief complaints and numerical vital signs, aiming to continuously monitor patients.

Methods

The dataset used in this study is the MIMIC-IV-ED dataset, which is a public database that contains multiple data tables, collected in the ED of Beth Isreal Deaconess Medical Center.[12] The authors have acquired credential access on PhysioNet. This study primarily focuses on the triage table, which contains information collected from the patient at the time of triage, including chief complaints and one set of vital signs measurement including body temperature, heart rate, respiratory rate, systolic and diastolic blood pressure. When relevant, a subjective pain level ranging from 0 to 10 is also reported by the patient. The table also contains an acuity value assigned by the nurses, which serves as label for training a triage assisting algorithm.[12]

The data undergoes the following preprocessing steps. First, overly unrealistic data are filtered out, for example, body temperature higher than 5000 Celsius. Table 1 shows the inclusion range of data. Second, a NEWS score is calculated for each data, without the criteria for consciousness or supplemental oxygen since those are not available in the table. Third, all the vital signs are scaled to unit variance, while all text is tokenized. Lastly, the data is split into 70% training data, 15% validation data, and 15% testing data. After preprocessing, 391533 datapoints remain. 274073 datapoints are used in training, 58730 datapoints are used in validation and another 58730 datapoints are used in testing.

As a baseline, a model is trained solely based on the vital signs. The model uses the mentioned vital signs (including pain), NEWS score and the number of abnormal vital signs as features and attempts to predict acuity level. Three multiclass classification models were compared during the building of the baseline model: a random forest classifier, a gradient boosting classifier, and an artificial neural network (ANN).

The ANN first processes text input and numerical input separately. The text input passes through an embedding layer, then a transformer unit. Meanwhile, the numerical input passes through one dense layer with Rectified Linear Unit (ReLU) activation. Then, both inputs are concatenated into one vector, and finally pass through a classification layer. The architecture is shown in Figure 1.

The transformer encoder is used to process text in the ANN. The unit contains multiple layers. The multi-head attention mechanism computes weights between each pair of elements in the input sequence, assigning importance to different words based on their relationships with each other.[13,14] Then the feed-forward neural network allows the model to capture complex patterns in the text. Transformer units have been widely used in natural language models. The structure of transformer encoder unit is shown in Figure 2.[14]

For the ANN, cross-entropy loss is used as loss function, as the models deal with multiclass classification. All models are mainly evaluated based on their accuracy, which is the percent of correct predictions out of all predictions. To show the model performance at individual levels, precision and recall are also calculated. Precision measures how many predictions are correct in that level, while recall measures how many correct predictions the algorithm can find within that level. A receiver operating characteristic curve (ROC) and a precision–recall curve (PRC) are also plotted for the parallel model.

All the hyperparameters in the model are tuned with the hyperband tuner, which is a novel hyperparameter tuning algorithm that optimizes a pure-exploration non-stochastic infinite-armed bandit problem.[15] Informally, it is a trial-and-error algorithm that tests out different combinations of hyperparameters and seeing which ones work best. However, instead of trying every possible combination, it quickly narrows down a set of hyperparameters that performs the best by training them for a short time, and gradually focusing on the more promising models.[15]

All the models are trained with batch size of 32. All the activation functions except for the output classification

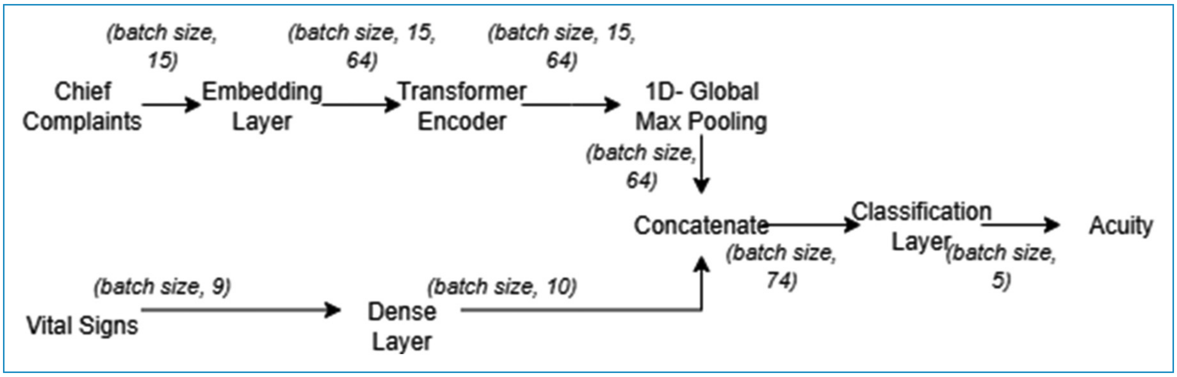


Figure 1. Architecture of the ANN, each block represents one single layer.

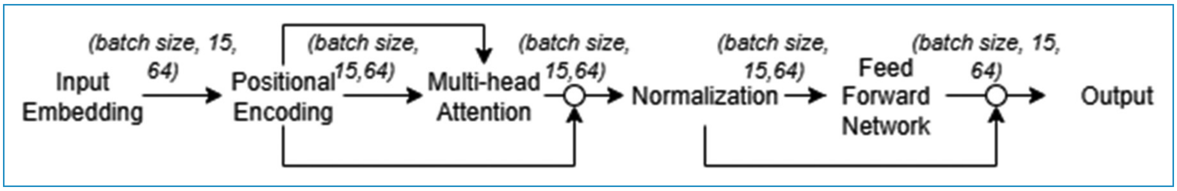


Figure 2. Architecture of the transformer encoder unit.[14]

layer are ReLu. The classification output layer uses SoftMax activation. The model is built using Python 3.11. We utilized scikit-learn.StandardScaler to perform normalization, keras.preprocessing.text.Tokenizer to tokenize text, tensorflow.keras to construct the neural network, numpy for numeric operations, pandas for dataframe utilities, shap for SHAP explanations, imblearn for extra analysis with imbalanced dataset, and matplotlib with seaborn for plotting functions. All trainings are performed locally on an encrypted laptop, with 32GB RAM, 13th Gen Intel Core i7-13700HX CPU, and NVIDIA GeForce RTX 4080 GPU.

To analyze the importance of each feature for the model, we adapt the SHapley Additive exPlanations (SHAP), which is a method based on game theory approaches, that assigns an importance value for each feature in each model predictions. The aggregate value of such values can then provide an overview of overall importance of each feature.[16] This is a model-agnostic method, making it fitting for complex deep neural networks, allowing us to see what's truly important in these model predictions.

From conceptualization stage to completion, this study lasted for 1 year, conducted in University of British Columbia (UBC). The study involves personnel from the UBC and Vancouver General Hospital, but no data are gathered from the hospital. By its very nature, this study is retrospective, and serves as a proof of concept, that models combining chief complaints and vital signs are promising in ED monitoring. This study neither claims nor aims to create a clinically deployable model at this stage.

Results

All the models have reached similar levels of accuracy around 59%. Also, something notable is that even during hyperparameter tuning, the accuracy did not significantly change. This indicates that the type and structure of these models do not have strong impact on the prediction accuracy. Rather, it is the data or features themselves that did not correlate strongly with the desired output. This is also evident during the triaging process, as nurses do not solely triage the patients based on their vital signs.

As shown in Figure 3, by incorporating chief complaints to the model, the accuracy has increased to over 70%, achieving a significantly higher accuracy than the baseline model (59%), further indicating that the poor performance of the pure vital-sign-based baseline model is caused by the lack of features. Allowing the model to access both textual and numerical data gives it opportunity to find correlations between vital signs, symptoms, and acuity, thus increasing predictive power of the model.

We performed data filtering by selecting data with identical NEWS score and chief complaints. In each of these data groups, the acuity level with highest occurrence is kept, and all other data are discarded. The objective is to ensure that if the chief complaints are identical, and vital signs are roughly at the same level (same NEWS level), the acuity level should be the same. This filtering step removed 85231 data points, which consists of about 20% data points. We perceive a significant increase in accuracy. Figures 4 and 5 show the significant increase in area under ROC (AUROC) after

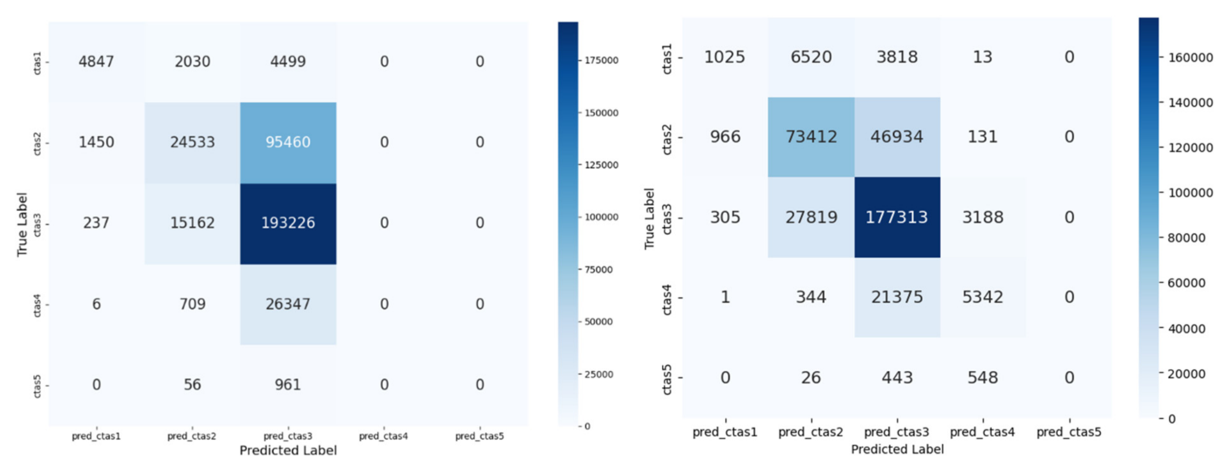


Figure 3. Comparison of confusion matrices of baseline model (left) and proposed model (right). Diagonal represents correct prediction, while data above the diagonal represent over-triaging, and below represents under-triaging.

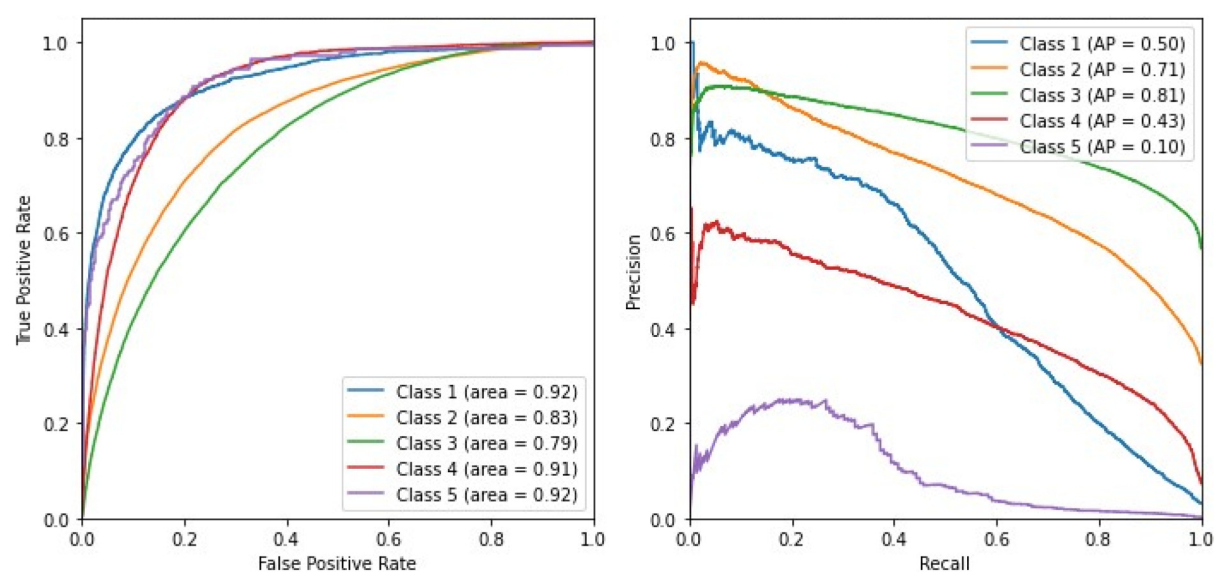


Figure 4. Receiver operating characteristic and precision–recall curve for the ANN.

performing data filtering. Figure 6 shows the increase in predictive power in most classes after filtering through the form of confusion matrix.

As seen on the confusion matrices, the classes are very imbalanced. To explore this, we use Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalance issue. SMOTE attempts to over-sample minority classes by finding its nearest neighbor, then creates a new synthetic sample on the linear segment connecting these data points.[17] The principle is to create synthetic samples of minority classes and essentially balance the class by adding these samples. However, introducing SMOTE created catastrophic effects on the model, as shown in Figure 7.

Lastly, we use the SHAP method to analyze the model both pre and post filtering. As shown in Figure 8, in unfiltered data, chief complaints (word features) have almost complete dominance over vital signs, as the vital signs have shown very negligible SHAP value compared to the word features. In contrast, after the inconsistent data are filtered out, we begin to see features such as NEWS score, number of abnormal vital signs, and heart rate begin to show importance in prediction.

Discussion

A notable phenomenon in the baseline models is that at true acuity level 2 and 4, there are more predictions at acuity 3

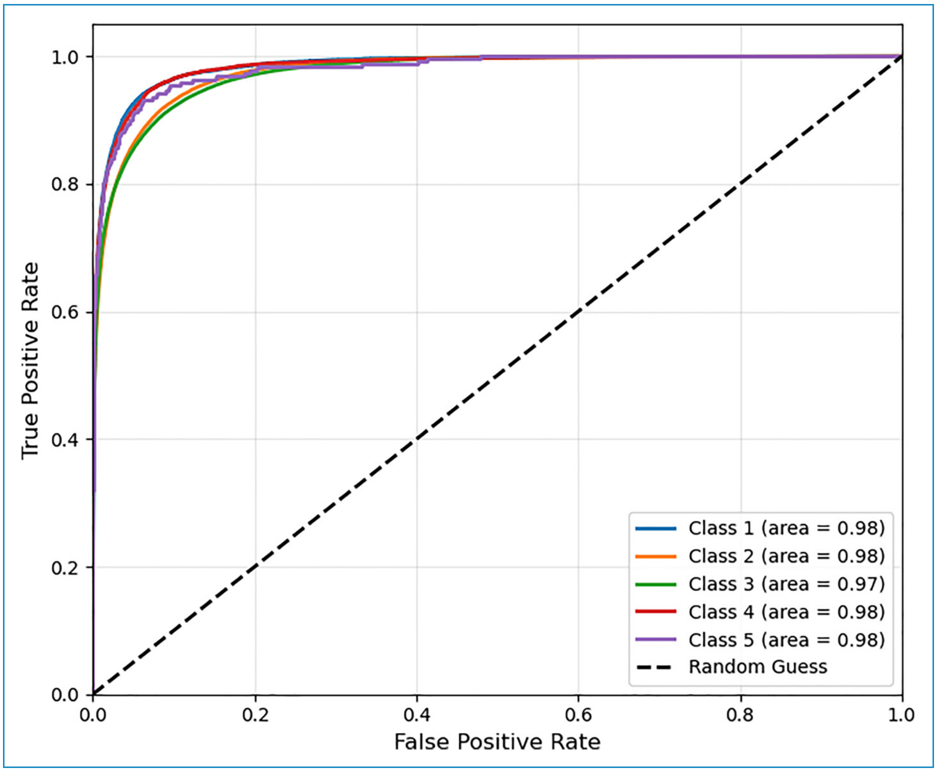
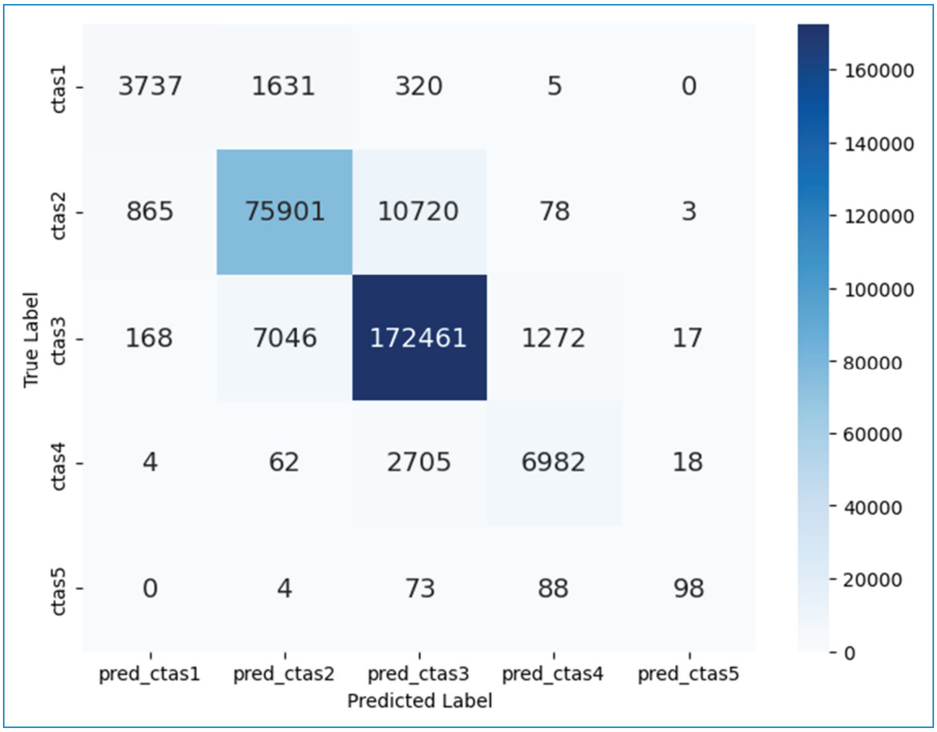


Figure 5. Receiver operating characteristic for the ANN post-filtering.



True Label	pred_ctas1	pred_ctas2	pred_ctas3	pred_ctas4	pred_ctas5
ctas1	3737	1631	320	5	0
ctas2	865	75901	10720	78	3
ctas3	168	7046	172461	1272	17
ctas4	4	62	2705	6982	18
ctas5	0	4	73	88	98

Figure 6. Confusion matrix for the ANN post-filtering. Diagonal represents correct prediction, while data above the diagonal represent over-triaging, and below represents under-triaging.

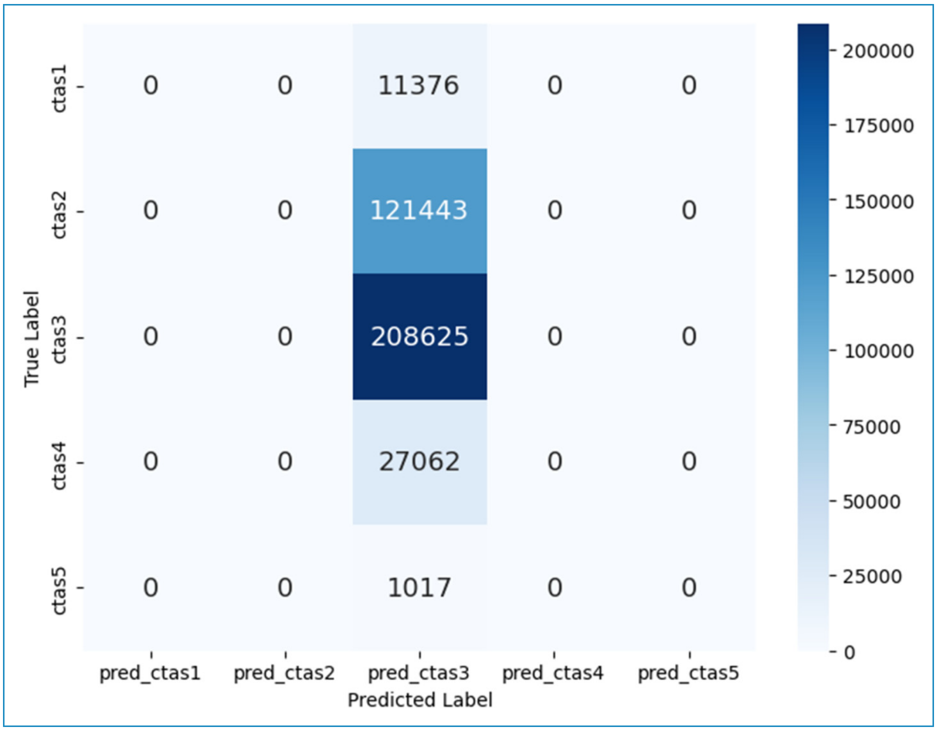


Figure 7. Confusion matrix for the ANN with SMOTE, trained on unfiltered data.

than correct predictions. This indicates an overlap between these acuity levels. There are many possible reasons. First, patient at these levels could have a wide range of vital signs. For example, altered mental status or confusion is commonly triaged as acuity 2, but their vital signs could be like normal people, as their mental status may not have affected their body functions. In contrast, patients with hypertension would have higher than normal blood pressure, but no immediate health risk so they could receive an acuity level of 4 when no symptoms are present.[5] Lastly, according to the CTAS training manual, there may be subconscious pressure to "down-triage" patients when ED is overcrowding. For example, a patient normally with acuity 2 may be assigned acuity 3 when ED is full, while nurses feel it unacceptable to assign level 2 patients in the waiting room.[18]

Nonetheless, by including chief complaints in features, we see a noticeable increase in accuracy (around 10% increase). The baseline model appears extremely biased as its predictions are almost exclusively limited to CTAS 2 and 3, while never once predicts a patient at CTAS 4 or 5. This indicates that the model has simply learned that CTAS 2 and 3 are most common, and is guessing them to get the highest accuracy. On the other hand, the ANN now actively predicts CTAS 4 and 5 based on some minority chief complaints, while having greatly improved performance in CTAS 2 prediction. While errors still exist, the mistakes are much more plausible, as the confusion is now primarily between adjacent classes (e.g. CTAS 2 vs CTAS 3, CTAS 3 vs CTAS 4), which also mirrors real-world ambiguity of triage.

It is also noticed that the sample size of each class is very imbalanced in the dataset. Roughly 90% of the data consists of acuity 2 and 3, while the rest consists of only about 10% of the data. This imbalance leads to some nuances in the model and its interpretation. First, as shown in Figure 4, the model has good AUROC across acuity 1, 4, and 5, but slightly lower at acuity 2 and 3. However, the average precision (AP) is much lower across all levels, especially for acuity 1, 4, and 5. This is also a phenomenon of an imbalanced dataset, as the baseline (lowest possible value) for AUROC is 0.5, and it is increased by both true positives and true negatives, while AP is only increased by true positives. When a model is imbalanced, the positive case for some classes is much less frequent than negative cases, leading to a lower baseline AP value.[19]

We also inspect the learning curve of the model, which reveals deeper issue than just imbalanced dataset.

As observed from the learning curves in Figure 9, there is a generalization gap in the model. Even though the model accuracy consistently increases while model loss decreases during training, the validation accuracy and loss stayed at roughly the same value, a phenomenon commonly known as "overfitting." While traditional methods to combat overfitting involve regularization or reducing model complexity, we hypothesize that the root cause in our problem is not excessive model complexity, but rather the intrinsic properties of the data, causing unrepresentative validation set.[20] Specifically, the severe class imbalance and the difficulty of separating the minority classes from the majority classes using available features.

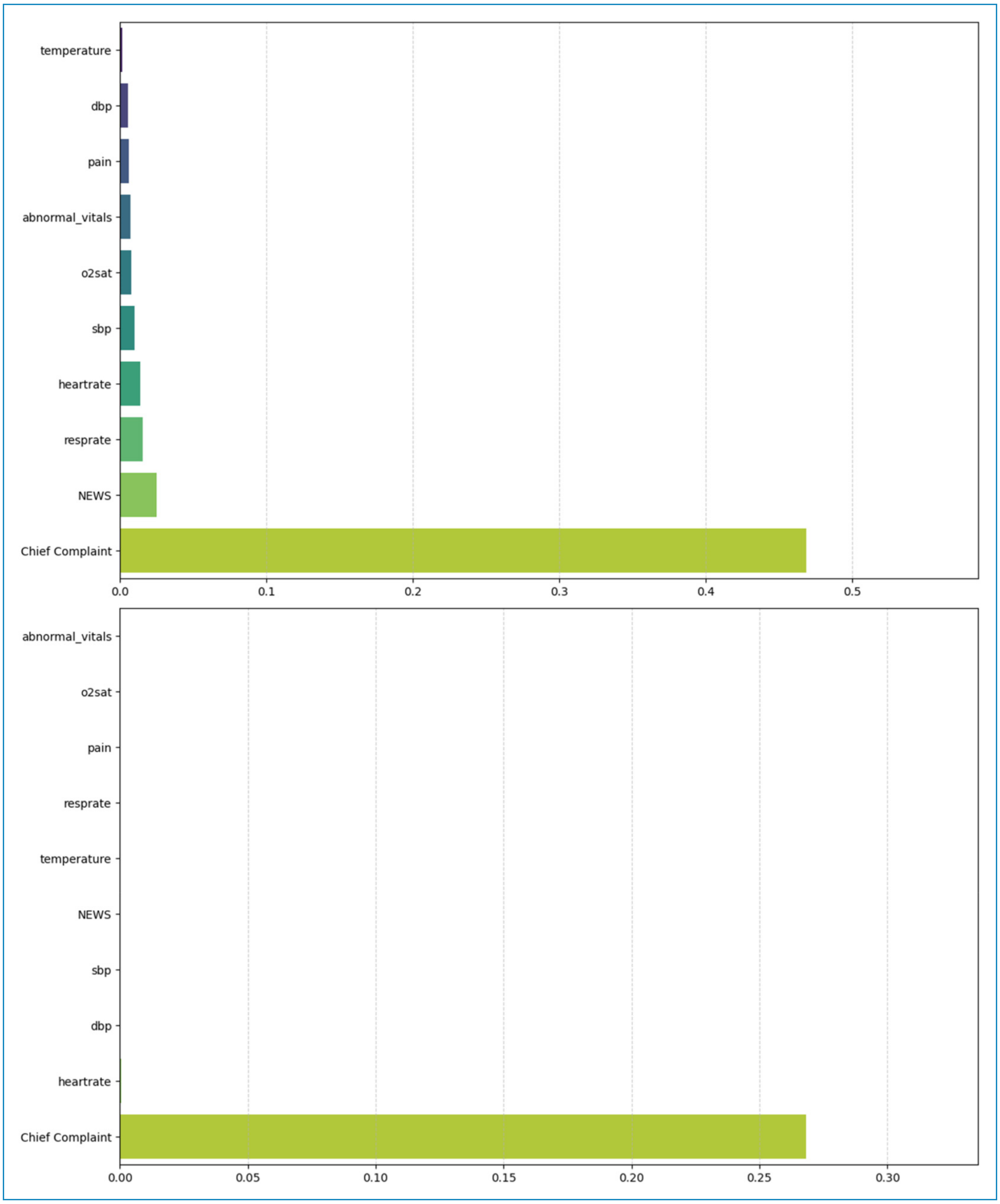


Figure 8. Comparison of mean absolute SHAP value of model with filtered (top) and unfiltered (bottom) data. *X*-axis represents SHAP value, which demonstrates the feature's impact on model output.

To investigate the impact of class imbalance, we implemented SMOTE. Standardly, SMOTE is expected to improve performance on minority classes.[17] However, in our case, its application led to a catastrophic failure in model training, causing the model to exclusively predict the majority class, CTAS 3. This outcome indicates that the feature space of the minority classes heavily overlaps with that of the majority classes. By creating synthetic samples between existing minority points, SMOTE inadvertently generated noisy and contradictory training examples located within dense majority-class clusters. This "noise amplification" made it impossible for the model to find a coherent decision boundary. Meanwhile, the instability shown in the validation curve can be interpreted as a milder symptom of this same root problem, as the model struggles to distinguish the islands of minority classes that are difficult to distinguish in the sea of majority classes.

Upon further investigation in the dataset, it is found that in some cases, even if the input is similar, the outputs could be completely different. One possible reason is that during

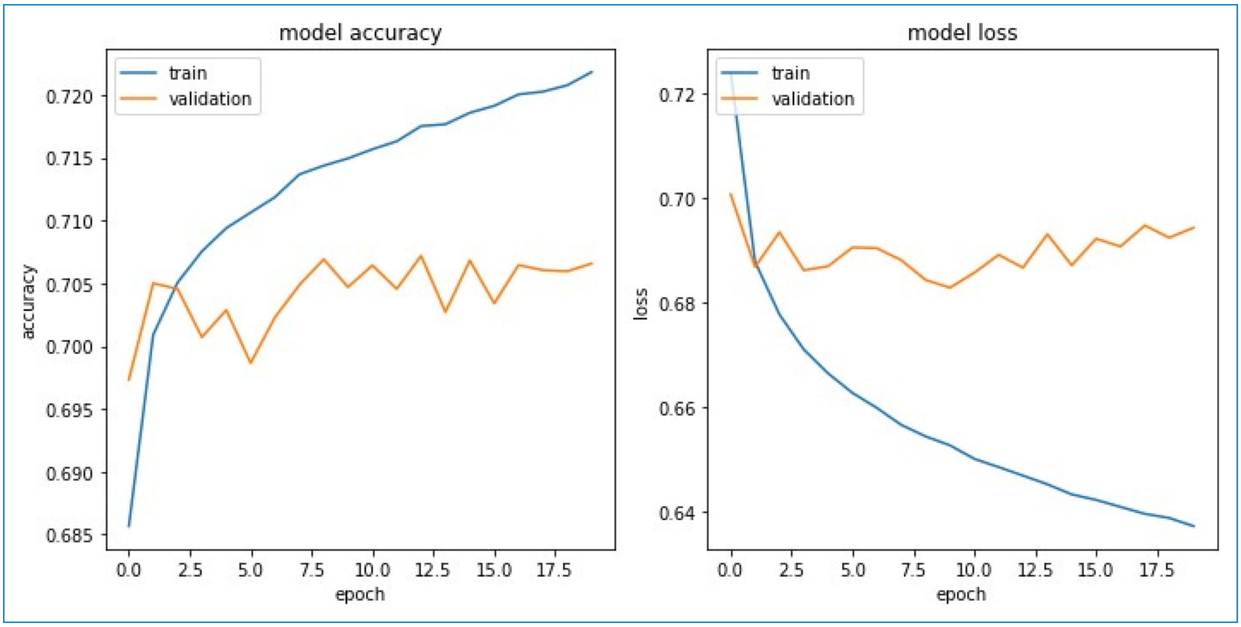


Figure 9. Learning curve for the model trained on unfiltered data.

triage, nurses have access to other information that was not recorded in the dataset. For example, crowded EDs can give subconscious pressure for nurses to down-triage patients, as mentioned previously, which could lead to similar input features but different acuity level. Another example is that nurses can see the patients and examine their conditions through details such as medical history, visible injuries or tone of speaking, none of which are recorded directly in dataset. Lastly, even if triage nurses are highly experienced, they still could make mistakes in triaging from time to time, or have inconsistency when triaging, as in many cases they are guided by experience, rather than algorithmic processing of data. This is further highlighted after we applied data filtering, removing about 20% of data but resulting in a significant accuracy improvement, proving that the original unfiltered dataset is indeed noisy.

Furthermore, when we apply SHAP method to explain the feature importance, we see a stark difference between filtered and unfiltered data. For the unfiltered data, chief complaints have complete reign over the acuity, while the vital signs have shown very negligible SHAP value. This phenomenon indicates that for that dataset, the correlation between vital signs and acuity is weak. This is not to claim that vital signs are unimportant features, rather it is to point out that in the unfiltered dataset, there are too many cases where similar vital signs are assigned different acuity values by nurses, also highlighting that patient's level of acuity relies on factors way beyond vital signs. In contrast, after the inconsistent data are filtered out, we begin to see features such as NEWS score, number of abnormal vital signs, and heart rate begin to show importance in prediction. However, it is still noticeable that the first three words in chief complaints have great importance. This further highlights the claim of this study: to develop a continuous monitoring algorithm in the ED, chief complaints must be used in conjunction with each other, to provide accurate predictions.

In summary, among all models, the baseline model has worst performance in terms of accuracy, while our proposed model reached higher accuracy, likely due to its ability to incorporate chief complaints. The accuracy has plateaued around 70% due to the inconsistency of data or lack of input features, as evident in the significant accuracy increase after additional data filtering. The summary of model performances can be viewed in Table 2.

Practical implications

This study proposes a novel approach to monitoring in the ED, which is to utilize chief complaints in conjunction with vital signs, to predict the acuity values. In the current practices, the CTAS level is only assigned at the start at triage. However, during the ED stay, nurses must rely on either early warning systems such as MEWS or NEWS, or some threshold values to provide algorithmic alarms, based on vital signs, or rely on repeated assessments and their own intuition.[2,5] Using an approach similar to this study would allow an algorithm to more accurately determine the patient's acuity over time, constantly providing risk stratifications, allowing nurses to focus on patients with higher risks at the time. When paired with continuous monitoring devices, it truly has potential to improve the workflow in the ED. However, by no means this type of model is meant to replace the role of triage nurses. Rather, it intends to serve as a triage assistant, providing nurses with the tool to more rapidly address a large number

Table 2. Performance of all evaluated models.

Model	Accuracy			Model type
	Training	Validation	Testing	
Baseline feed-forward network	0.6034	0.6009	0.6014	ANN
Baseline gradient boosting model	0.5922	N.A	0.5941	Gradient boosting
Baseline random forest model	0.5908	N.A	0.5932	Random forest
Baseline feed-forward network with filtering	0.6785	0.6805	0.6784	ANN
ANN model with chief complaints	0.7265	0.7069	0.7065	ANN with a transformer encoder
Parallel model with chief complaints and filtered data	0.9065	0.8781	0.8803	ANN with a transformer encoder

of patients. Lastly, even if this particular model does not have enough predictive power or enough validation to be used in a real clinical environment, it shows the importance of such multimodal approaches, allowing future studies to try incorporating chief complaints into their model development.

Limitations and future work

The models utilized in this study exhibit several limitations. Firstly, due to the absence of time series data, the vital signs used in the model relies solely on those from initial triaging. This limitation arises because the time series data within MIMICS-IV-ED lacks labels, necessitating manual labeling for supervised learning, while unsupervised learning is unsuitable for the triaging task. Moreover, the reliance on labels provided by "expert clinicians" introduces the risk of inaccurately triaged cases, potentially compromising the decision-making process—a scenario often described as "Garbage in, garbage out."[21] Additionally, the exclusive use of data sourced from the MIMIC-IV-ED database introduces the potential for inherent bias towards this dataset, which may result in statistical disparities when extrapolated to real ED scenarios. Lastly, it is imperative to recognize the retrospective nature of this study, implying that the model's performance might vary when implemented in clinical settings.[6]

To address these limitations, several avenues for improvement are worth exploring. Firstly, supplementing the MIMIC-IV-ED dataset with data collected directly from hospitals could diversify the dataset. Furthermore, integrating time series data into the model is essential. As the MIMIC-IV-ED lacks labels for time series data, it might be necessary to find or create datasets with clear definition of clinical deterioration. Lastly, to validate the accuracy of the model, clinical trials should be conducted to evaluate its performance in real-life situations.

Conclusion

In this study, data collected from MIMIC-IV-ED is used to train several models. Among these models, ones that included chief complaints as input had significantly higher accuracy (72.65%) than the baseline model (60.34%) which only uses numerical data, supporting the claim that the inability to incorporate chief complaints leads to the inaccuracy of EWS. Model accuracy increased further (90.65%) when the inconsistent data are removed, further highlighting the potential of this algorithm when human inconsistency during triaging is minimal. The findings in this work have significant real-world application, having the potential to enhance patient monitoring and early detection of deterioration in ED waiting rooms, ultimately improving patient outcomes. Future research should focus on incorporating more diverse datasets and utilizing time series data to further refine and validate these models in clinical settings, paving the way for AI-driven advancements in emergency medicine.

Acknowledgements

The authors would like to thank the TxMED team for providing guidance and support and reviewing the manuscripts. The authors would also like to thank Vancouver Coastal Health and Rogers Communications for making this study possible.

ORCID iDs

Zongrun Li https://orcid.org/0009-0005-9566-2318
Julie Lockington https://orcid.org/0009-0009-4346-5127
Samya Torres https://orcid.org/0009-0003-5869-4375
Nooshin Jafari https://orcid.org/0000-0001-8107-7779
Michael Lim https://orcid.org/0009-0007-8669-0418

Ethical considerations

This study only uses data from a publicly available database, does not involve human participants. Thus, ethics approval was not required.

Author contributions

Zongrun Li: data curation, formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing. Julie Lockington: investigation, methodology, writing—review and editing. Samya Torres: project administration, writing—review and editing. Nooshin Jafari: conceptualization, writing—review and editing. Michael Lim: conceptualization, data curation, supervision, writing—review and editing. Dragan Andjelic: formal analysis, software, writing—review and editing. Edmond Cretu: conceptualization, funding acquisition, project administration, supervision, writing—review and editing. Kendall Ho: conceptualization, funding acquisition, project administration, supervision, writing—review and editing. Bhushan Gopaluni: conceptualization, funding acquisition, project administration, supervision, writing—review and editing.

Funding

The authors disclosed the following financial support for the research, authorship, and/or publication of this article: This work was supported by Mitacs Accelerate [grant number IT42089] and Rogers Communications.

Declaration of conflicting interest

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

Supplemental material

Supplemental material for this article is available online.

References

1. Kwon J, et al. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc* 2018; 7(XX): 1–10.
2. Royal College of Physicians. National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS. 2017. https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2
3. Dale J, et al. Primary care: nurse-led telephone triage and advice out-of-hours. *Nurs Stand* 1998; 12: 39–43.
4. Semigran H, et al. Evaluation of symptom checkers for self diagnosis and triage: audit study. *Br Med J* 2015; 351: h3480.
5. Ministry of Health and Long-Term Care. Prehospital CTAS paramedic guide (Version 2.0). Ontario, Canada. n.d.
6. Muralitharan S. Machine learning-based early warning systems for clinical deterioration: systematic scoping review. *J Med Internet Res* 2021; 23: e25187.
7. Ong M, Lee Ng CH, Goh K, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Crit Care* 2012; 16: R108.
8. Raita Y. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019; 23: 1–10.
9. Ivanov O. Improving ED emergency severity index acuity assignment using machine learning and clinical natural language processing. *J Emerg Nurs* 2021; 47: 1–10.
10. Luo R,, Sun L, Xie Y, et al. BioGPT: generative pretrained transformer for biomedical text generation and mining. *Brief Bioinform* 2022; 23: 1–10.
11. Pasli S. ChatGPT-supported patient triage with voice commands in the emergency department: a prospective multicenter study. *Am J Emerg Med* 2024; 94: 63–70.
12. Johnson A. MIMIC-IV-ED (Version 2.2) [Dataset]. PhysioNet. 2023.
13. Chollet F. *Deep learning with Python*. Shelter Island, NY: Manning Publications, 2018.
14. Vaswani A, et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp.6000–6010.
15. Li L and Jamieson K. Hyperband: a novel bandit-based approach to hyperparameter optimization. *J Mach Learn Res* 2018; 18: 1–52.
16. Scott L and Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp.4768–4777.
17. Chawla N, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357.
18. Canadian Association of Emergency Physicians. Canadian Triage and Acuity Scale (CTAS) participant's manual. 2012.
19. Draelos R. Measuring performance: AUPRC and average precision. Glass Box Medicine. 2019, March 2. https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/
20. Brownlee J. How to use learning curves to diagnose machine learning model performance. *Mach Learn Mastery* 2019; 1. https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/
21. Mueller B. Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute Med Surg* 2022; 9: e735.