Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: A Machine Learning-Based Multimodal Approach

Hsin-Ying Lee, Po-Chih Kuo, Frank Qian, Chien-Hung Li, Jiun-Ruey Hu, Wan-Ting Hsu, Hong-Jie Jhou, Po-Huang Chen, Cho-Hao Lee, Chin-Hua Su, Po-Chun Liao, I-Ju Wu, Chien-Chang Lee

Submitted to: JMIR Medical Informatics
on: May 19, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: A Machine Learning-Based Multimodal Approach

Hsin-Ying Lee[1*]; Po-Chih Kuo[2*]; Frank Qian[3]; Chien-Hung Li[2]; Jiun-Ruey Hu[4]; Wan-Ting Hsu[5]; Hong-Jie Jhou[6]; Po-Huang Chen[7]; Cho-Hao Lee[8]; Chin-Hua Su[9]; Po-Chun Liao[9]; I-Ju Wu[9]; Chien-Chang Lee[10]

[1]Department of Medicine, College of Medicine, National Taiwan University Taipei TW
[2]Department of Computer Science, National Tsing Hua University Hsinchu TW
[3]Department of Medicine, Beth Israel Deaconess Medical Center Boston US
[4]Department of Internal Medicine, Yale School of Medicine New Haven US
[5]Department of Epidemiology, Harvard T.H. Chan School of Public Health Boston US
[6]Department of Neurology, Changhua Christian Hospital Changhua TW
[7]Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center Taipei TW
[8]Division of Hematology and Oncology Medicine, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center Taipei TW
[9]Department of Emergency Medicine, National Taiwan University Hospital Taipei TW
[10]Department of Emergency Medicine, National Taiwan University Hospital Taipei City TW
[*]these authors contributed equally

Corresponding Author:
Chien-Chang Lee
Department of Emergency Medicine, National Taiwan University Hospital
No.7,8, Zhongshan S. RD., Zhongzheng District
Taipei City
TW

Abstract

Background: Early identification of impending in-hospital cardiac arrest (IHCA) improves clinical outcomes but remains elusive for practicing clinicians.

Objective: We aimed to develop a multimodal machine learning algorithm based on ensemble techniques to predict the occurrence of IHCA.

Methods: Our model was developed by the MIMIC-IV database and validated in the eICU-CRD database. Baseline features consisting of patient demographics, presenting illness, and comorbidities were collected to train a Random Forest (RF) model. Next, vital signs were extracted to train a long short-term memory (LSTM) model. A Support Vector Machine (SVM) algorithm then stacked the results to form the final prediction model.

Results: Of 23,909 patients in the MIMIC-IV database and 10,049 patients in the eICU database, 452 and 85 patients had incident IHCA. Up to 13 hours in advance of an IHCA event, our algorithm maintained an area under the ROC curve above 0.78. Satisfactory results were also seen in validation from two external databases and comparison to existing warning systems.

Conclusions: Using only vital signs and information available in the electronic medical record, our model demonstrates it is possible to detect a trajectory of clinical deterioration up to 13 hours in advance. This predictive tool, which has undergone external validation, could forewarn and help clinicians identify patients in need of assessment to improve their overall prognosis.

(JMIR Preprints 19/05/2023:49142)
DOI: https://doi.org/10.2196/preprints.49142

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

~~Machine Learning-Based Multimodal Prediction of In-Hospital Cardiac Arrest in the ICU~~ Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: A Machine Learning-Based Multimodal Approach

[¶a]Hsin-Ying Lee MD, [¶b]Po-Chih Kuo PhD, [c,d]Frank Qian MD, MPH, [b]Chien-Hung Li MS, [e]Jiun-Ruey Hu MD, MPH, [f]Wan-Ting Hsu MS, [g]Hong-Jie Jhou MD, [h]Po-Huang Chen MD, [i]Cho-Hao Lee MD, [j]Chin-Hua Su MSc, [j]Po-Chun Liao MSc, [j]I-Ju Wu MD, [*j,k]Chien-Chang Lee MD, ScD

[a]Department of Medicine, College of Medicine, National Taiwan University, Taiwan
[b]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[c]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[d]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[e]Department of Internal Medicine, Yale School of Medicine, USA
[f]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[g]Department of Neurology, Changhua Christian Hospital, Changhua, Taiwan
[h]Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
[i]Division of Hematology and Oncology Medicine, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
[j]Department of Emergency Medicine, National Taiwan University Hospital, Taipei, Taiwan
[k]Center of Intelligent Healthcare, National Taiwan University Hospital, Taipei, Taiwan
[¶]The two authors contribute equally to this work

***Correspondence to:**
Chien-Chang Lee, MD, ScD (Harvard)
Professor and Attending Physician, Department of Emergency Medicine, National Taiwan University Hospital
Principal Investigator, Health Data Science Research Group, National Taiwan University Hospital
Deputy Director, Center of Intelligent Healthcare, National Taiwan University Hospital
No.7, Chung Shan S. Rd., Zhongzheng Dist., Taipei City 100, Taiwan.
Email: hit3transparency@gmail.com/cclee100@gmail.com
TEL: +886-2-2312-3456 ext. 63485

Running title: Prediction of In-Hospital Cardiac Arrest in the ICU
Word count for the text: ~~3688~~ 3778/10000
Word count for the abstract: 229/450
References: ~~30~~ 34/50
Main tables and figures: 5
Supplementary table: 1
Supplementary figures: 11

Abstracts

Background

Early identification of impending in-hospital cardiac arrest (IHCA) improves clinical outcomes but remains elusive for practicing clinicians.

Objective

We aimed to develop a multimodal machine learning algorithm based on ensemble techniques to predict the occurrence of IHCA.

Methods

Our model was developed by the MIMIC-IV database and validated in the eICU-CRD database. Baseline features consisting of patient demographics, presenting illness, and comorbidities were collected to train a Random Forest (RF) model. Next, vital signs were extracted to train a long short-term memory (LSTM) model. A Support Vector Machine (SVM) algorithm then stacked the results to form the final prediction model.

Results

Of 23,909 patients in the MIMIC-IV database and 10,049 patients in the eICU database, 452 and 85 patients had incident IHCA. At 13 hours in advance of an IHCA event, our algorithm had already demonstrated an area under the ROC (AUROC) curve of 0.85 (0.815-0.885) in the MIMIC-IV database. External validation over the eICU and National Taiwan University Hospital databases also presented satisfactory results, showing AUROCs of 0.81 and ~~0.934~~ 0.945, respectively.

Conclusions

Using only vital signs and information available in the electronic medical record, our model demonstrates it is possible to detect a trajectory of clinical deterioration up to 13 hours in advance. This predictive tool, which has undergone external validation, could forewarn and help clinicians identify patients in need of assessment to improve their overall prognosis.

Keywords: cardiac arrest; machine learning; intensive care; mortality

Introduction

Prognosis of IHCA is dismal as it represents the culmination of heterogeneous multi-organ dysfunction, with scarce treatments.[1] IHCA has an incidence of 9 to 10 per 1000 admissions and a mortality rate of 80-100%.[2] Therefore, clinical guidelines emphasize the urgent need for early identification of patients at risk for IHCA.[3] Early Warning Scores (EWS) were developed to facilitate early identification of impending clinical deterioration and trigger rapid interventions.[4] However, many traditional EWS are limited by considerable variation in discrimination in different populations and are often not sufficiently sensitive.[5]

Recent research indicates that the implementation of the Electronic Cardiac Arrest Risk Triage (eCART) score has significantly decreased the incidence of IHCA at UChicago Medicine.[6] However, the inclusion of laboratory data in eCART substantially diminishes the practicality and immediacy of this scoring system. Moreover, other studies have reported that calculating the Modified Early Warning Score (MEWS) system 0.5 hours before a cardiac arrest can significantly increase the survival-to-discharge rate in patients experiencing IHCA.[7] Nonetheless, a 0.5-hour lead time is often insufficient for a prompt reaction during a patient's rapid deterioration. Given the continuously generated real-time information, such as vital signs, a time-varying model could be constructed for more timely and early identification of IHCA.

The aim of our study was to develop an RNN-based model using the electronic health records (EHRs) of a single tertiary medical center to predict incident IHCA. We hypothesized that variation in physiological parameters, evaluated in the context of known comorbidities, could help to predict incident cardiac arrest. We also aimed to validate the model in an independent cohort and compare it to a previous scoring system.

Methods

Data source

Predictive models were developed using the Multiparameter Intelligent Monitoring of Intensive Care (MIMIC)-IV v0.4 database and were externally validated using the electronic ICU Collaborative Research Database (eICU-CRD) v2.0.[8,9] Pre-existing institutional review board (IRB) approval was waived given the de-identified nature of this public data set. (Massachusetts Institute of Technology, No. 0403000206; Beth Israel Deaconess Medical Center, 2001-P-001699/14).[10] One author who completed the Collaborative Institutional Training Initiative examination (certificate number: 57186438 for author H. J. Jhou) obtained access to the database and performed the data extraction. To assess the performance of our model in practical applications, we collected clinical data from the electronic medical records of National Taiwan University Hospital, spanning from 2008 to 2018 ~~2018 to 2019~~. Given the retrospective study design, the Research Ethics Committee of the National Taiwan University Hospital approved this study (project approval number 202206108RINB) and waived the requirement for obtaining informed consent. To decrease patient heterogeneity and feature variability, we applied the same inclusion criteria and data processing workflow to the three databases. We extracted data on patients aged over 20 who were hospitalized in intensive care units (ICUs) for at least 24 hours. Patients were excluded if they were encoded with a deceased status but without a IHCA labeling defined as below. We employed 5-fold cross-validation in our training cohort, randomly dividing the dataset into five equally-sized subsets. Four of these folds (80% of the MIMIC-IV cohort) were used for training, while the remaining fold (20% of the MIMIC-IV cohort) was reserved for internal validation. Performance metrics were recorded for each iteration, resulting in five distinct performance scores. These scores were then averaged to derive a singular, more robust performance estimate for the model. Finally, external validation was performed on the entire eICU-CRD cohort.

Disease outcome ascertainment

In the MIMIC-IV cohort, patients were marked with IHCA if they were either (i) labeled with time-stamped database-specific procedure code: 225466 (Cardiac arrest) or (ii) diagnosed with International Classification of Diseases, 9th revision, Procedure Coding System (ICD-9-PCS) codes: 9960 (Cardiopulmonary resuscitation, not otherwise specified). Although the MIMIC-IV database contained both ICD-9 and ICD-10 codes, we failed to convert ICD-9-PCS code: 9960 to ICD-10-PCS code, as the most approximately equivalent indicated code 5A1.2012 (Performance of Cardiac Output, Single, Manual) represented variable definitions. For the eICU-CRD cohort, patients were classified with IHCA if they either (i) presented with a time-stamped, database-specific procedure note indicating CPR, or (ii) were administered epinephrine, either as a bolus of 1mg/10ml or an infusion rate of 30mg/250ml at 100ml/hr, with an associated administration time. In both the MIMIC-IV and eICU-CRD cohorts, the control group was defined as patients who were not labeled as having experienced an IHCA nor deceased, and the reference time was set as the ICU discharge time. For IHCA patients with multiple labelings, we only selected the time of the first label as the reference time. The data collection method in the National Taiwan University Hospital database involves identifying patients with specific ICD codes (ICD-9 427.5, ICD-10 T46.2, 145.8, 146.9) ~~ICD-10 codes (I46.0, I46.1, I46.2, I46.8, I46.9, I49.01, I49.02, I49.03, I49.1, I49.3, I49.8, I49.9)~~. Patients who have been diagnosed with the aforementioned codes followed by the initiation of CPR or bolus epinephrine injection will be classified as IHCA patients.

Data curation and features extraction

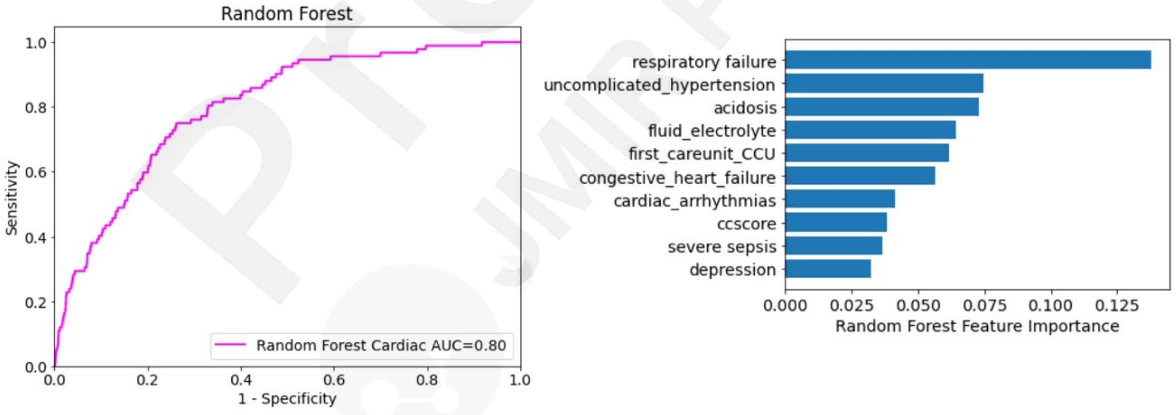
Two types of features were extracted: time-independent baseline features and time-varying physiologic readings from bedside monitors. Baseline features, which are variables registered at the time of admission, consisted of three types: (1) demographic information such as gender, age, ethnicity, type of ICU admission and BMI; (2) chronic comorbidities, as identified by combined comorbidity score and Elixhauser comorbidity index; [11,12] (3) presenting illness, as identified by ICD codes for acute cardiac disease, respiratory insufficiency, sepsis and potential reversible causes of cardiac arrest, popularly known as the “H’s (Hyperkalemia, Hypokalemia, Hypothermia, Hypoxemia, Hypovolemia, Hydrogen ion e.g. acidosis) and T’s (Spontaneous tension pneumothorax, Thrombosis, Cardiac tamponade)” by resuscitation guidelines.[13] Physiologic readings, which consisted of 6 vital signs: heart rate (HR), respiratory rate (RR), O2 saturation (SpO2), systolic blood pressure (sBP), diastolic blood pressure (dBP) and mean arterial pressure (MAP), were extracted on an hourly basis. For all patients, vital signs in the 24 hours prior to the reference time were recorded. To balance model utility with adequate accuracy, we only investigated the risk of cardiac arrest starting from 13 hours prior to the event. To overcome the time series' irregularity, specific rules were applied to combine multiple vital signs in the same hour (Supplementary material). The remaining missing values in vital signs were filled with the last observation carried forward method. To eliminate the misguidance of our imbalanced dataset, we tested two following remedies: Synthetic Minority Oversampling Technique (SMOTE) and near-miss algorithm (NearMiss).[14,15] We employed SMOTE in the following training with a nearest neighbor interpolation of 1 as it yielded a better performance compared to NearMiss (Supplementary Figure 1). After applying SMOTE, the numbers of IHCA patients and control patients were equal, signifying data balance.

Model development

Our predictive model was encoded in three layers (Figure 1). First, Random Forest (RF) was responsible for classifying the baseline features.[16] For hyperparameter optimization, the number of estimators was set to 5, the maximum depth was set to 20, and Gini impurity was used to determine the split. Nodes are expanded until all leaves contain fewer than 2 samples.[17] Second, RNN with the Long Short-Term Memory (LSTM) architecture stored the vital signs trajectories in an hourly pattern.[18] There were 3 hidden layers and 8 cells each, with a tangent and a sigmoid activation function. Learning rate was set to 0.001 and a dropout rate of 0.4 was applied for regularization.[19] The Adam algorithm was adapted for optimizing network weights.[20] Last, the Support Vector Machine (SVM) with a radial basis function kernel integrates RF and LSTM models to generate the final prediction. The SVM predicts the identical target outcome by learning the relationship between the predictions from two base models (RF and LSTM) and the target outcomes in the training set.[21] All the models were implemented in Python 3.8.3 with TensorFlow 2.1.0, pandas 1.1.2, scikit-learn 0.24.2, and NumPy 1.19.1 libraries.

Figure 1: Illustration of the modeling framework.

Each patient's data from the EHR is used as input for our model. Four pre-processing steps are carried out on the vital signs to obtain fixed interval data. All features go through SMOTE to overcome data imbalance and are split into training and testing groups. Baseline features are inputted to Random Forest and vital signs are inputted into LSTM for prediction. Support vector machine then integrates both models.



Evaluation strategy

To identify the perfect algorithm, the following ML techniques were evaluated in terms of prediction performance. First, based on baseline data's time-independency and binary structure, logistic regression (LR), K-nearest neighbor (KNN), extreme gradient boosting tree (XgBoost), and SVM were compared with RF for model fitness. In the LR model, we applied an L2 penalty with a stopping tolerance set at 1e-4, and the model underwent a maximum of 100 iterations. For the KNN algorithm, we set the parameter K to 2, utilizing Euclidean distance as the chosen metric. In the XGBoost model, the number of estimators was configured to 5 with a maximum depth of 5 and a learning rate of 0.1. Hyperparameter optimization was carried out through grid search. In the SVM, we utilized a radial basis function with an L2 penalty, setting the regularization parameter to 1. The SVM model was executed with a stopping tolerance of 1e-3, and no limit was imposed on the maximum number of iterations. For the time-dependent vital signs trajectories, the incorporation of memory gates in LSTM indicates its superiority in handling long sequence data. Thus, no other model comparison was made. To compare different stacking techniques, logistic regression (LR) was also implemented for comparison with SVM. Last, as we aim to use neural networks to accommodate our feature's complexity, we connected this three-layer model by engaging a deep neural network (DNN) in baseline data prediction and final stacking. The hyperparameters of DNN were set at an epoch of 30, batch size of 24 and the Adam algorithm as optimizer. Model performance was assessed based on discrimination and calibration using the internal validation cohort, as quantified by the AUROC with mean values and 95% confidence intervals (CIs).[22] Sensitivity and specificity metrics are presented by two binary classification, including a predefined threshold of 0.5 and an optimal cut-off determined by the Youden index.[23] We used the Brier Score to assess accuracy and visualized calibration curves across deciles based on observed and expected cardiac arrest numbers.[24].

Model interpretation

The importance of baseline features in the RF model was ranked based on “gain”, the cumulative improvement in accuracy of the nodes attributed to a specific feature. To focus more on the local impact of each vital signs at the patient level, we employed the SHapley Additive exPlanations

(SHAP) method to explain how our LSTM model makes predictions during a specific time point.[25]

Comparison with previous prediction score

Cardiac Arrest Risk Triage (CART), a commonly used cardiac arrest prediction model, was calculated to put the prediction results in perspective with prior studies.[26] A previously described "early warning score efficiency curve" was created to compare CART and our prediction model.[27] By plotting the percentage of detected events within 13 hours followed by the observations above the predefined threshold, a 0.5 probability in our model and a score of 20 in the CART model, we could demonstrate the changes of cumulative incidence as the event time approached. Due to the large number of missing data for temperature and neurological status in our development cohort, we were unable to compare our risk prediction tool against the MEWS or Acute Physiology and Chronic Health Evaluation (APACHE).

Results

Patient characteristics

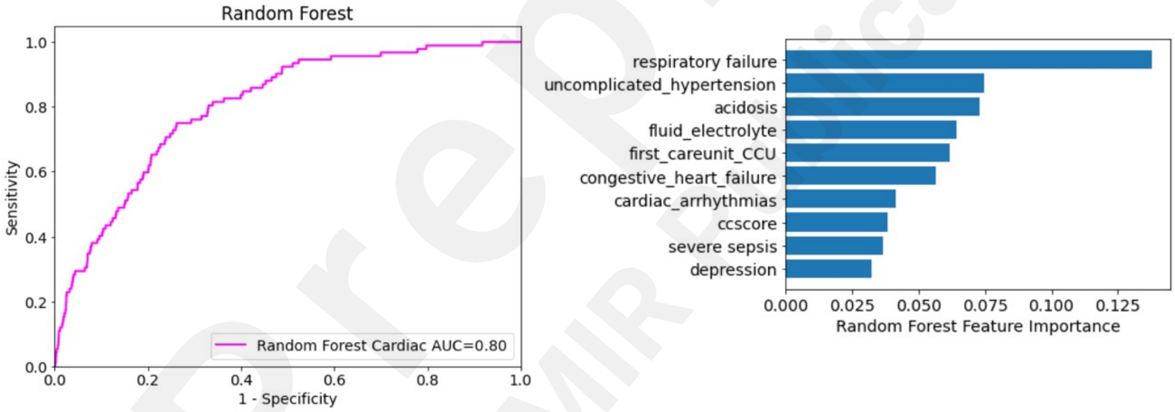
A total of 34,633 patients in the MIMIC-IV database and 79,643 patients in the eICU-CRD database were included in our analysis. After processing the vital signs data, a total of 452 IHCA patients and 23,457 control patients from MIMIC-IV were used for model development, whereas 85 IHCA patients and 9,964 control patients from eICU-CRD were used for external validation. Supplementary Table 1 shows the baseline characteristics of the IHCA group and the control group for the two cohorts. IHCA patients were significantly older ($p<0.001$), scored higher on combined comorbidity scores and the Elixhauser comorbidity index. In terms of presenting illness, myocardial infarction, pneumonia, respiratory failure and the "5 H's and 5 T's" were more prevalent in IHCA patients than among control patients.

Prediction from time-independent data

Patient demographics, comorbidities and presenting illness were first classified by RF. Figure 2 demonstrates the discrimination of RF model with an AUROC: 0.80 (0.779-0.844), sensitivity: 0.71, specificity 0. 78, f1-score: 0.79. Top five important features listed by RF include the presence of respiratory failure or acidosis, comorbid uncomplicated hypertension, comorbid fluid and electrolyte disorder, and initial ICU being the cardiac intensive care unit.

Figure 2: Prediction from baseline features.

(A) AUROC for evaluating the discriminatory ability of Random Forest on baseline features. (B) Feature importance derived from Random Forest model.



Modeling of time-dependent data

Trajectories of six vital signs were modeled with respect to time. Supplementary Figure 2a illustrates that in the MIMIC-IV cohort, the control group exhibited a constant value of all six vital signs throughout the 24-hour collecting period. However, vital signs of the IHCA patients were characterized by progressive deterioration in the last several hours. Of note, throughout the 24-hour monitoring period, patients who developed cardiac arrest exhibited on average, a 12 mmHg lower sBP, 1.5% lower SpO2 and a 9 bpm higher resting HR compared to the control group. However, the exact timing of the start of deterioration could not be clearly marked on the plot. A similar vital signs trajectory was seen in the eICU cohort (Supplementary Figure 2b).

Prediction from time-dependent data

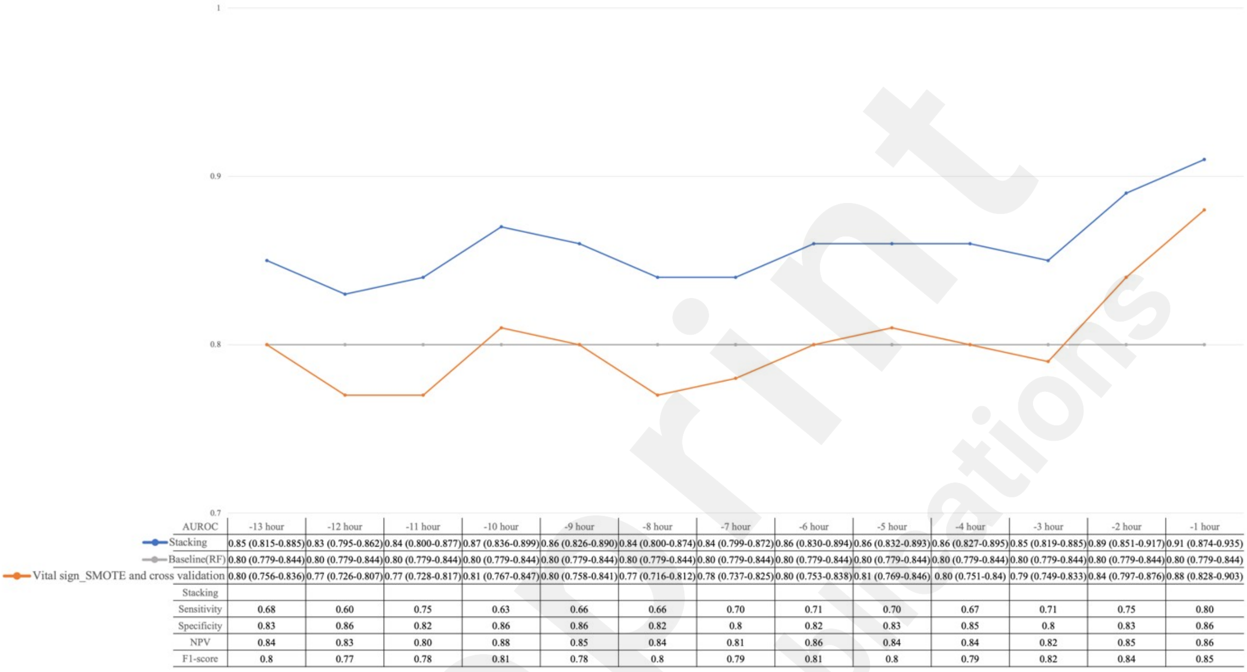
The hourly AUROC values for predicting cardiac arrest are presented in Supplementary Figure 3, which show the results after SMOTE and cross validation. A steady rise in AUROC was observed in the hours leading up to cardiac arrest with a sharp increase in the preceding three hours.

Performance of the SVM-based stacking model

In the final step of model construction, we stacked the LSTM model with the RF model and combined both predictions from baseline features and vital signs. AUROCs of the stacked model exhibited consistently better predictions compared with the baseline and vital signs only model, with the highest AUROC of 0.91 (0.874-0.935), sensitivity of 0.80, specificity of 0.86, and f1-score of 0.85 at one hour prior to the event. Further evaluation of the stacked model presented an increase in sensitivity, specificity, negative predictive value (NPV) and f1-score by the reduction of time interval (Figure 3). However, the calibration plot showed a risk of overestimation and a steadily low Brier Score throughout the 13 hours prediction time (Supplementary Figure 4). Additionally, in Supplementary Figure 5 we compared the model performance using different cutoffs. We found that the optimal cutoff defined by the Youden index (at 13 hours, 0.29; at 12 hours, 0.25; at 11 hours, 0.38; at 10 hours, 0.25; at 9 hours, 0.28; at 8 hours, 0.26; at 7 hours, 0.28; at 6 hours, 0.30; at 5 hours, 0.26; at 4 hours, 0.38; at 3 hours, 0.30; at 2 hours, 0.34; at 1 hour, 0.35.) presented with a better sensitivity compared with the predefined 0.5 cutoff, the largest difference was 14% at 12 hours prior to the event.

Figure 3: Performance of the stacked model in the MIMIC-IV database.

AUROCs of LSTM model with vital signs as input (orange plot), Random Forest model with baseline features as input (gray plot), and stacked model after integration of Random Forest and LSTM (blue plot) are shown. The exact values of the AUROCs of three models, sensitivity, specificity, NPV and f1-score of the stacked model are listed in the following table.



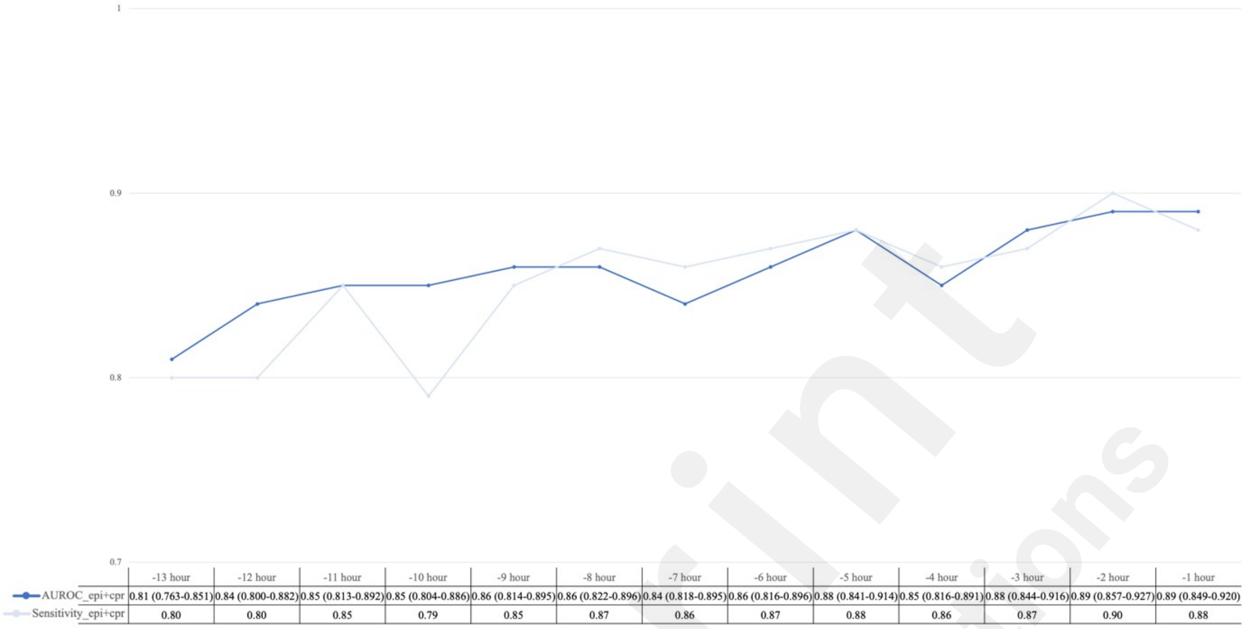
AUROC	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
Stacking	0.85 (0.815-0.885)	0.83 (0.795-0.862)	0.84 (0.800-0.877)	0.87 (0.836-0.899)	0.86 (0.826-0.890)	0.84 (0.800-0.874)	0.84 (0.799-0.872)	0.86 (0.830-0.894)	0.86 (0.832-0.893)	0.86 (0.827-0.895)	0.85 (0.819-0.885)	0.89 (0.851-0.917)	0.91 (0.874-0.935)
Baseline(RF)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)
Vital sign_SMOTE and cross validation	0.80 (0.756-0.836)	0.77 (0.726-0.807)	0.77 (0.728-0.817)	0.81 (0.767-0.847)	0.80 (0.758-0.841)	0.77 (0.716-0.812)	0.78 (0.737-0.825)	0.80 (0.753-0.838)	0.81 (0.769-0.846)	0.80 (0.751-0.84)	0.79 (0.749-0.833)	0.84 (0.797-0.876)	0.88 (0.828-0.903)
Stacking													
Sensitivity	0.68	0.60	0.75	0.63	0.66	0.66	0.70	0.71	0.70	0.67	0.71	0.75	0.80
Specificity	0.83	0.86	0.82	0.86	0.86	0.82	0.8	0.82	0.83	0.85	0.8	0.83	0.86
NPV	0.84	0.83	0.80	0.88	0.85	0.84	0.81	0.86	0.84	0.84	0.82	0.85	0.86
F1-score	0.8	0.77	0.78	0.81	0.78	0.8	0.79	0.81	0.8	0.79	0.82	0.84	0.85

External validation

We performed external validation of the stacked model in the eICU-CRD database. The results showed the best performance at one hour prior to IHCA with an AUROC of 0.89 (0.849-0.920), sensitivity of 0.79, specificity of 0.83, and an f1-score of 0.81. These findings align closely with the AUROC obtained from the MIMIC-IV dataset. (Figure 4). To further validate our model in an actual clinical scenario, we identified ~~50~~ 1935 IHCA patients and ~~50~~ 3692 control patients from the ICU of National Taiwan University Hospital (NTUH). Additionally, our model demonstrated high prediction sensitivity and an AUROC of ~~0.94~~ 0.945 when predicting IHCA one hour prior to its occurrence (Figure 5).

Figure 4: Performance of the stacked model in the eICU-CRD database.

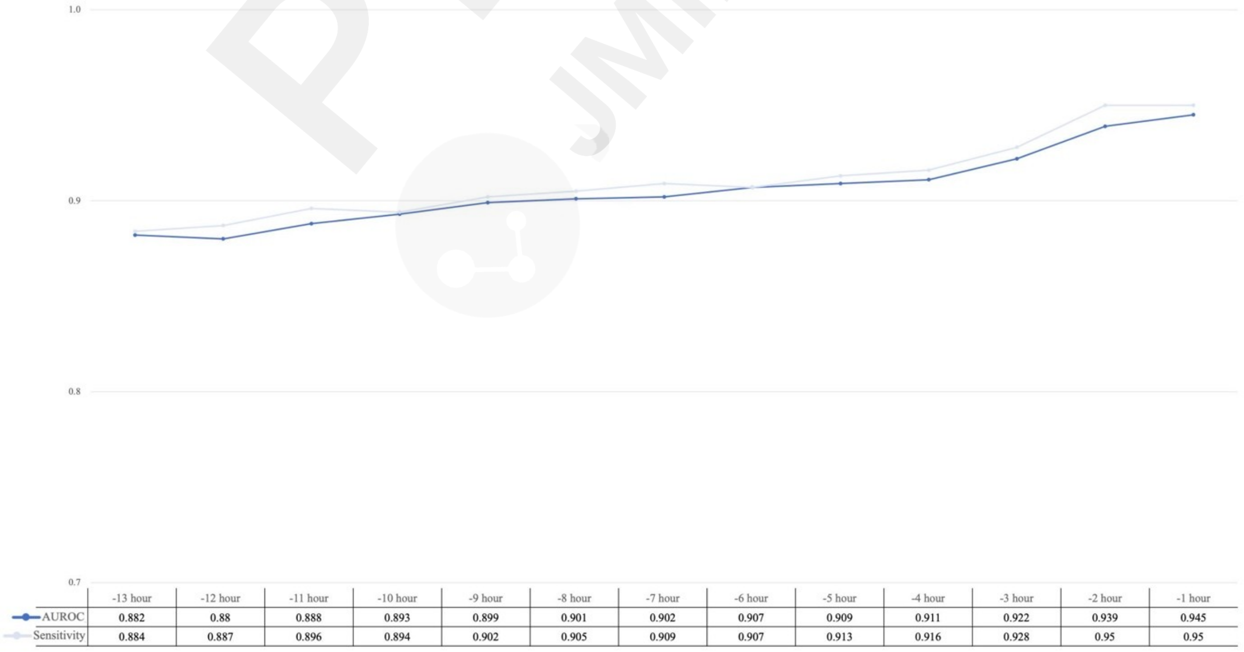
External validation of the stacked model is performed on the eICU-CRD database. AUROC is plotted in a blue line; sensitivity is plotted in a gray line.



	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
AUROC_epi+cpr	0.81 (0.763-0.851)	0.84 (0.800-0.882)	0.85 (0.813-0.892)	0.85 (0.804-0.886)	0.86 (0.814-0.895)	0.86 (0.822-0.896)	0.84 (0.818-0.895)	0.86 (0.816-0.896)	0.88 (0.841-0.914)	0.85 (0.816-0.891)	0.88 (0.844-0.916)	0.89 (0.857-0.927)	0.89 (0.849-0.920)
Sensitivity_epi+cpr	0.80	0.80	0.85	0.79	0.85	0.87	0.86	0.87	0.88	0.86	0.87	0.90	0.88

Figure 5: Performance of the stacked model in the clinical scenario

External validation of the stacked model is performed using data from ~~50~~ 1935 in-hospital cardiac arrest patients and ~~50~~ 3692 control patients collected from the National Taiwan University Hospital. AUROC is plotted in a blue line; sensitivity is plotted in a gray line.



	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
AUROC	0.882	0.88	0.888	0.893	0.899	0.901	0.902	0.907	0.909	0.911	0.922	0.939	0.945
Sensitivity	0.884	0.887	0.896	0.894	0.902	0.905	0.909	0.907	0.913	0.916	0.928	0.95	0.95

Local interpretation of LSTM model

We adopted the SHAP method to enable model explanation from an individual patient's perspective. In each box, SHAP values for specific vital signs are assigned, with red (SHAP value above 1) indicating a risk factor and blue (SHAP value below 1) indicating a protective factor. Supplementary Figure 6A represents a patient from the MIMIC-IV database experiencing IHCA at time 0hr. As IHCA approaches, an increase in sBP from its average contributes to an elevated risk, with the most significant effect occurring 6 hours prior to IHCA. However, at one hour before IHCA, the most significant risk becomes a decrease in sBP from its average. Supplementary Figure 6B illustrates another IHCA patient from the eICU database. In contrast to Figure 6A, the most prominent feature at 1 hr prior to IHCA is a decrease in HR and SpO2 from its baseline value. These figures showcase diverse presentations leading to IHCA in various patients, providing a valuable guideline for medical staff to identify the specific organ failure responsible for IHCA. The significance lies in enabling a swift response, incorporating timely interventions such as intubation for saturation drop and the administration of inotropic agents for decreased sBP. This approach ensures that medical staff will not delay necessary treatments while determining the cause of IHCA.

Performance compared with different ML and deep learning algorithms

Conventional statistics and supervised machine learning algorithms were compared to predict IHCA using only baseline features. RF demonstrated superior performance in terms of AUROC compared with XgBoost, LR, KNN, and SVM (Supplementary Figure 7). SVM also presented preferable results during the stacking operation compared with LR throughout the 13 hours prediction period. AUROCs at one hour prior to the incidence of IHCA were 0.91 versus 0.80, respectively (Supplementary Figure 8). Finally, using a Neural Network to connect baseline, vital signs, and stacking predictions did not reveal an improving outcome (Supplementary Figure 9). After comparing several algorithms and combinations, RF, LSTM, and SVM predictions still yielded the most satisfactory results.

Detection efficacy compare to previous prediction score

We compared the performance of our proposed model to that of the CART score. Overall, our model demonstrated better AUROC throughout the prediction period (Supplementary Figure 10). As illustrated in Supplementary Figure 11, it is evident that at the time of 12 hours prior to cardiac arrest, our model was able to detect over 70% of patients at risk for IHCA, compared to the CART score which did not surpass 65% of detection rate even one hour prior to IHCA.

Discussion

Principal Findings

In this retrospective study of 34,633 patients in the MIMIC-IV database, we constructed a high-performance multimodal model (AUROC: 0.91 (0.874-0.935)) that can predict IHCA up to 13 hours in advance using EHR and high-resolution time-series physiological readings. As the time of cardiac arrest approached, our model yielded a steady increase in detection rate, finally reaching 89 percent one hour prior to the event. We also illustrated the impact of each vital signs on the prediction of cardiac arrest associated with individual patients through the use of SHAP values. Furthermore, we demonstrated the advantage of this ML algorithm over the CART score, which was derived using traditional regression models.

Comparison to Prior Work

As a ubiquitous activity in the hospital, several studies had demonstrated the importance of vital signs measurement in determining a patient’s disease course.[28] dBP, RR, and maximum HR have all been found to be significant and independent predictors of cardiac arrest.[29] However, maintaining a minimal model with only vital signs or adding lab data as predictors at the cost of decreasing model adaptability remains a dilemma.[30,31] Lactic acid level is the most representative laboratory biomarker in circulatory failure, but had a high rate of missingness in the MIMIC-IV database (16,317/23,909 (68.2%)). This motivated us to abandon utilizing lab results and assess if a nimbler model could be constructed with vital signs trends alone, overlaying the easily obtainable ICD codes and patient demographics as baseline features. Unsurprisingly, a significant increase in AUROC was discovered by adding demographics and comorbidities to the vital signs only model. Furthermore, a SVM-based stacked model can address the predictive capabilities of underlying conditions and dynamic changes during disease deterioration. Stacking proves advantageous by compensating for the weaknesses of both models, with RF potentially struggling with highly correlated data while LSTM exceling in handling timely intricate information.

Distinct Advantages of Our Approach

The reason for not establishing an end-to-end neural network throughout the prediction stood out as supervised ML algorithms retained the ability to determine the importance of each predictor and have a better model explainability. Moreover, in the ensemble technique, stacking excels over both boosting and bagging due to its versatility in integrating diverse data domains and combining various types of models. Late fusion at the model level is also preferred over other fusion methods for mitigating feature discrepancies and enabling independent model training between the time-independent baseline and time-dependent vital signs. Additionally, the outperformance of SVM over LR in the stacking operation could be attributed to better data handling using the nonlinear Kernel function. To evaluate the external validity of our model, we tested it on two distinct datasets: the eICU database and NTUH database, both representing patient groups with diverse ethnicities and disease backgrounds. Over a 10-year duration, we identified 1935 IHCA cases (34.3%) in NTUH. In contrast to prior IHCA prediction studies, such as Kwon et al.'s 2.3% over 7 years (1233 cases), Chae et al.'s 1.3% over 4 years (1154 cases), and Ding et al.'s 23.09% over 5 years (1796 cases), our clinical database demonstrated a higher IHCA incidence yet fewer cases.[32–34] This disparity is attributed to our ICU-focused validation database, in contrast to earlier studies that encompassed all hospitalized patients. Consequently, our approach ensures heightened data precision and a more nuanced understanding of patient dynamics through continuous monitoring within this critically ill cohort. Nevertheless, our high prediction quality ~~observed~~ in both independent databases ensure the credibility of our model across various demographic groups and subpopulations. The consistent performance across these datasets not only minimizes the possibility of overfitting but also validates the generalizability of our predictions.~~ However, we do not recommend administering the identical model in all types of real-world ICUs. In fact, our system should be viewed as alterable, and it is the similar methodology that should be carried out in various local data along with extra fine-tuning to reach optimal prediction. ~~

Limitations of Our Methodology

Our study had limitations because we used data collected from one medical center. First, due to the nature of EHR, we were unable to determine the reason for the multi-scale gaps and different frequencies of each input. Second, we did not include clinical interventions, body temperature, and mental status in our model. Clinical interventions may change the disease course or even terminate the deterioration process. Nevertheless, the complexity of the treatment record and the high frequency of missing values in temperature and mental status compelled us to omit these valuable predictors. Third, our identification of IHCA relied on time-labeled: (1) databases specific procedure code, (2) ICD-procedure codes, (3) administration of epinephrine in resuscitation dosages. In real-time clinical scenarios, delays in data entry may occur as documentation is considered secondary to patient care. Additionally, the accuracy of these codes is often operator-dependent and may vary across different ICU policies. To minimize recording biases, we manually reviewed all IHCA vital signs data and only included reasonable measurements, ensuring that the identified IHCA timepoints correlated with the worst patient vital signs.

Conclusion

We built a multimodal ML model based on time serial vital signs and three types of baseline features, which were all easily accessible in the intensive care unit. Our model showed high accuracy in detecting clinical deterioration leading to development of IHCA, up to 13 hours in advance, in both the internal and external validation cohorts. A model like this could be integrated into a hospital's EHR system to identify high-risk patients and provide clinical decision support.

Author contributions

C-CL has full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis, concept and design, critical revision of the manuscript for important intellectual content, obtaining funding and supervision. H-YL, P-CK, FQ, H-JJ, P-HC, C-HL, and I-JW were responsible for drafting the manuscript, interpretation of the data, and critical revision of the manuscript for important intellectual content. CH-L was responsible for statistical analysis. J-RH, and W-TH were responsible for critical revision of the manuscript for important intellectual content. All authors read and approved the final manuscript.

Funding statement

This study was funded by the Grant (MOST-110-2622-8-002-017). No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

All authors are fully included throughout the research process and declare no conflicts of interest.

Data availability statement

The datasets analyzed during the current study are available in the MIMIC IV repository, https://mimic.mit.edu and eICU-CRD repository, https://eicu-crd.mit.edu/about/eicu/

Code availability statements

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

References

1. Sinha SS, Sukul D, Lazarus JJ, Polavarapu V, Chan PS, Neumar RW, Nallamothu BK. Identifying Important Gaps in Randomized Controlled Trials of Adult Cardiac Arrest Treatments: A Systematic Review of the Published Literature. 2017;17.
2. Holmberg MJ, Ross CE, Fitzmaurice GM, Chan PS, Duval-Arnould J, Grossestreuer AV, Yankama T, Donnino MW, Andersen LW, for the American Heart Association's Get With The Guidelines–Resuscitation Investigators*, Chan P, Grossestreuer AV, Moskowitz A, Edelson D, Ornato J, Berg K, Peberdy MA, Churpek M, Kurz M, Starks MA, Girotra S, Perman S, Goldberger Z, Guerguerian A-M, Atkins D, Foglia E, Fink E, Lasa JJ, Roberts J, Bembea M, Gaies M, Kleinman M, Gupta P, Sutton R, Sawyer T. Annual Incidence of Adult and Pediatric In-Hospital Cardiac Arrest in the United States. Circ: Cardiovascular Quality and Outcomes 2019 Jul;12(7). doi: 10.1161/CIRCOUTCOMES.119.005580
3. Morrison LJ, Neumar RW, Zimmerman JL, Link MS, Newby LK, McMullan PW, Hoek TV, Halverson CC, Doering L, Peberdy MA, Edelson DP. Strategies for Improving Survival After In-Hospital Cardiac Arrest in the United States: 2013 Consensus Recommendations: A Consensus Statement From the American Heart Association. Circulation 2013 Apr 9;127(14):1538–1563. doi: 10.1161/CIR.0b013e31828b2770
4. Spångfors M, Molt M, Samuelson K. In-hospital cardiac arrest and preceding National Early Warning Score (NEWS): A retrospective case-control study. Clin Med (Lond) 2020 Jan;20(1):55–60. PMID:31941734
5. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger' systems. Resuscitation 2008 May;77(2):170–179. doi: 10.1016/j.resuscitation.2007.12.004
6. Bartkowiak B, Snyder AM, Benjamin A, Schneider A, Twu NM, Churpek MM, Roggin KK, Edelson DP. Validating the Electronic Cardiac Arrest Risk Triage (eCART) Score for Risk Stratification of Surgical Inpatients in the Postoperative Setting: Retrospective Cohort Study. Annals of Surgery 2019 Jun;269(6):1059–1063. doi: 10.1097/SLA.0000000000002665
7. Wang A-Y, Fang C-C, Chen S-C, Tsai S-H, Kao W-F. Periarrest Modified Early Warning Score (MEWS) predicts the outcome of in-hospital cardiac arrest. Journal of the Formosan Medical Association 2016 Feb;115(2):76–82. doi: 10.1016/j.jfma.2015.10.016
8. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000 Jun 13;101(23). doi: 10.1161/01.CIR.101.23.e215
9. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Dec;5(1):180178. doi: 10.1038/sdata.2018.178
10. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000 Jun 13;101(23). doi: 10.1161/01.CIR.101.23.e215
11. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol 2011 Jul;64(7):749–759. PMID:21208778
12. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care United States; 1998 Jan;36(1):8–27. PMID:9431328
13. Soar J, Nolan JP, Böttiger BW, Perkins GD, Lott C, Carli P, Pellis T, Sandroni C, Skrifvars MB, Smith GB, Sunde K, Deakin CD, Koster RW, Monsieurs KG, Nikolaou NI. European Resuscitation Council Guidelines for Resuscitation 2015. Resuscitation 2015 Oct;95:100–147. doi: 10.1016/j.resuscitation.2015.07.016

14. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 Dec;14(1):106. doi: 10.1186/1471-2105-14-106
15. Mani I, Zhang I. kNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of workshop on learning from imbalanced datasets ICML; 2003. p. 1–7.
16. Liaw A, Wiener M. Classification and Regression by randomForest. 2002;2:5.
17. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. :25.
18. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. MIT Press; 1997. p. 1735–1780. Available from: https://doi.org/10.1162/neco.1997.9.8.1735
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. :30.
20. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:14126980 [cs] 2017 Jan 29; Available from: http://arxiv.org/abs/1412.6980 [accessed Feb 16, 2021]
21. Wang J, Feng K, Wu J. SVM-Based Deep Stacking Networks. AAAI 2019 Jul 17;33:5273–5280. doi: 10.1609/aaai.v33i01.33015273
22. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015 Jan 6;162(1):W1. doi: 10.7326/M14-0698
23. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. Biom J 2008 Jun;50(3):419–430. doi: 10.1002/bimj.200710415
24. Rolke W, Gongora CG. A Chi-square Goodness-of-Fit Test for Continuous Distributions against a known Alternative. Comput Stat 2021 Sep;36(3):1885–1900. doi: 10.1007/s00180-020-00997-x
25. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. :10.
26. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs*: Critical Care Medicine 2012 Jul;40(7):2102–2108. doi: 10.1097/CCM.0b013e318250aa5a
27. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013 Apr;84(4):465–470. doi: 10.1016/j.resuscitation.2012.12.016
28. Smith GB. Vital signs: Vital for surviving in-hospital cardiac arrest? Resuscitation 2016 Jan;98:A3–A4. doi: 10.1016/j.resuscitation.2015.10.010
29. Churpek MM, Yuen TC, Huber MT, Park SY, Hall JB, Edelson DP. Predicting Cardiac Arrest on the Wards. Chest 2012 May;141(5):1170–1176. doi: 10.1378/chest.11-1301
30. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using Time Series Analysis to Predict Cardiac Arrest in a PICU: Pediatric Critical Care Medicine 2015 Nov;16(9):e332–e339. doi: 10.1097/PCC.0000000000000560
31. Ueno R, Xu L, Uegami W, Matsui H, Okui J, Hayashi H, Miyajima T, Hayashi Y, Pilcher D, Jones D. Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: A single-center retrospective cohort study. Cheungpasitporn W, editor. PLoS ONE 2020 Jul 13;15(7):e0235835. doi: 10.1371/journal.pone.0235835
32. Kwon J, Lee Y, Lee Y, Lee S, Park J. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. J Am Heart Assoc 2018 Jul 3;7(13). doi: 10.1161/JAHA.118.008678
33. Chae M, Han S, Gil H, Cho N, Lee H. Prediction of In-Hospital Cardiac Arrest Using Shallow and Deep Learning. Diagnostics 2021 Jul 13;11(7):1255. doi: 10.3390/diagnostics11071255
34. Ding X, Wang Y, Ma W, Peng Y, Huang J, Wang M, Zhu H. Development of early prediction model of in-hospital cardiac arrest based on laboratory parameters. BioMed Eng OnLine 2023

Dec 6;22(1):116. doi: 10.1186/s12938-023-01178-9

35.

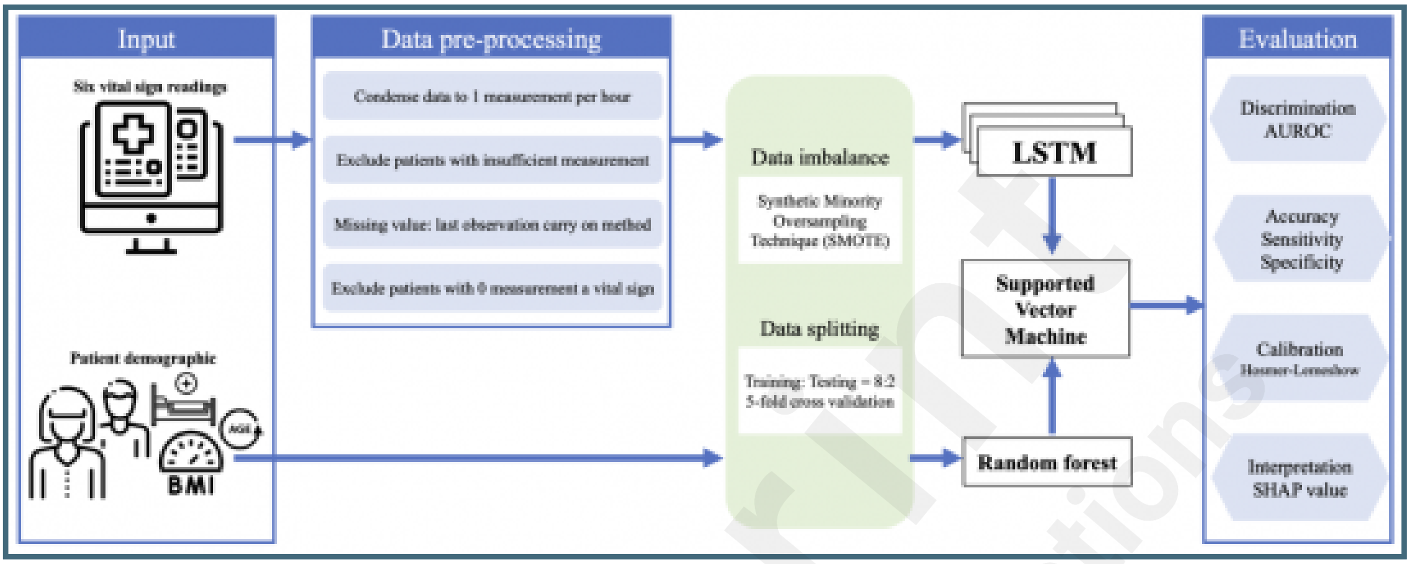
Abbreviations

AUROC, Area under the receiver operating characteristic curve;
CART, Cardiac Arrest Risk Triage;
dBP, Diastolic blood pressure;
DNN, Deep neural network;
EHRs, Electronic health records;
eICU-CRD, Electronic ICU Collaborative Research Database;
EWS, Early warning scores;
HR, Heart rate;
ICD, International Classification of Diseases;
ICUs, Intensive care units;
IHCA, In-hospital cardiac arrest;
KNN, K-nearest neighbor;
LR, Logistic regression;
LSTM, Long short-term memory;
mAP, Mean arterial pressure;
MEWS, Modified Early Warning Score;
MIMIC, Multiparameter Intelligent Monitoring of Intensive Care;
ML, Machine learning;
NPV, Negative predictive value;
NTUH, National Taiwan University Hospital
PCS, Procedure Coding System;
RF, Random forest;
RNN, Recurrent neural networks;
RR, Respiratory rate;
sBP, Systolic blood pressure;
SHAP, SHapley Additive exPlanations;
SMOTE, Synthetic Minority Oversampling Technique;
SpO2, Oxygen saturation;
SVM, Support vector machine;
XgBoost, Extreme gradient boosting;

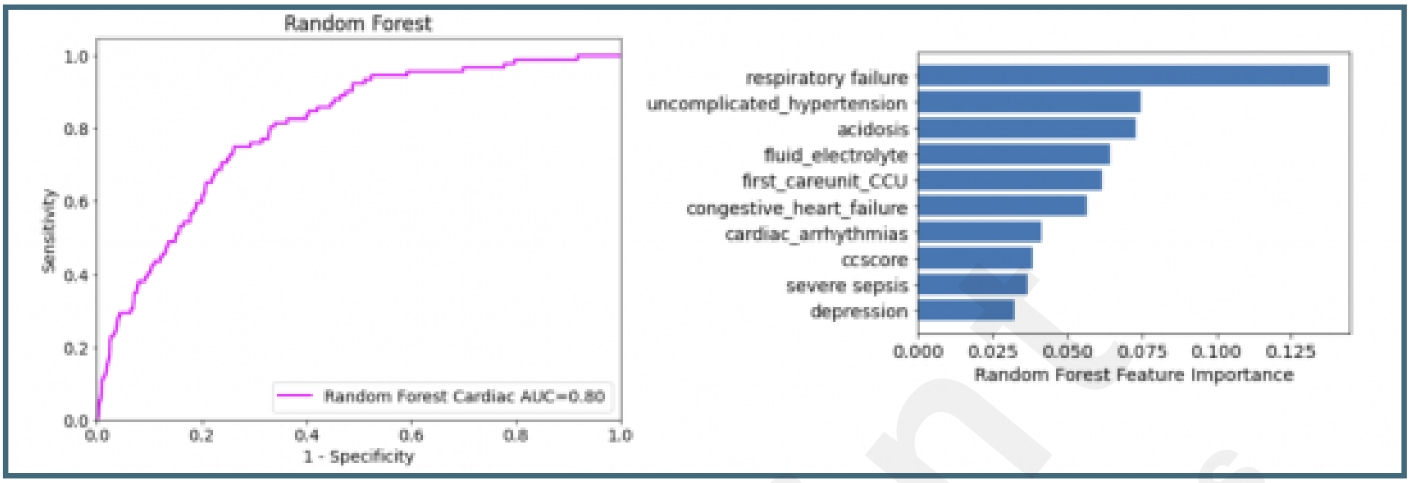
Supplementary Files

Figures

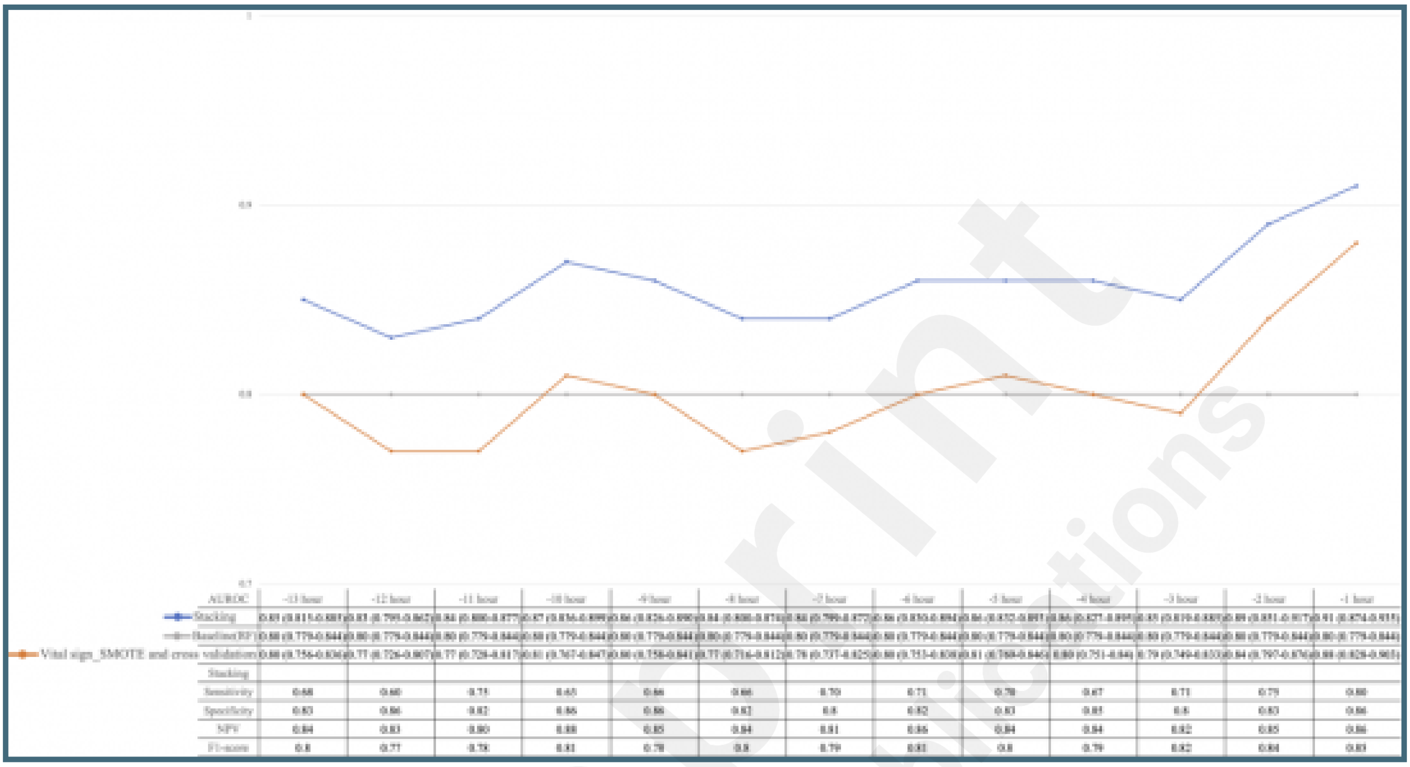
Each patient’s data from the EHR is used as input for our model. Four pre-processing steps are carried out on the vital signs to obtain fixed interval data. All features go through SMOTE to overcome data imbalance and are split into training and testing groups. Baseline features are inputted to Random forest and vital signs are inputted into LSTM for prediction. Support vector machine then integrates both models.


Input
Six vital sign readings
Patient demographic
AGE
BMI
Data pre-processing
Condense data to 1 measurement per hour
Exclude patients with insufficient measurement
Missing value: last observation carry on method
Exclude patients with 0 measurement a vital sign
Data imbalance
Synthetic Minority Oversampling Technique (SMOTE)
Data splitting
Training: Testing = 8:2
5-fold cross validation
LSTM
Supported Vector Machine
Random forest
Evaluation
Discrimination AUROC
Accuracy Sensitivity Specificity
Calibration Hosmer-Lemeshow
Interpretation SHAP value

(A) AUROC for evaluating the discriminatory ability of Random forest on baseline features. (B) Feature importance derived from Random forest model.

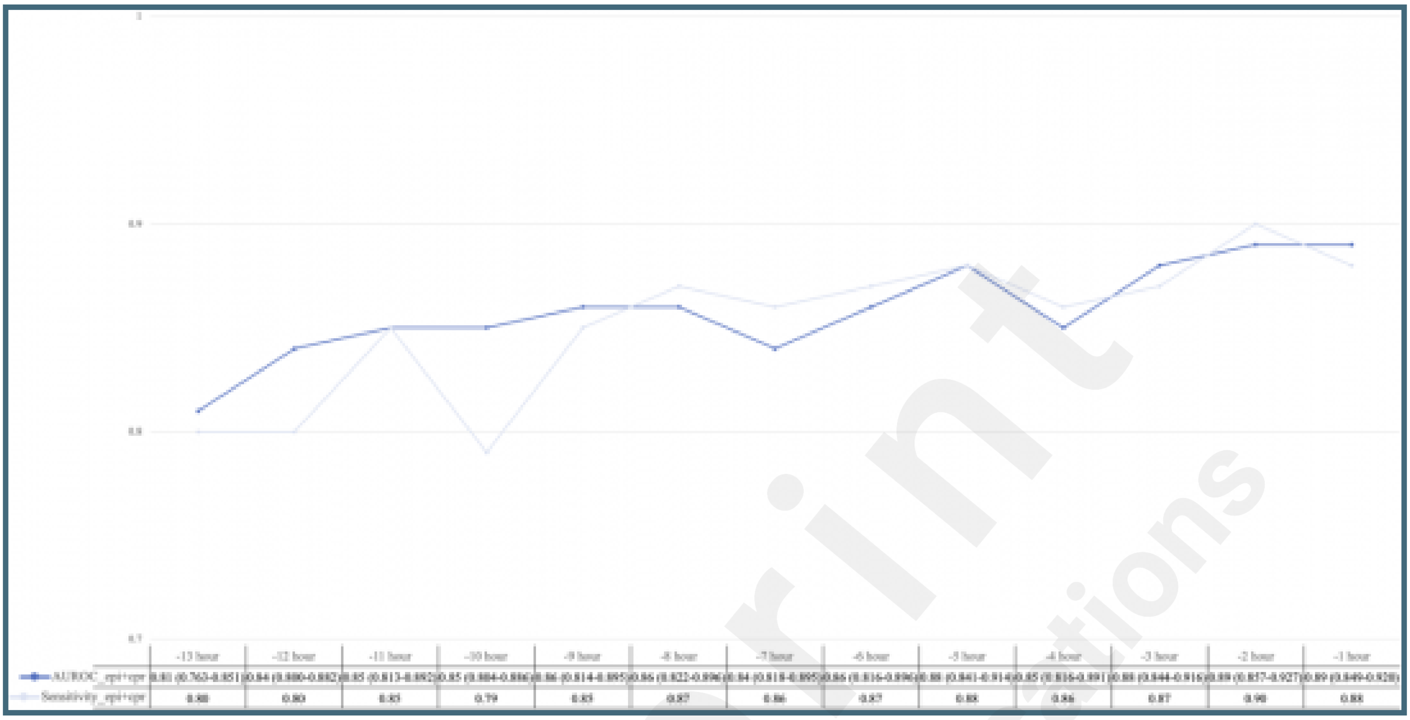


AUROCs of LSTM model with vital signs as input (orange plot), Random Forest model with baseline features as input (gray plot), and stacked model after integration of Random Forest and LSTM (blue plot) are shown. The exact values of the AUROCs of three models, sensitivity, specificity, NPV and f1-score of the stacked model are listed in the following table.



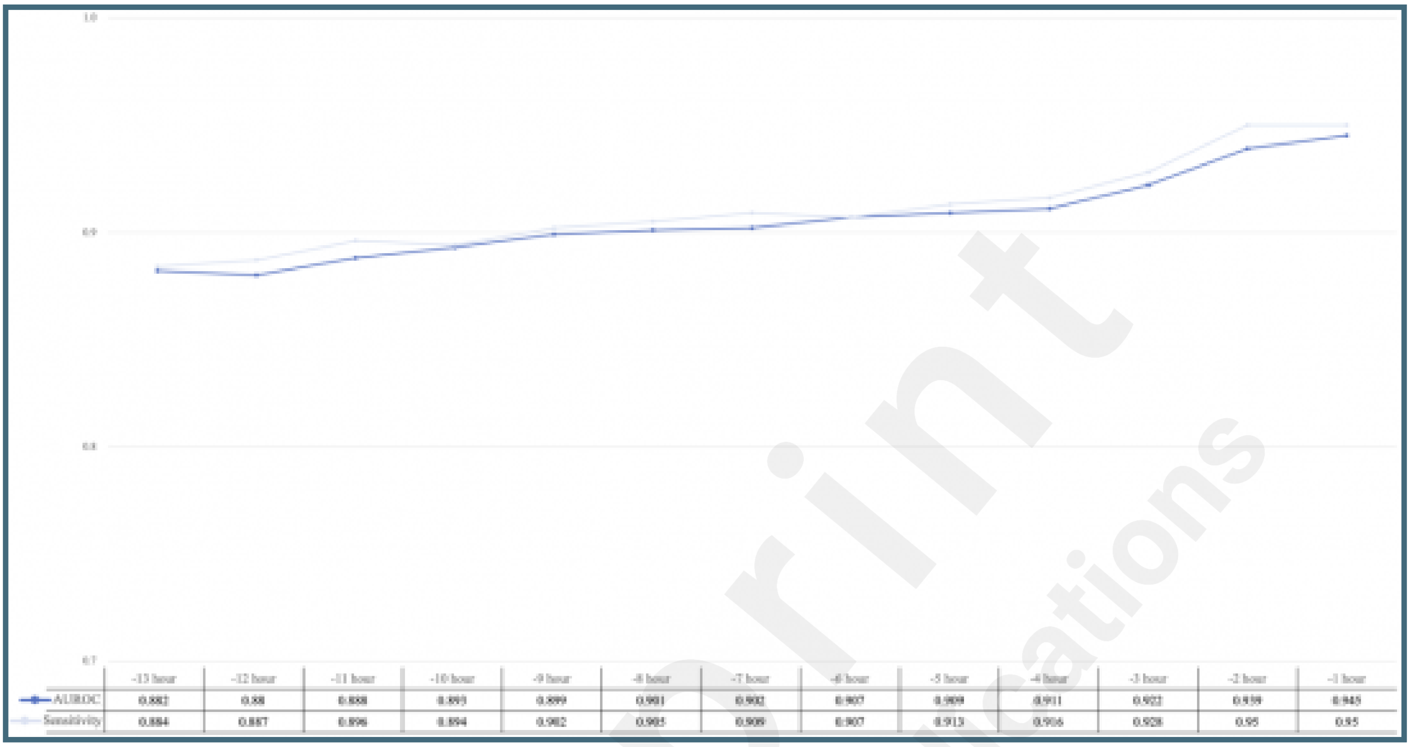
AUROC	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
Stacking	0.85 (0.815-0.885)	0.83 (0.795-0.862)	0.84 (0.800-0.877)	0.87 (0.836-0.899)	0.86 (0.826-0.890)	0.84 (0.808-0.874)	0.84 (0.799-0.877)	0.86 (0.830-0.894)	0.86 (0.832-0.891)	0.86 (0.827-0.895)	0.85 (0.819-0.885)	0.89 (0.851-0.917)	0.91 (0.874-0.935)
Baseline RF	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)	0.80 (0.779-0.844)
Vital sign_SMOTE and cross validation	0.80 (0.756-0.836)	0.77 (0.726-0.807)	0.77 (0.729-0.817)	0.81 (0.767-0.847)	0.80 (0.758-0.841)	0.77 (0.716-0.812)	0.78 (0.737-0.825)	0.80 (0.753-0.838)	0.81 (0.769-0.846)	0.80 (0.751-0.84)	0.79 (0.749-0.833)	0.84 (0.797-0.876)	0.88 (0.828-0.905)
Stacking													
Sensitivity	0.68	0.60	0.73	0.65	0.66	0.66	0.70	0.71	0.70	0.67	0.71	0.75	0.80
Specificity	0.83	0.86	0.82	0.86	0.86	0.82	0.8	0.82	0.83	0.85	0.8	0.83	0.86
NPV	0.84	0.83	0.80	0.88	0.85	0.84	0.81	0.86	0.84	0.84	0.82	0.85	0.86
F1-score	0.8	0.77	0.78	0.81	0.78	0.8	0.79	0.81	0.8	0.79	0.82	0.84	0.83

External validation of the stacked model is performed on the eICU-CRD database. AUROC is plotted in a blue line; sensitivity is plotted in a gray line.



	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
AUROC_epi+cpr	0.81 (0.763-0.851)	0.84 (0.800-0.882)	0.85 (0.813-0.892)	0.85 (0.804-0.886)	0.86 (0.814-0.895)	0.86 (0.822-0.896)	0.84 (0.818-0.895)	0.86 (0.816-0.896)	0.88 (0.841-0.914)	0.85 (0.816-0.891)	0.88 (0.844-0.916)	0.89 (0.857-0.927)	0.89 (0.849-0.928)
Sensitivity_epi+cpr	0.80	0.80	0.85	0.79	0.85	0.87	0.86	0.87	0.88	0.86	0.87	0.90	0.88

External validation of the stacked model is performed using data from 1935 in-hospital cardiac arrest patients and 3692 control patients collected from the National Taiwan University Hospital. AUROC is plotted in a blue line; sensitivity is plotted in a gray line.



	-13 hour	-12 hour	-11 hour	-10 hour	-9 hour	-8 hour	-7 hour	-6 hour	-5 hour	-4 hour	-3 hour	-2 hour	-1 hour
AUROC	0.882	0.88	0.888	0.893	0.899	0.901	0.902	0.907	0.909	0.911	0.922	0.939	0.945
Sensitivity	0.884	0.887	0.896	0.894	0.902	0.905	0.909	0.907	0.913	0.916	0.928	0.95	0.95

Multimedia Appendixes

Supplementary Table, Supplementary Figures, Supplementary Material.
URL: http://asset.jmir.pub/assets/6f7dd928948c79e834a64b2d5bbc2406.docx