ORIGINAL ARTICLE

Machine-learning models are superior to severity scoring systems for the prediction of the mortality of critically ill patients in a tertiary medical center

Ruey-Hsing Chou[a,b,c], Benny Wei-Yun Hsu[d], Chun-Lin Yu[e], Tai-Yuan Chen[d], Shuo-Ming Ou[c,f,g], Kuo-Hua Lee[c,f,g], Vincent S. Tseng[h,*], Po-Hsun Huang[a,b,c,*], Der-Cherng Tarng[c,f,g,i,*]

[a]Department of Critical Care Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC [b]Cardiovascular Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC [c]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC [d]Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC [e]Institute of Data Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC [f]Division of Nephrology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC [g]Center for Intelligent Drug Systems and Smart Bio-Devices (IDS2B), National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC [h]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ROC [i]Department and Institute of Physiology, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC

ABSTRACT

Background: Intensive care unit (ICU) mortality prediction helps to guide therapeutic decision making for critically ill patients. Several scoring systems based on statistical techniques have been developed for this purpose. In this study, we developed a machine-learning model to predict patient mortality in the very early stage of ICU admission.

Methods: This study was performed with data from all patients admitted to the intensive care units of a tertiary medical center in Taiwan from 2009 to 2018. The patients' comorbidities, co-medications, vital signs, and laboratory data on the day of ICU admission were obtained from electronic medical records. We constructed random forest and extreme gradient boosting (XGBoost) models to predict ICU mortality, and compared their performance with that of traditional scoring systems.

Results: Data from 12,377 patients was allocated to training (n = 9901) and testing (n = 2476) datasets. The median patient age was 70.0 years; 9210 (74.41%) patients were under mechanical ventilation in the ICU. The areas under receiver operating characteristic curves for the random forest and XGBoost models (0.876 and 0.880, respectively) were larger than those for the Acute Physiology and Chronic Health Evaluation II score (0.738), Sequential Organ Failure Assessment score (0.747), and Simplified Acute Physiology Score II (0.743). The fraction of inspired oxygen on ICU admission was the most important predictive feature across all models.

Conclusion: The XGBoost model most accurately predicted ICU mortality and was superior to traditional scoring systems. Our results highlight the utility of machine learning for ICU mortality prediction in the Asian population.

Keywords: Intensive care units; Machine learning; Mortality

1. INTRODUCTION

The prediction of mortality in intensive care units (ICUs) helps to guide therapeutic decision making and resource allocation. It may also be useful for the counseling of family members and provision of prognostic information about critically ill patients.[1] Several tools have been applied to predict the mortality of these patients; they include the Acute Physiology and Chronic Health Evaluation (APACHE) II,[2] the Sequential Organ Failure Assessment (SOFA),[3] and the Simplified Acute Physiology Score (SAPS) II.[4] However, most such scoring systems were developed with Caucasian populations, and their accuracy when applied to Asian populations is unclear. Furthermore, these systems are based on traditional statistical techniques, with which the management of the abundance of data collected in the ICU is difficult, and do not utilize comprehensive patient information. In contrast, machine-learning techniques enable the analysis of complex signals in data-rich environments.[5] This single-center study was conducted to develop a machine-learning model for the prediction of mortality in the very early stage of ICU admission using large-scale data collected from patients' electronic

* Address correspondence. Dr. Vincent S. Tseng, Department of Computer Science, National Yang Ming Chiao Tung University, 1001, University Road, Hsinchu 300, Taiwan, ROC. E-mail address: vtseng@cs.nctu.edu.tw (V. S. Tseng).

Dr. Po-Hsun Huang, Department of Critical Care Medicine, Taipei Veterans General Hospital, 201, Section 2, Shi-Pai Road, Taipei 112, Taiwan, ROC. E-mail address: huangbsvgh@gmail.com (P.-H. Huang).

Dr. Der-Cherng Tarng, Department of Medicine, Taipei Veterans General Hospital, 201, Section 2, Shi-Pai Road, Taipei 112, Taiwan, ROC. E-mail address: dctarng@vghtpe.gov.tw (D.-C. Tarng).

Author contributions: Dr. Ruey-Hsing Chou and Dr. Wei-Yun Hsu contributed equally to this work.

Conflicts of interest: Dr. Po-Hsun Huang and Dr. Der-Cherng Tarng, editorial board members at Journal of the Chinese Medical Association, have no roles in the peer review process of or decision to publish this article. The other authors declare that they have no conflicts of interest related to the subject matter or materials discussed in this article.Journal of Chinese Medical Association. (2024) 87: 369-376.

Received September 6, 2022; accepted June 11, 2023.

doi: 10.1097/JCMA.0000000000001066

medical records and by physiological monitoring. We hypothesized that the machine-learning model would be more accurate than traditional scoring systems.

2. METHODS

2.1. Study population

We retrospectively screened the records and other data collected from all patients aged >20 years who were admitted to the medical and surgical ICUs of Taipei Veterans General Hospital from 2009 to 2018. The collected data included demographic characteristics, medical histories, vital signs, and laboratory findings from patients' ICU stays. Philips IntelliSpace Critical Care and Anesthesia systems, which enable the collection of rich data about patients' conditions streamed automatically from bedside monitors and input manually by health care providers, were used in the ICUs. These data included hemodynamic and ventilation parameters (e.g., from electrocardiographic monitors, pulse oximeters, and multiparameter monitors), nutrition prescriptions, information about medications administered, and regular notes from medical staff. We obtained data not recorded in the Intellispace Critical Care and Anesthesia systems (e.g., on medicines administered in the outpatient department and adverse events occurring after ICU discharge) from the hospital's electronic medical records system. In order to address the missing values issue, an initial step involved the removal of features exhibiting a substantial proportion of missing data, specifically those with a missing rate exceeding 50%. Subsequently, for the remaining features, mean imputation was applied to continuous variables to fill in the missing values. Notably, the continuous variables which needed to impute missing fields, such as serum sodium, calcium concentrations, and central venous pressure, were Gaussian (normal) distribution. Additionally, no missing values were present in the categorical variables after eliminating features with high missing rates. All distributions of those features were symmetric, and the mean and median are all at the exact center value.

To evaluate disease severity, APACHE II[2] and SOFA[3] and SAPS II scores[4] were calculated within 24 hour after ICU admission. The lowest mean arterial pressure and the highest heart rate (HR) within 24 h after ICU admission were recorded. The use of inotropes or vasopressors, such as norepinephrine and dopamine, was also recorded. White blood cell counts and blood chemistry studies were performed on ICU admission using routine laboratory methods. Sepsis was defined as organ dysfunction reflected by a ≥2-point increase in the SOFA score,[3] consequent to infection.[6] Shock was defined as hypotension requiring vasopressors to maintain a mean arterial pressure ≥65 mmHg and a serum lactate concentration >18 mg/dL, despite fluid resuscitation.[6] For patients required mechanical ventilation, assist-control mode was used initially with a tidal volume 6 mL/kg ideal body weight, fraction of inspired oxygen (FiO_2) 100%, and positive end expiratory pressure 5 to 8 cmH_2O. The FiO_2 would be adjusted hourly to achieve oxygen saturation (SpO_2) 90 to 95% or partial pressure of oxygen (PaO_2) 60 to 80 mmHg. The ventilator settings within first 24 hours of ICU admission were recorded and entered for analysis. This study was conducted according to the principles of the Declaration of Helsinki and was approved by the Research Ethics Committee of Taipei Veterans General Hospital (no. 2019-09-006BC), with waiver of the requirement for informed consent.

2.2. Development of machine-learning models

We used the extreme gradient boosting (XGBoost) and random forest (RF)[7,8] ensemble methods to construct ICU mortality prediction models. XGBoost and RF are representative tree-based machine-learning methods. Training data were split into n subsets to build n trees (learners), and the results of the trees were then aggregated to generate the final results; in this way, many weak learners are incorporated to generate a strong learner. The difference between XGBoost and RF is the core algorithm. XGBoost is based on the boosting algorithm: given n subsets $\{S_1, \ldots, S_n\}$ and n trees $\{T_1, \ldots, T_n\}$, T_k is trained on S_k, and the weights of the trained T_k are passed to T_k+1; in other words, a learner receives the learning results from the previous learner. RF is based on the bagging algorithm, which centers on "voting." T_k is still trained on S_k, but the results are voted on by each T_k.

XGBoost and RF have shown remarkable performance in many classification tasks.[9–12] One characteristic of these methods is the output of feature importance, which indicates the features or attributes that are the main factors affecting the classification. This output makes machine-learning models more explainable than models generated with deep-learning methods, which have been extensively used in many applications.[13] However, only showing feature rankings is insufficient for deep analysis. Therefore, we applied SHapley Additive exPlanations (SHAP) values,[14] which reflect the positive or negative influence of each feature in addition to its rank, for in-depth analysis and plotting. Thus, the feature ranking from top to bottom on the summary plots generated represents high to low degrees of significance for classification, and the SHAP values along the x axis represent positive and negative impacts on the models.

We obtained optimal hyperparameters for the models by grid search, selecting those with the best average results of k-fold cross validation (k=5) using different scoring methods. That is, in the training phase, we split the data into five folds and then took four folds as training data; one fold as validation data. Finally, the procedure was repeated five times. Specifically, the maximum depths of RF configurations 1 and 2 were 13 and 26, and the corresponding maximum features were 16 and 19, respectively. For XGBoost, configurations 1 and 2 had the same gamma and scale_pos_weight values (5 and 2, respectively), and maximum depths of 2 and 3, respectively. Default values were used for the remaining hyperparameters. Details of the settings for the two XGBoost configurations are provided in the Additional files: http://links.lww.com/JCMA/A234.

2.3. Statistical analysis

The enrolled patients were allocated to training (n = 9901) and testing (n = 2476) datasets. The data of two datasets were compared using the Mann–Whitney U test for continuous variables (expressed as medians and interquartile ranges), and Fisher's exact test for categorical variables (expressed as counts and percentages). Areas under receiver operating characteristic curves (AUCs) were used to evaluate the accuracy of the severity scores and machine-learning models in predicting mortality of critically ill patients. The accuracy (ACC), positive predictive value (PPV), and negative predictive value (NPV) of each model were calculated. To investigate the performance of machine-learning models modified by varying conditions, we performed subgroup analyses with the cohort stratified by patients' age, APACHE II scores, SOFA scores, and the usage of mechanical ventilator. In addition, we performed logistic regression analysis to confirm the independence of clinical variables composed the machine-learning models. The five most predictive variables used in RF and XGBoost models were further adjusted in multivariate logistic regression analysis. The analyses were performed with SPSS (ver. 18.0; SPSS Inc., Chicago, IL, USA) and SAS (ver. 9.3; SAS Institute Inc., Cary, NC, USA). A p value < 0.05 was considered to indicate significance.

3. RESULTS

3.1. Sample characteristics

Of 13,187 cases screened, data from 810 patients aged <20 years or with missing labels were excluded, leaving a sample of data from 12,377 patients. A flowchart of patient enrollment

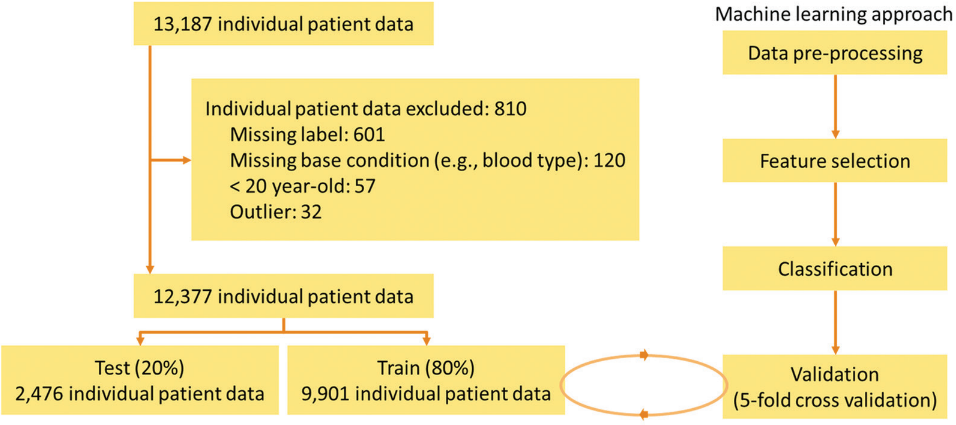
13,187 individual patient data
Individual patient data excluded: 810
Missing label: 601
Missing base condition (e.g., blood type): 120
< 20 year-old: 57
Outlier: 32
12,377 individual patient data
Test (20%)
2,476 individual patient data
Train (80%)
9,901 individual patient data
Machine learning approach
Data pre-processing
Feature selection
Classification
Validation
(5-fold cross validation)

Fig. 1 Flowchart of patient enrollment and classification.

and classification is provided as **Fig. 1**. The distributions of age, sex, comorbidities, medications administered, disease severity, vital signs, and laboratory results were similar between the training and testing datasets (**Table 1**). The median age of the enrolled patients was 70.0 years, and 7966 (64.36%) patients were male. In total, 8360 (67.54%) of the patients were admitted to the medical ICUs (ICU-A and ICU-C) and 4017 (32.46%) patients were admitted to the surgical ICU (ICU-B). In medical ICUs, the highest proportion of patients' subspecialties was the division of hematology and oncology (20.0%), followed the divisions of infectious diseases (14.9%), gastroenterology (13.3%), and nephrology (8.4%). The most common reasons for patients admitted to the medical ICUs were acute respiratory failure (44.1%), shock (21.8%), and acute renal failure (14.1%). On the other hand, the highest proportion of patients' subspecialties in the surgical ICU was the division of general surgery (31.1%), followed the divisions of colorectal surgery (15.4%), transplantation surgery (14.2%), and oral and maxillofacial surgery (8.7%). The most common reasons for patients admitted to surgical ICU were intensive care after major surgery (53.3%), acute respiratory failure (18.8%), and shock (18.1%). In the medical and surgical ICUs, 9210 (74.41%) patients were under mechanical ventilation and 3327 (26.88%) patients were under vasopressor (norepinephrine) treatment. The median APACHE II score of the study population was 23.5. Details of the sample characteristics are provided in Supplementary File 1, http://links.lww.com/JCMA/A234.

3.2. Performance of machine-learning models and traditional scoring systems

Receiver operating characteristic curves characterizing the ability of the different models to predict ICU mortality are presented in **Fig. 2**, and corresponding c-statistics are provided in **Table 2**. The XGBoost model had the greatest area under the curve (AUC 0.880, sensitivity 0.802, specificity 0.805; ACC 0.874, PPV 0.619, NPV 0.921), followed by the RF model (AUC 0.876, sensitivity 0.815, specificity 0.777; ACC 0.871, PPV 0.621, NPV 0.911). In contrast, SOFA scores (AUC 0.747, sensitivity 0.815, specificity 0.549; ACC 0.592, PPV 0.258, NPV 0.939), SAPS II scores (AUC 0.743, sensitivity 0.872, specificity 0.462; ACC 0.528, PPV 0.237, NPV 0.950), and APACHE II scores (AUC 0.738, sensitivity 0.820, specificity 0.505; ACC 0.556, PPV 0.241, NPV 0.936) had much smaller AUCs and less specificity. **Table 3** shows the 5-fold cross-validation results, and it demonstrates that XGBoost with AUC 0.900 (95% CI, 0.893-0.907), sensitivity 0.806, and specificity 0.829, which outperformed the performance of conventional scoring systems (relative increase of 18% in AUC).

3.3. Feature importance and independence of variables in the machine-learning models

The variables used in the machine-learning models and their relative importance are shown in **Figs. 3 and 4**. The hyperparameters and setting details for the two XGBoost configurations are provided in Supplementary Files 2 and 3, http://links.lww.com/JCMA/A234. Norepinephrine usage in the ICU, the FiO_2, the lowest and highest of Richmond Agitation and Sedation Scale (RASS) scores, and prothrombin time (PT) were the five most predictive variables in the RF models. Twenty-four-hour urine output, FiO_2, PT, HRs, and platelets were the five most predictive variables in the XGBoost models. Above variables were all significantly associated with of ICU mortality in the univariate logistic regression analysis. In the multivariate logistic regression analysis, norepinephrine usage, FiO_2, the highest RASS scores, PT, 24-hour urine output, HRs, and platelets were still independently associated with the incidence ICU mortality (showed in **Table 4**).

3.4. Subgroup analysis

The results of subgroup analysis were summarized in **Table 5**. Our machine-learning models were with AUC more than 0.8 among most of subgroups. However, the XGBoost models were with less predictive performance in subjects with relatively lower disease severity, such as patients with SOFA score less than 6 (AUC 0.786, sensitivity 0.813, specificity 0.554), or in patients without mechanical ventilation (AUC 0.843, sensitivity 0.800, specificity 0.738). Relative importance of the XGBoost model variables for prediction of ICU mortality in nonventilated patients was illustrated in Supplementary File 4, http://links.lww.com/JCMA/A234. The five most predictive variables among nonventilated critical patients were Glasgow Coma Scale, body weights on admission, serum sodium, total bilirubin concentrations, and HRs.

4. DISCUSSION

In this longitudinal study of data from 12,377 critically ill patients, we developed four machine-learning models to predict ICU mortality, which were much more accurate and specific than traditional severity scoring systems. The XGBoost model

Table 1

Summary of patients' baseline characteristics

	Total (n=12,377)	Training dataset (n=9901)	Testing dataset (n=2476)	*p*
(1) Age	70 (57–82)	70 (57–82)	69 (56–82)	0.256
(2) Gender (male)	7966 (64.36)	6345 (64.08)	1621 (65.47)	0.205
(3) Height	162.8 (156–169)	162.5 (156–168.9)	163 (156–169)	0.175
(4) Weight	60.2 (52–70)	60.1 (52–70)	60.25 (52–70)	0.830
(5) Renal function				
Creatinine	1.43 (0.92–2.74)	1.44 (0.92–2.78)	1.42 (0.92–2.62)	0.405
eGFR	40.6 (22–53)	40.60 (22–53)	40.6 (23.75–53)	0.431
(7) Urine output				
UO, in 8h	400 (90–890)	400 (90–900)	400 (80–850)	0.148
UO, in 24 h	1280 (550–2190)	1280 (560–2210)	1260 (510–2120)	0.069
(8) I/O balance, in 24 h	1864.43 (542.96–2795.47)	1864.43 (531–2767.8)	1864.43 (586.55–2914.12)	0.192
(9) Vital signs at ICU admission				
Highest BT	37.7 (37.2–38.3)	37.7 (37.2–38.3)	37.7 (37.2–38.3)	0.596
Heart rate	110 (96–127)	110 (96–127)	111 (96–128)	0.755
Respiratory rate	25 (22–30)	25 (22–29)	26 (22–30)	0.047
SBP	82 (70–96)	83 (70–97)	81 (70–95)	0.055
DBP	43 (35–51)	43 (34–51)	43 (35–51)	0.939
(10) Mechanical ventilation				
Use mechanical ventilator	9210 (74.41)	7392 (74.66)	1818 (73.42)	0.216
Tidal volume	547.09 (507–573)	547.09 (507–574)	547.09 (509–571)	0.581
FiO2	41.79 (35–41.79)	41.79 (30–41.79)	41.79 (35–41.79)	0.199
Rate	11.46 (11.46–12)	11.46 (11.46–12)	11.46 (11.46–12)	0.036
PEEP	6.23 (5–6.23)	6.23 (5–6.23)	6.23 (5–6.23)	0.604
Peak	23.64 (22–24)	23.64 (22–24)	23.64 (22–24)	0.372
(11) Oxygenation				
PaO2	133 (86.7–173.8)	133 (87–173.2)	131.95 (85.57–176.07)	0.874
(12) Acidosis and electrolyte				
pH	7.41 (7.38–7.46)	7.41 (7.38–7.46)	7.41 (7.37–7.46)	0.440
PaCO2	31.9 (27.1–35.8)	31.9 (27.1–35.8)	31.9 (27.2–35.9)	0.827
HCO3	21 (18.1–23.2)	21 (18.1–23.3)	21 (18.1–23.2)	0.820
(13) Hemodynamic status				
CVP	10.31 (9–10.31)	10.31 (9–10.31)	10.31 (10–10.31)	0.442
Usage of norepinephrine	3327 (26.88)	2630 (26.56)	697 (28.15)	0.116
(14) Liver function				
Total bilirubin	0.95 (0.5–2.06)	0.94 (0.5–2.06)	0.96 (0.52–2.06)	0.386
AST	53 (25–187.95)	52 (25–187.95)	56 (26–187.95)	0.133
ALT	26 (15–68)	26 (15–68)	27 (15–67)	0.158
(15) Inflammatory markers				
WBC	9900 (6500–14,200)	9900 (6500–14,200)	9900 (6500–14,100)	0.815
CRP	10.85 (5.02–13.23)	10.85 (4.93–13.12)	10.85 (5.38–13.57)	0.100
(16) Platelet (k)	164 (100–229)	164 (102–230)	163 (96–226)	0.166
(17) Albumin	2.78 (2.4–3.1)	2.78 (2.4–3.1)	2.78 (2.4–3.1)	0.443
(20) GCS	9 (4–14)	9 (4–14)	9 (4–14)	0.397
(24) Hemoglobin	10 (8.6–11.7)	10 (8.6–11.7)	9.9 (8.5–11.7)	0.487
(28) Comorbidities				
Atrial fibrillation	451 (3.64)	363 (3.67)	88 (3.55)	0.853
Coronary artery disease	2618 (21.15)	2120 (21.41)	498 (20.11)	0.161
Chronic kidney disease	2533 (20.47)	2016 (20.36)	517 (20.88)	0.578
COPD	1277 (10.32)	1026 (10.36)	251 (10.14)	0.766
Cancer	4776 (38.59)	3811 (38.49)	965 (38.97)	0.661
Cerebral vascular disease	1539 (12.43)	1211 (12.23)	328 (13.25)	0.173
Chronic liver disease	1539 (12.43)	1215 (12.27)	324 (13.09)	0.276
Diabetes mellitus	3013 (24.34)	2424 (24.48)	589 (23.79)	0.479
GI Bleeding	1477 (11.93)	1198 (12.1)	279 (11.27)	0.268
Heart failure	1313 (10.61)	1072 (10.83)	241 (9.73)	0.118
Hypertension	3998 (32.3)	3225 (32.57)	773 (31.22)	0.203
Myocardial infarction	440 (3.55)	354 (3.58)	86 (3.47)	0.851
Peptic ulcer disease	1843 (14.89)	1457 (14.72)	386 (15.59)	0.283
(30) Ward				
ICU-A	4002 (32.33)	3186 (32.18)	816 (32.96)	0.471

(Continued)

Table 1

(Continued.)

	Total (n=12,377)	Training dataset (n=9901)	Testing dataset (n=2476)	*p*
ICU-B	4017 (32.46)	3253 (32.86)	764 (30.86)	0.058
ICU-C	4358 (35.21)	3462 (34.97)	896 (36.19)	0.259
ICU mortality	1993 (16.1)	1594 (16.1)	399 (16.11)	0.977

ALT=alanine aminotransferase; AST=aspartate aminotransferase; BT=body temperature; COPD=chronic obstructive pulmonary disease; CRP=C-reactive protein; CVP=central venous pressure; DBP=diastolic blood pressure; eGFR=estimated glomerular filtration rate; FiO_2= fraction of inspired oxygen; GCS=Glasgow Coma Scale; GI=gastrointestinal; HCO_3=bicarbonate; ICU=intensive care unit; I/O=intake/output; $PaCO_2$=partial pressure of carbon dioxide; PaO_2=partial pressure of oxygen; PEEP=positive end expiratory pressure; SBP=systolic blood pressure; UO=urine output; WBC=white blood cells.

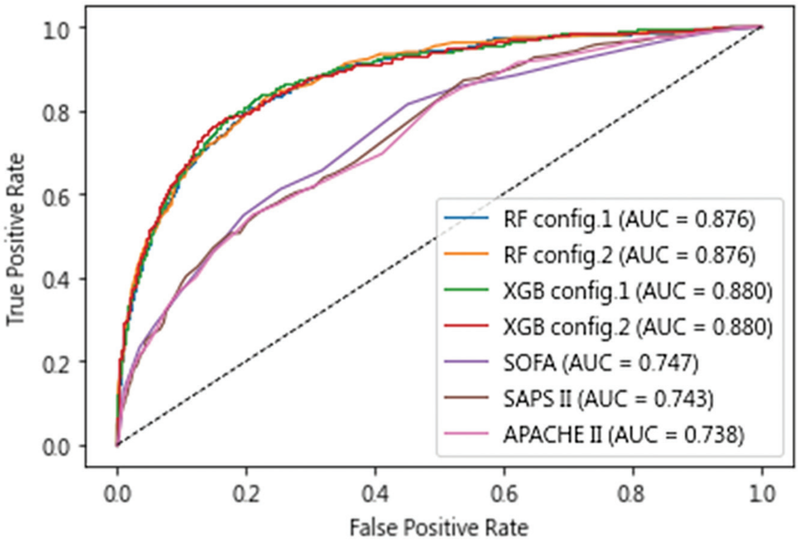

Fig. 2 Pairwise comparison of ROC curves for the models for the prediction of mortality in intensive care units. APACHE II=Acute Physiology and Chronic Health Evaluation II, AUC=areas under ROC curve, RF=random forest; ROC=receiver operating characteristic, SAPS II=Simplified Acute Physiology Score II, SOFA=Sequential Organ Failure Assessment, XGBoost=extreme gradient boosting.

Table 2

C-statistics for models for the prediction of mortality in the intensive care unit (testing database)

Method	AUC	Sensitivity	Specificity	ACC	PPV	NPV
Random forest (config. 1)	0.876	0.805	0.798	0.868	0.596	0.918
Random forest (config. 2)	0.876	0.815	0.777	0.871	0.621	0.911
XGBoost (config. 1)	0.880	0.802	0.805	0.870	0.598	0.922
XGBoost (config. 2)	0.880	0.802	0.792	0.874	0.619	0.921
SOFA score	0.747	0.815	0.549	0.592	0.258	0.939
SAPS II score	0.743	0.872	0.462	0.528	0.237	0.950
APACHE II score	0.738	0.820	0.505	0.556	0.241	0.936

ACC=accuracy; AUC=area under the ROC curve; APACHE II=Acute Physiology and Chronic Health Evaluation II; NPV=negative predictive value; PPV=positive predictive value; SAPS II=Simplified Acute Physiology Score II; SOFA=Sequential Organ Failure Assessment; XGBoost=extreme gradient boosting.

was the most accurate. The FiO_2 at the time of ICU admission was consistently most important for prediction across all models. Our results highlight the utility of machine learning for the prediction of outcomes in a large population of critically ill Asian patients.

ICU mortality prediction and disease severity stratification using traditional scoring systems[2,3] was not sufficiently accurate in all cases in this study. Certain conditions, such as diabetic ketoacidosis,[15] may generate high APACHE II scores despite not generally resulting in high degrees of mortality. Attempts to improve the accuracy of such systems would likely increase their complexity to the degree that manual calculation would be difficult.[16] With technology advancements, large-scale data analysis by machine learning provides a new solution for this dilemma. In contrast to the use of traditional descriptive statistics, the application of machine-learning techniques to clinical data stored in electronic medical records is a data-driven approach to the identification of adverse outcomes that does not require mathematical calculation or an understanding of the mechanisms involved.[17]

The predictive performance of our machine-learning models was superior to that of traditional scoring systems for the following reasons. First, we developed them using a much larger-scale dataset based on electronic medical records and data science. The original studies of the APACHE II[2] and SOFA[3] scores were conducted with data from only 5815 and

Table 3

C-statistics for models for the prediction of mortality in the intensive care unit (5-fold cross validation on training database)

Method	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	ACC (95% CI)	PPV (95% CI)	NPV (95% CI)
Random Forest (config. 1)	0.890 (0.881-0.899)	0.803 (0.802-0.804)	0.819 (0.800-0.838)	0.881 (0.872-0.890)	0.641 (0.614-0.670)	0.922 (0.916-0.929)
Random Forest (config. 2)	0.893 (0.885-0.901)	0.811 (0.805-0.817)	0.816 (0.786-0.846)	0.883 (0.876-0.889)	0.669 (0.648-0.689)	0.914 (0.910-0.920)
XGBoost (config. 1)	0.900 (0.893-0.907)	0.806 (0.801-0.811)	0.829 (0.809-0.849)	0.888 (0.883-0.894)	0.662 (0.647-0.676)	0.929 (0.922-0.935)
XGBoost (config. 2)	0.900 (0.890-0.911)	0.803 (0.802-0.804)	0.827 (0.802-0.852)	0.928 (0.922-0.934)	0.474 (0.439-0.509)	0.652 (0.636-0.669)
SOFA score	0.760 (0.749-0.771)	0.848 (0.832-0.864)	0.553 (0.541-0.565)	0.600 (0.591-0.609)	0.267 (0.263-0.271)	0.950 (0.945-0.954)
SAPS II score	0.712 (0.693-0.731)	0.833 (0.819-0.847)	0.464 (0.458-0.470)	0.524 (0.520-0.528)	0.230 (0.227-0.233)	0.935 (0.931-0.940)
APACHE II score	0.744 (0.732-0.756)	0.849 (0.833-0.865)	0.519 (0.497-0.541)	0.573 (0.556-0.589)	0.254 (0.248-0.260)	0.947 (0.943-0.951)

ACC=accuracy; AUC=area under the ROC curve; APACHE II=Acute Physiology and Chronic Health Evaluation II; NPV=negative predictive value; PPV=positive predictive value; SAPS II=Simplified Acute Physiology Score II; SOFA=Sequential Organ Failure Assessment; XGBoost=extreme gradient boosting.

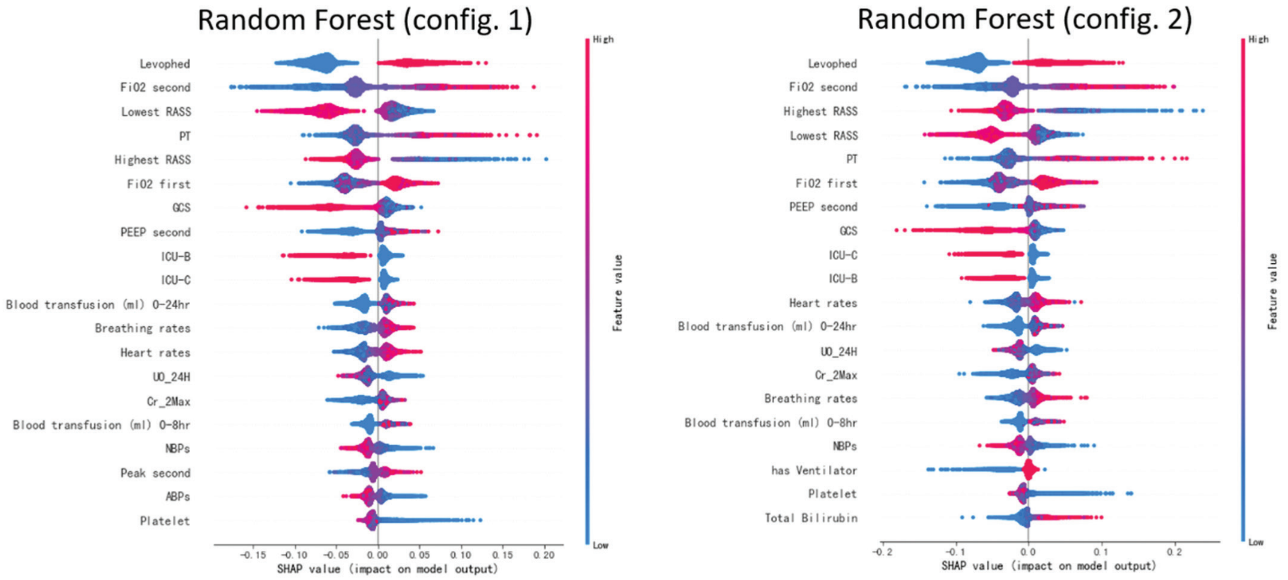

Fig. 3 Relative importance of the random forest model variables for the prediction of mortality in ICUs. ICUs=intensive care units.

1643 critically ill patients, respectively. In contrast, our models were developed with data from 12,377 patients admitted to the ICU and contained more than 140 variables, including vital signs, comorbidities, information about comedication, laboratory data, and hemodynamic parameters. In addition, the APACHE II score has been reported to be less accurate for patients with head injuries or nontraumatic intracranial hemorrhage, underestimating mortality rates in such cases.[18] Our sample included data from 4017 patients who were admitted to the surgical ICU and contained variables that are essential for the evaluation of neurological outcomes, such as Glasgow Coma Scale scores and hemorrhagic stroke. Finally, several variables included in traditional scoring systems may be altered by resuscitative therapy, leading to biased outcome prediction. The six variables that accounted for the most lead-time bias were the HR, blood pressure, respiratory rate, oxygenation, pH, and blood glucose level.[19] Our models also incorporated interpretation of the effects of comedication in the ICU. Unsurprisingly, vasopressor (norepinephrine) use was the variable most predictive of ICU mortality in the RF models.

Several other studies have examined the use of machine-learning models for the prediction of the mortality of critically ill patients.[20–23] However, most of those models were developed with Caucasian populations, and their accuracy when applied to Asian populations has not been validated. Deep-learning technology was applied to data from ICU patients in China in one study,[24] but the sample was small (clinical data from 4000 patients) and detailed information about the model features was not provided. Our machine-learning models were developed with data from a large Asian population. Furthermore, we applied SHAP values, which showed the ranking of feature importance and the positive or negative influence of each feature. Norepinephrine use, the FiO_2, and the RASS score were the most predictive variables in the RF models, whereas the 24-h urine output, FiO_2, and PT were the most predictive variables

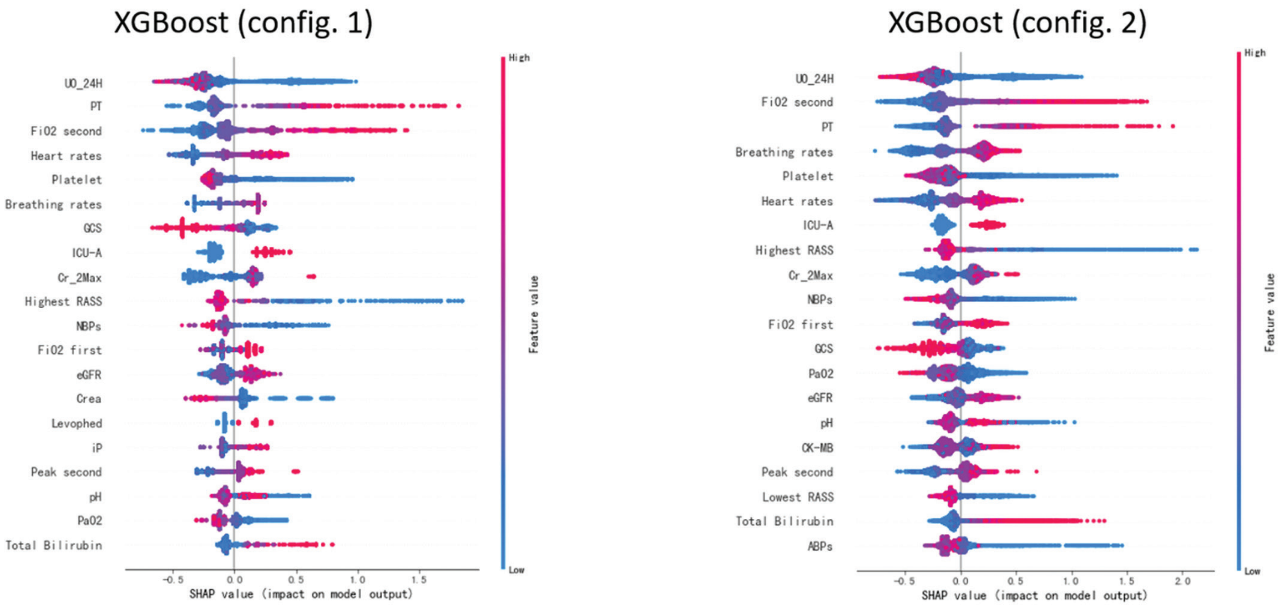

Fig. 4 Relative importance of the XGBoost model variables for the prediction of mortality in ICUs. ICUs=intensive care units, XGBoost=extreme gradient boosting.

Table 4

Univariate and multivariate logistic regression analysis to interpret the association between ICU mortality and clinical variables in machine-learning models

	Univariate		Multivariate*	
	Crude OR (95% CI)	***p***	**Adjusted OR (95% CI)**	***p***
Usage of norepinephrine	0.154 (0.083-0.288)	<0.001	0.333 (0.143-0.772)	0.010
FiO_2	1.055 (1.037-1.074)	<0.001	1.051 (1.026-1.075)	<0.001
Lowest RASS	0.493 (0.399-0.610)	<0.001	0.634 (0.474-0.847)	0.634
Highest RASS	0.605 (0.498-0.734)	<0.001	0.745 (0.565-0.981)	0.036
Prothrombin time	1.155 (1.067-1.249)	<0.001	1.102 (1.014-1.198)	0.023
UO, in 24 h	0.998 (0.998-0.999)	0.001	0.999 (0.998-1.000)	0.024
Heart rates	1.035 (1.022-1.049)	<0.001	1.019 (1.001-1.036)	0.035
Platelet	1.000 (1.000-1.000)	0.004	1.000 (1.000-1.000)	0.021
Breathing rates	1.017 (0.988-1.046)	0.248	1.011 (0.961-1.063)	0.679

* Adjusted for usage of norepinephrine, FiO2, lowest RAAS, highest RAAS, PT, UO in 24 h, heart rates, platelet, and breathing rates.

FiO_2=fraction of inspired oxygen; ICU=intensive care units; OR=Odd Ratio; RASS=Richmond agitation and sedation scale; UO=urine output.

Table 5

Subgroup analysis of machine-learning models for the prediction of mortality in the intensive care unit

	Random forest (config. 1)			Random forest (config. 2)			XGBoost (config. 1)			XGBoost (config. 2)		
Subgroup	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**
Age ≥75												
Yes (n=1016)	0.880	0.804	0.788	0.879	0.818	0.778	0.877	0.804	0.817	0.872	0.804	0.771
No (n=1460)	0.874	0.801	0.800	0.874	0.813	0.782	0.881	0.801	0.818	0.884	0.801	0.824
APACHE II≥24												
Yes (n=1133)	0.828	0.802	0.692	0.824	0.802	0.673	0.839	0.806	0.725	0.841	0.802	0.746
No (n=1343)	0.877	0.802	0.794	0.882	0.802	0.837	0.880	0.802	0.835	0.880	0.802	0.755
SOFA≥6												
Yes (n=1846)	0.872	0.804	0.767	0.869	0.804	0.763	0.873	0.801	0.790	0.872	0.801	0.792
No (n=630)	0.767	0.813	0.667	0.786	0.813	0.520	0.786	0.813	0.554	0.804	0.813	0.651
Use ventilator												
Yes (n=1818)	0.874	0.802	0.788	0.874	0.814	0.775	0.880	0.802	0.816	0.881	0.802	0.832
No (n=658)	0.846	0.800	0.791	0.835	0.800	0.781	0.843	0.800	0.738	0.837	0.800	0.688

AUC=area under the ROC curve; APACHE II=Acute Physiology and Chronic Health Evaluation II; SOFA=Sequential Organ Failure Assessment; XGBoost=extreme gradient boosting.

in the XGBoost models. These features are related to circulation failure or acute organ damage,[3] and thus are reasonably associated with the mortality of critically ill patients.

This study has several limitations. First, it was a single-center study, and the patients admitted to the ICUs of our hospital, a tertiary medical center, were relatively old and tended to have multiple comorbidities. Thus, the generalizability of our findings is limited, and external validation is needed to confirm the extrapolatory results. Second, to enable the prediction of mortality in the very early stage of ICU admission, our models included clinical data obtained within the first 24 h of patients' ICU stays. We did not use or evaluate the impact of follow-up serial data on vital signs or laboratory parameters. Finally, the data were collected over a long and potentially heterogeneous period, which may have led to bias due to changes in treatment guidelines or the improvement of care quality over time.

In conclusion, in this large-scale study, machine-learning models showed much greater accuracy and specificity than did traditional severity scores for ICU mortality prediction. The FiO_2 was the most important predictive feature across all models. Although external validation of our findings is required, our results highlight the strength and usefulness of machine learning for the prediction of outcomes for critically ill patients.

ACKNOWLEDGMENTS

This study was supported, in part, by research grants from the Ministry of Science and Technology of Taiwan (MOST 106-2314-B-350-001-MY3 and 109-2314-B-075-097-MY3); the Novel Bioengineering and Technological Approaches to Solve Two Major Health Problems in Taiwan program, sponsored by the Taiwan Ministry of Science and Technology's Academic Excellence Program (MOST 108-2633-B-009-001); the Ministry of Health and Welfare (MOHW106-TDU-B-211-113001); and Taipei Veterans General Hospital (V105C-207, V106C-045, V108C-195, V109B-010, V109D50-003-MY3-1, V110C-149, and V108D42-004-MY3-3). The funding institutions took no part in the study design, data collection or analysis, publication intent, or manuscript preparation.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data related to this article can be found at http://links.lww.com/JCMA/A234.

REFERENCES

1. Afessa B, Keegan MT. Predicting mortality in intensive care unit survivors using a subjective scoring system. *Crit Care* 2007;11:109.
2. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985;13:818–29.
3. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996;22:707–10.
4. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270:2957–63.
5. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs (Project Hope)* 2014;33:1163–70.
6. Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive Care Med* 2017;43:304–77.
7. Biau G, Scornet E. A random forest guided tour. *Test* 2016;25:197–227.
8. Chen T, Guestrin C, editors. Xgboost: a scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 2016;2016: 785–94.
9. Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans Comput Biol Bioinf* 2019;17:2131–40.
10. Binson V, Subramoniam M, Sunny Y, Mathew L. Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. *IEEE Sens J* 2021;21:20886–95.
11. Luo Y, Wang Z, Wang C. Improvement of APACHE II score system for disease severity based on XGBoost algorithm. *BMC Med Inform Decis Mak* 2021;21:1–12.
12. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart health* 2021;20:100178.
13. Bengio Y, Lecun Y, Hinton G. Deep learning for AI. *Commun ACM* 2021;64:58–65.
14. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neur IPS* 2017;30:1–10.
15. Safari S, Rahmani F, Soleimanpour H, Bakhtavar HE, Esfanjani RM. Can APACHE II score predict diabetic ketoacidosis in hyperglycemic patients presenting to emergency department? *Anesth Pain Med* 2014;4:e21365.
16. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 1—interpretation and accuracy of outcome prediction scoring systems. *Chest* 2012;141:245–52.
17. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, et al. Artificial intelligence in pathology. *J Pathol Transl Med* 2019;53:1–12.
18. Niskanen M, Kari A, Hernesniemi J, Vapalahti M, Iisalo E, Kaukinen L, et al. Contribution of non-neurologic disturbances in acute physiology to the prediction of intensive care outcome after head injury or non-traumatic intracranial haemorrhage. *Intensive Care Med* 1994;20:562–6.
19. Tunnell R, Millar B, Smith G. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care—a pilot study. *Anaesthesia* 1998;53:1045–53.
20. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Crit Care Med* 2018;46:e481–8.
21. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:1–13.
22. Holmgren G, Andersson P, Jakobsson A, Frigyesi A. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. *J Intensive Care* 2019;7:1–8.
23. Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, et al; RISC-19-ICU Investigators. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *J Intensive Med* 2021;1:110–6.
24. Li M, Chen H, Yan S, Xu X, Xu H. Application of Deep Learning Technology in Predicting the Risk of Inpatient Death in Intensive Care Unit. *J Healthc Eng* 2021;2021:6169481.