



Pitfalls of developing medical artificial intelligence: prediction of inpatient's unexpected deterioration using the hidden implication of vital parameters

Y.H. Chou^{1, 2}, W.J. Liaw^{3, 4}, C.H. Yu^{5, 6}, D.Y. Yan⁷, I.C. Lin⁷, W.C. Shen⁷

¹Department of Medical Imaging and Radiological Sciences, Chung Shan Medical University, ²Department of Radiation Oncology, ³Department of Medical Quality, Chung Shan Medical University Hospital, ⁴College of Medicine, Chung Shan Medical University, ⁵Public Relations Department, Chung Shan Medical University Hospital, ⁶School of Nursing, College of Medicine, Chung Shan Medical University, ⁷Artificial Intelligence Center, Chung Shan Medical University Hospital, Taichung, Taiwan

Introduction

- To demonstrate the challenges encountered in creating an artificial intelligence (AI) to predict unexpected deterioration of inpatients who stayed in the general wards, primarily focusing on the impact of data leakage on the model's performance.

◆What is Data Leakage?

As shown in Figure 1, the initial purpose is to develop an AI model that can be used to detect whether the medical lines were installed on a patient from the input chest X-ray (CXR) image.

- The trained AI model exhibited a good prediction performance in detection. The heatmap was then adopted to show the key features used in the prediction to understand the reasons for judgments made by AI.
- Unfortunately, the word "Portable" marked on CXR was critical in identifying a patient with the medical lines.
- The decision made by the AI model is not based on the identification of the medical lines but on the hidden implication provided by the annotations on the CXR.

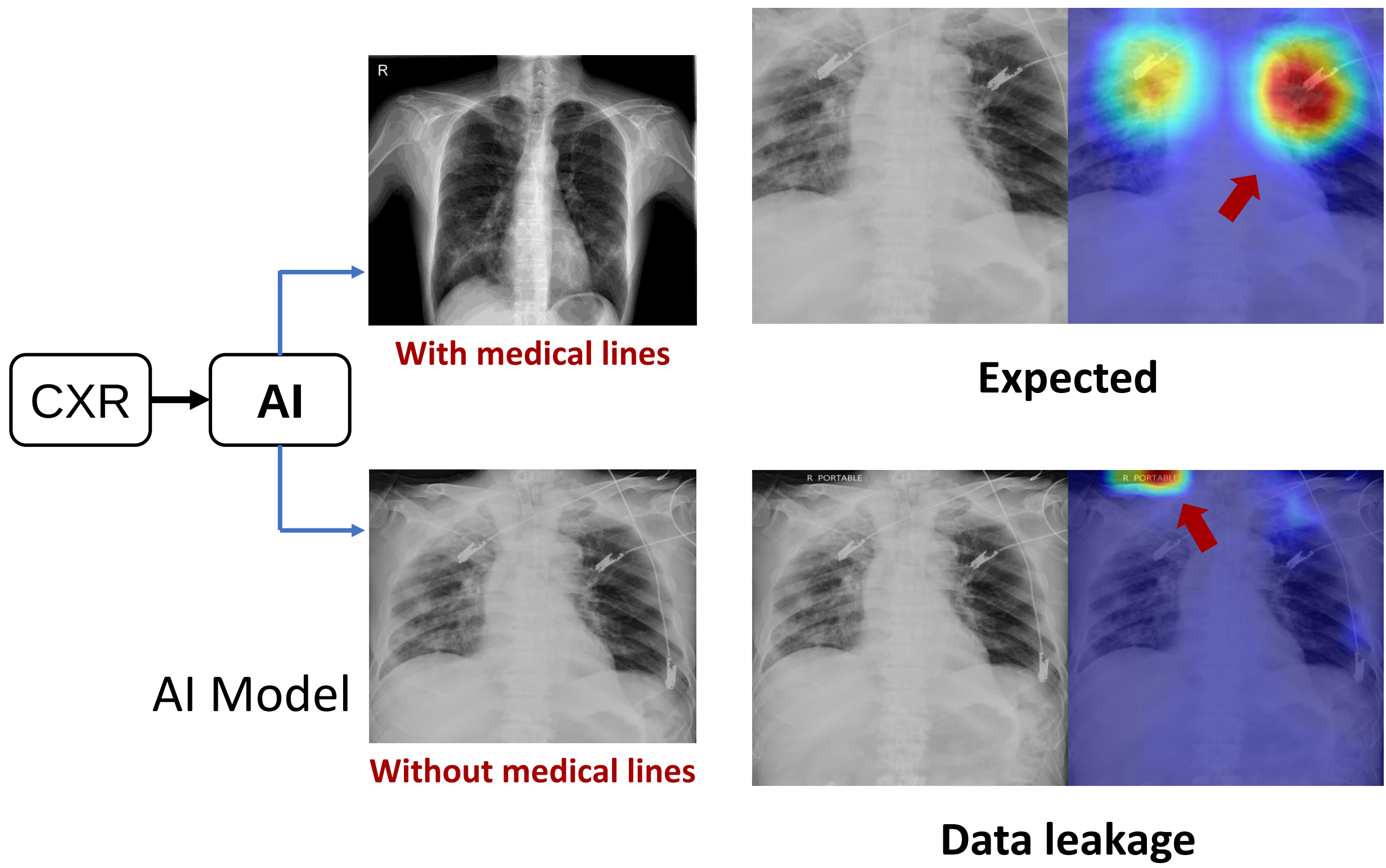


Figure 1. An example of data leakage in the training data

Methods

- The Research Ethics Committee of Chung Shan Medical University Hospital approved the study (No. CS1-22196).
- Adult inpatients of our hospital in 2021 were included, and their values for eight vital parameters (VPs) were collected.
- The unexpected deterioration was defined by unplanned intensive care unit (ICU) admission.

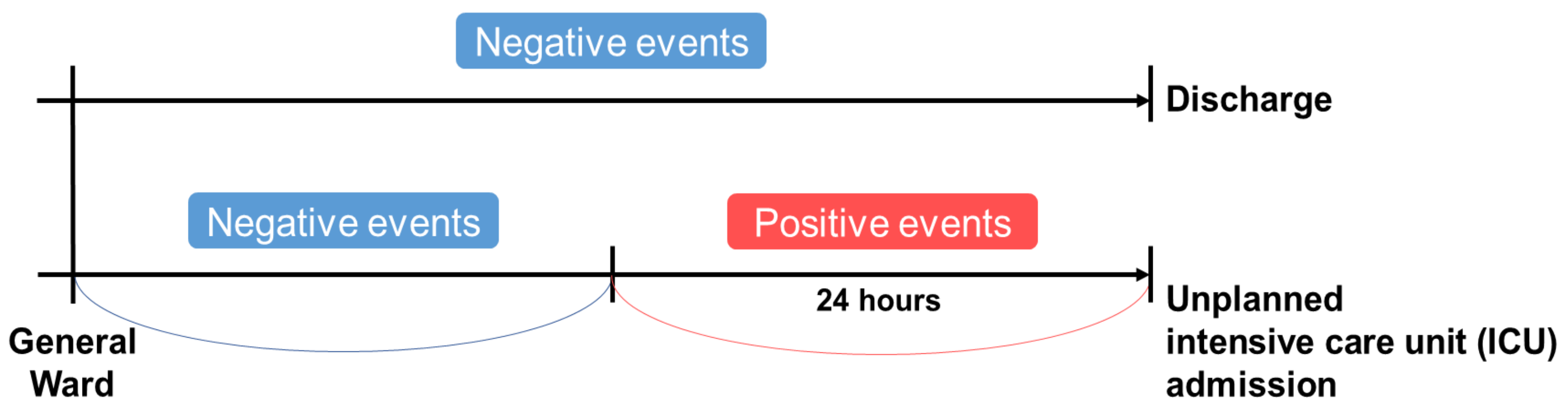


Figure 2. The definition of the positive events in this study.

- The assessment of certain VPs may be omitted for general ward inpatients, particularly for GCS and SpO₂.
- These omitted parameters were recorded as missing values in the training data.

Vital parameters (VPs)	AI ₄	AI ₅	AI ₈
Body temperature	⊙	⊙	⊙
Heart rate	⊙	⊙	⊙
Respiratory rate	⊙	⊙	⊙
Systolic blood pressure	⊙	⊙	⊙
Blood oxygen saturation (SpO ₂)		⊙	⊙
The change in consciousness (Glasgow coma scale, GCS)			
Eye			⊙
Verbal			⊙
Motor responses			⊙

Figure 3. Eight vital parameters (VPs) in three different AI models.

- Five-fold cross-validation was employed to verify the predictive performance of the three AI models.
- After five rounds, the predictive probabilities of a positive event in the five validation sets were merged.
- Finally, the area under the receiver operating characteristic curve (AUC) analysis was conducted to assess the merged results as the predictive performance of AI models.

Results

- A total of 27,579 inpatients with 539,245 times of VP assessments were included. About 0.5% (2,601) of these VP assessments were labeled positive events.
- Within 24 hours before an unplanned admission, GCS and SpO₂ measurements frequencies increased from 0.3 to 1.1 and from 0.3 to 4.8 times per day, respectively.
- The AUCs of AI₈, AI₅, and AI₄ for predicting unexpected deterioration within 24 hours were .914, .916, and .860, respectively.
- The pairwise comparison among the three models revealed that the significance levels were .382 between AI₈ and AI₅, <.0001 between AI₈ and AI₄, and <.0001 between AI₅ and AI₄.

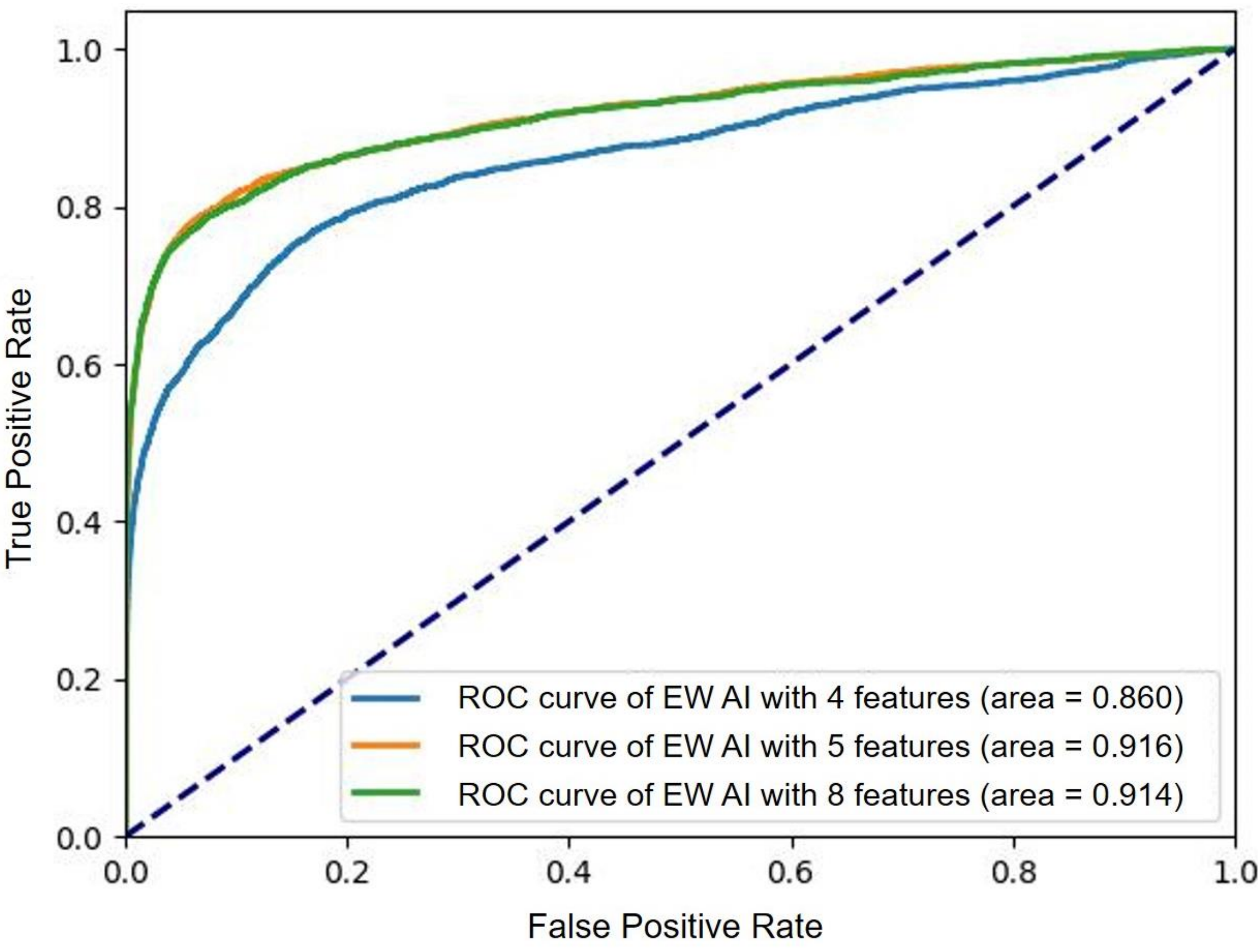


Figure 4. The AUCs of AI₈, AI₅, and AI₄ for predicting unexpected deterioration within 24 hours.

- In this study, the AI model determined whether an unexpected deterioration occurred based on the measurement frequency of the vital parameters rather than the parameters (Table 1).

Table 1. Frequency of vital sign measurement

	Daily vital parameters measurements					
	SBP	Pulse	Respiratory	BT	SpO ₂	GCS
Normal inpatients	3.3	3.3	3.2	3.2	0.3	0.3
Too early	4.3	4.3	4.1	4.0	1.8	0.6
Buffer time	6.7	6.8	6.5	5.8	4.8	1.1

Conclusion

The AI₈ and AI₅ perceived the subtle and indirect hints from the change of missing values and then adjusted the predicted probability to gain a better performance in training. However, the presence of data leakage in the training data may have introduced uncertainty in the performance of AI models, even though verified by cross-validation.