2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Interpretable Intensive Care Unit Mortality Risk Calculator

Eugene T. Y. Ang[1], Mila Nambiar[2], Yong Sheng Soh[1] and Vincent Y. F. Tan[1]

Abstract— **Mortality risk is a major concern to patients who have just been discharged from the intensive care unit (ICU). Many studies have been directed to construct machine learning models to predict such risk. Although these models are highly accurate, they are less amenable to interpretation and clinicians are typically unable to gain further insights into the patients' health conditions and the underlying factors that influence their mortality risk. In this paper, we use patients' profiles extracted from the MIMIC-III clinical database to construct risk calculators based on different machine learning techniques such as logistic regression, decision trees, random forests, k-nearest neighbors and multilayer perceptrons. We perform an extensive benchmarking study that compares the most salient features as predicted by various methods. We observe a high degree of agreement across the considered machine learning methods; in particular, age, blood urea nitrogen level and the indicator variable - whether the patient is discharged from the cardiac surgery recovery unit are commonly predicted to be the most salient features for determining patients' mortality risks. Our work has the potential to help clinicians interpret risk predictions.**

I. Introduction

Risk calculators are tools used by healthcare providers to estimate the probabilities of chronically ill or Intensive Care Unit (ICU) patients experiencing adverse clinical outcomes such as health complications, readmission to hospital, or mortality, and to then identify and screen high-risk patients. In the context of patient mortality, severity scores such as Simplified Acute Physiology Score (SAPS-II) [1] and sequential organ failure assessment (SOFA) [2] have traditionally been used to predict hospital mortality. These scoring systems predict mortality by feeding a fixed set of predictor variables, including various laboratory measurements and vitals, into simple models such as logistic regression. The simplicity of these models limits their accuracy, and benchmarking studies [3] have found that automated risk calculators based on machine learning, that train supervised learning models on Electronic Health Record (EHR) data, tend to be far more accurate.

However, a drawback of using nonparametric machine learning models is that these models tend to be less interpretable than the simple models used in traditional scoring systems, i.e. it is not always transparent what motivates their predictions. This paper investigates the problem of building a risk calculator that is not only accurate but interpretable. We focus on the task of predicting the 28-day mortality of ICU patients at discharge, but note that the observations and principles applied to this work are broadly applicable to the tasks of predicting other adverse clinical outcomes using EHR data.

There is a considerable body of literature on the problem of ICU mortality prediction. Typically, this literature uses data from the first 24 to 48 hours of a patient's stay to predict in-hospital mortality. The 2012 PhysioNet Computing in Cardiology Challenge called for machine learning solutions to address the task of predicting in-hospital mortality for patients who stayed in the ICU for at least 48 hours [4]. The winning team developed a novel Bayesian ensemble learning solution and achieved an AUC of 0.86 [5]. In the present paper, however, we study the task of 28-day mortality prediction at discharge—thus, the accuracies obtained are not directly comparable to the results in these other papers. Another work that investigates risk prediction at discharge is [6], in which the authors use long short-term memory models to predict the readmission of ICU patients within 28 days of their discharge, based on the last 48 hours of data from their stays.

A different stream of literature has looked at designing risk calculators that are not only accurate but interpretable. A number of papers [7], [8] have proposed attention neural networks, where attention mechanisms are used to identify important features while retaining the accuracy of deep neural networks. Some other works [9]–[11] have also employed methods such as Shapley values, which are also used in this paper, and local interpretable model-agnostic explanation (LIME) to enhance interpretability. More specifically, [9] combines Shapley values with XGBoost to predict the mortality of elderly ICU patients, while [10] combines Shapley values with Convolutional Neural Networks [12] to predict ICU mortality.

A key difference between our paper and the literature on interpretable risk calculators is that rather than focusing on enhancing the interpretability of any particular machine learning method, we perform a benchmarking study that compares the factors influencing the predictions made by various methods. We develop a series of risk calculators using Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (k-NN), and Multilayer Perceptron (MLP); see Bishop's book [13] for detailed descriptions of these methods. In Section IV, we extract the most influential set of variables that each calculator relies on to make predictions. We observe that all of these calculators arrive at a similar prediction and there is a high degree of commonality among the learned sets of variables. From a clinical perspective, our results are reassuring because they suggest that these data-driven methods for performing

[1]Department of Mathematics, National University of Singapore e0175081@u.nus.edu, matsys@nus.edu.sg, vtan@nus.edu.sg

[2]Institute for Infocomm Research, Agency for Science Technology and Research Milashini_Nambiar@i2r.a-star.edu.sg

prediction in clinical settings make similar conclusions, even if the precise manner in which each method arrives at a decision can be quite different.

In Section V, we show how the set of influential factors derived in Section IV may be used to guide a clinician as to how a prediction was made. However, it is important for clinicians to understand the limitations of each model when using the risk calculator, so to gain better insights to the patients' health conditions. As LR constructs linear decision boundaries to distinguish patients' class, this model would not be that effective if the training data is severely not linearly separable. Even though the DT model is highly interpretable, slight changes in the training data may result in a completely different tree. Unlike DT, RF is more stable, however, it requires more computational resources to construct and aggregate various trees. Despite its simplicity, k-NN also requires a significant amount of memory to store all the training data and to compute the prediction given a large dataset of patients' profiles. While the MLP can deal with training data that is not linearly separable, it is computationally intensive to train. It is thus suggested for clinicians to use different models according to their needs.

II. Methodology

A. Cohort Selection

The data used in this study are extracted from the MIMIC-III Critical Care Database [14], which contains the health records of patients in Beth Israel Deaconess Medical Center ICU from 2001-2012. For patients with multiple admissions, we considered every admission independently. There were a total of 61532 ICU records.

B. Feature Extraction

Our choice of features followed from studies done in the MIMIC-III research community and in MIT Critical Data [15]. In this study, our target variable was a binary flag, which indicated the patient's mortality within 28 days of their discharge from the ICU. For our input features, we selected 4 different data categories, namely, demographic, laboratory measurements, severity scores and vitals.

For demographic features, we extracted the patients' height, weight, age, ethnicity, length of ICU stay, gender and the service unit that the patients were admitted to, namely the Coronary Care Unit (CCU), the Cardiac Surgery Care Unit (CSRU), the Medical Intensive Care Unit (MICU), the Surgical Intensive Care Unit (SICU) and the Trauma/Surgical Intensive Care Unit (TSICU). We calculated the patients' BMIs by querying the patients' height and weight measurements that were last taken before the patients were discharged from the ICU.

For laboratory measurement features, we extracted blood urea nitrogen (BUN), chloride, creatinine, hemoglobin, platelet, potassium, sodium, total carbon dioxide (TotalCO2) level and white blood cells (WBC) count of the patients. As these measurements might change due to the treatment the patients received during their ICU stay, we chose to query the patients' measurements that were last taken before the patients were discharged from the ICU.

We also calculated severity scores such as the patients' SOFA score [2] and the estimated Glomerular Filtration Rate (eGFR). The latter was calculated based on the CKD-EPI Creatinine Equation [16].

For vitals features, we extracted the temperature, heart rate, blood oxygen level (SpO2), systolic blood pressure (SysBP), diastolic blood pressure (DiasBP) and mean arterial pressure (MAP). Due to the non-stationary nature of the time series of vital signs, we chose to take the median value of the time series.

C. Data Processing

Out of the 61532 records, the majority of the entries had missing height and weight features while a handful had missing features such as MAP, SpO2 and SysBP. Furthermore, there were anomalies in the certain features such as age and weight. We define an outlier as a value that is more than 3 standard deviations from the mean. We dropped records that contained these anomalies and those that had missing feature values, except for height, leaving 36330 data entries.

We then treated missing height entries as follows. As is well known, the BMI can be expressed as the ratio of weight and squared height. We used ordinary least squares linear regression to regress BMI against weight and impute the missing height entries using the derived formula

$$\widetilde{\text{BMI}} = 5.6925 + 0.2769 \times \text{weight}, \tag{1}$$

and the height features through the BMI ratio. We found a strong positive correlation between BMI and weight, with an R^2 value of 0.737, justifying that our imputation procedure is fairly accurate.

An issue that we faced with the dataset was that the number of positively labelled records (patients who died within 28 days of discharge) represented 7.6% of the total data sample, contributing to a severe class imbalance. One approach to tackle such class imbalance is Synthetic Minority Oversampling Technique (SMOTE) [17], which oversamples the examples in the minority class, by creating new examples from the minority class in the training data set. This technique chooses a random minority class instance and finds its k nearest minority class neighbors, based on the Euclidean distance. In our data processing step, we used the default setting for k, where $k = 5$, in the imblearn library. As the generated synthetic instance is a convex combination of that instance and one of its randomly chosen neighbors, it would likely contain some characteristics of the minority class due to its proximity (according to Euclidean distance).

As some of our features were categorical, we used SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) technique [17] to deal with the categorical features in the dataset. SMOTE-NC sets the categorical values of the generated data by choosing the most frequent category of the nearest neighbors during the generation process.

D. Model Selection and Prediction

After splitting our data so that 75% of the original dataset is treated as the training data and the remaining 25% as test data, we applied SMOTE-NC on the training set and trained the models 30 times.

The features were used to train LR, DT, RF, k-NN and MLP in order to determine whether a patient would survive within 28 days of discharge from the ICU. The parameters of the model were estimated using 5-fold cross-validation (CV). The hyperparameters of each model were optimised through random and grid search. Also, in this optimisation, a different 5-fold CV on the training set was performed. The tuned models were tested on the test set and performances were evaluated using the area under the receiving operating characteristic curve (AUC), test accuracy (ACC) and recall (REC). We then calculated the mean of the performance metrics across the 30 trials and used the standard deviation as the uncertainty bound. All modelling and analyses were performed with Python, and in particular the sklearn library.

III. Model Performances

In Table I, we show the performance of our trained classifiers at predicting mortality of an ICU patient in the test data set. With the exception of the Decision Tree model, all our trained classifiers attained a AUC of 0.76 or greater, and an accuracy (denoted by ACC) of 0.71 or greater. LR achieved the best performance in terms of AUC and REC, with 0.8 and 0.713 respectively, while the MLP prediction model had the best performance in terms of ACC at 0.8. Our results suggest that all trained classifiers provide fairly accurate and reasonable performance for predicting mortality.

TABLE I

Performance on predicting ICU mortality on test data set across different classifiers.

	LR	DT	RF	kNN	MLP
AUC	**0.8 ±0.008**	0.696 ± 0.02	0.764 ± 0.007	0.761 ± 0.008	0.76 ± 0.009
ACC	0.744 ± 0.004	0.675 ± 0.065	0.742 ± 0.007	0.713 ± 0.005	**0.8 ±0.007**
REC	**0.713 ±0.02**	0.608 ± 0.076	0.616 ± 0.023	0.673 ± 0.017	0.509 ± 0.025

IV. Extracting Influential Features

Next, we extract the most influential features from each trained classifier. We exhibit the high degree of overlap among the influential features across different classifiers.

A. Logistic Regression

In LR, the trained classifier predicts mortality according to the following relationship

$$P(Y=1|\mathbf{x}) = 1/\big(1+\exp\big(-(\boldsymbol{\theta}^T\mathbf{x}+\theta_0)\big)\big). \quad (2)$$

Here, $\boldsymbol{\theta}$ represents the weights of the input features, $\mathbf{x}$ represents the patient's normalized data, and θ_0 represents the offset of the decision boundary. In particular, a larger value of $\boldsymbol{\theta}^T\mathbf{x}+\theta_0$ corresponds to a higher probability of mortality. As such, we can infer the most influential features as those corresponding to the largest coefficients θ_i in magnitude. In Table II, we show the features corresponding to the five largest coefficients and the five smallest coefficients.

TABLE II

Top 5 most influential features by LR

	Results			
	Positive Influence		Negative Influence	
1	Age	0.586	CSRU	-1.176
2	BUN	0.336	TSICU	-0.444
3	Male	0.325	SICU	-0.296
4	MICU	0.285	BMI	-0.288
5	Heart Rate	0.273	Hemoglobin	-0.198

We observed from Table II that all service units except MICU had a negative influence on the mortality risk. We also observed that older patients, male patients, patients with higher blood urea nitrogen level or heart rate would tend to have a greater mortality risk within 28 days after their discharge. Similarly, patients that had higher BMI or hemoglobin level would tend to have lower mortality risk.

To increase the interpretability of the model, we penalize the logistic regression problem with an ℓ_1-norm regularizer, i.e., we solve

$$\min_{\boldsymbol{\theta},\theta_0} \sum_{t=1}^{n} \log\Big(1+\exp\big(-y_t(\boldsymbol{\theta}^T\mathbf{x}_t+\theta_0)\big)\Big) + \lambda\|\boldsymbol{\theta}\|_1. \quad (3)$$

Here, $\lambda > 0$ is the regularization parameter, which we tune via CV.

After tuning the hyperparameters with 5-fold CV and running the model 30 times, the best estimator achieved a test score of 0.72 ± 0.005, recall of 0.734 ± 0.021 and AUC of 0.797 ± 0.009. We annihilated those features which had a coefficient of zero more than half of the time [18].

TABLE III

Top 5 most influential features with non-zero coefficients in L1-penalised LR

	Results			
	Positive Influence		Negative Influence	
1	Age	0.494	CSRU	-1.362
2	BUN	0.236	BMI	-0.198
3	SOFA	0.236	Hemoglobin	-0.123
4	Heart Rate	0.22	Temperature	-0.1
5	MICU	0.101	SysBP	-0.016

We observed that 11 features were annihilated and clinicians could determine to mortality risk using the top 5 most influential features as shown in Table III. As extreme values of features such as BMI are not desirable for the patients, the logistic model could not detect this intricacy because the outcome depends *linearly* on the features. Hence, clinicians need to be aware of such limitations when applying this model.

B. Decision Tree

To obtain an interpretable model, we restricted the learned DT model to a maximum depth of 5 tiers. One approach to find out which features are more influential in a DT model

is to compute the Gini Importance of each feature. This is given as the decrease in Gini impurity weighted by the node probability where the *Gini impurity* of a certain node is

$$G(p_1, p_2) = 1 - p_1^2 - p_2^2, \tag{4}$$

where p_i is the probability of picking a data instance from class $i \in \{1, 2\}$ in that node. Hence, the larger the feature's Gini importance, the more important the feature is [19]. The top 5 features with the highest Gini importances are shown in Table IV.

TABLE IV
TOP 5 FEATURES OF THE HIGHEST GINI IMPORTANCE

Results		
	Features	Gini Importance
1	Length of stay	0.224773
2	CSRU	0.214003
3	BUN	0.174614
4	Age	0.052457
5	eGFR	0.047148

We observed from Table IV that length of stay, service unit CSRU, blood urea nitrogen level, age and estimated GFR had the highest Gini Importance. This means that most of the decisions that were made on how to split the data were mostly based on these features, and that the samples in the following nodes were relatively pure. Hence, clinicians could obtain a reasonable prediction of the risk of mortality by looking just at this smaller subset of features.

C. Random Forest

A random forest is an ensemble of decision tree models, and is hence less amenable to interpretation. Although there are many methods to compute feature importance such as entropy and the Gini impurity measure (as described in (4)), we have chosen to employ Shapley values here. This technique is versatile as it can be used to interpret outcomes from both the RF and MLP models and can explicitly state the marginal risk contribution of every feature. As such, we compute the Shapley value of each feature to infer its influence. The concept of Shapley values were first introduced in the field of cooperative game theory, and it measures the marginal contribution that each feature (player, in the context of game theory) to a subset of features (coalition), averaged across all possible subsets of features [20]. The Shapley value of i-th feature is given as

$$\phi_i(f) = \sum_{S \subseteq \{x_j\}_{j=1}^{p} \setminus \{x_i\}} \frac{|S|!(p-|S|-1)!}{p!} (f(S \cup \{x_i\}) - f(S)). \tag{5}$$

Thus $\phi_i(f)$ is the contribution of the i-th feature based on f, which calculated the economic output of all feature values in S, a subset of the features used in the model, p is the number of features and x_i represents the corresponding feature values of a data point to be explained. Intuitively, a feature has a higher Shapley value if the inclusion of this feature in the prediction model results in a greater change in predictions, as compared to the inclusion of other features.

For every data instance, the average marginal contribution of each feature is given by its Shapley value. In order to find out the influence of the feature, we can determine the SHAP feature importance by averaging the Shapley value for every features across all data instances [21].

Note that the Shapley value requires us to sum over all possible subsets of features. This becomes intractable if the number of features is moderately large. In our numerical implementations, we approximately compute the Shapley value by averaging over a small number of randomly sampled subsets of features. For the RF model, we implemented a TreeExplainer from the shap library, which the class disregarded decision paths that involved missing features instead of computing the output through a selection of the features [22]. In Fig. 1 we show the SHAP feature importance values across the top 20 features.

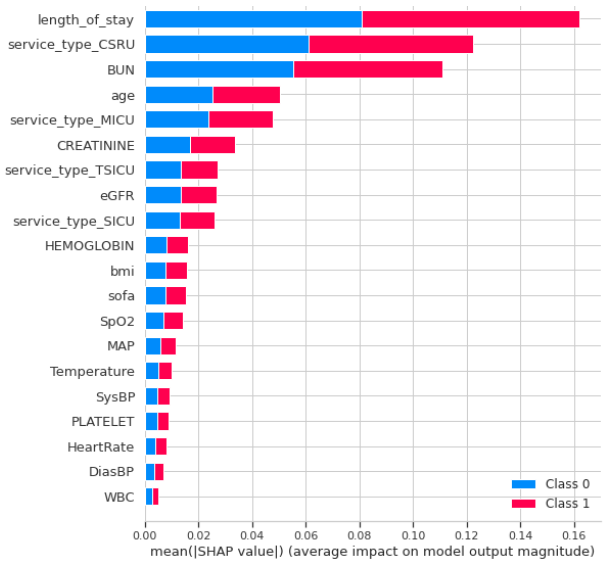

Fig. 1. SHAP feature importance for Random Forest

We observed from Fig. 1 that length of stay had the largest SHAP feature importance and was responsible for changing the predicted absolute ICU mortality risk prediction on average by around 8% (0.08 on x-axis) regardless of classes. As the larger the SHAP feature importance, the bigger the average marginal contribution to risk prediction the feature had. We could also observe that length of stay, service unit CSRU, blood urea nitrogen level, age and service unit MICU were the 5 most influential features. In fact, for many instances, clinicians could accurately deduce the final predicted risk mortality with our trained RF model with these 5 features while choosing the remaining features randomly.

D. k-Nearest Neighbors

As our implementation of k-NN uses weights trained using the RF model, the most influential variables using k-NN are precisely those found using these other methods. As such, we omit discussing k-NN.

E. Multilayer Perceptron

The MLP is a black-box model, and hence is less amenable to interpretation. Hence, we computed Shapley values using the KernelExplainer from the shap library to interpret the model [23]. In order to compute the SHAP feature importance, we had to take the average of the Shapley values of every feature across all data instances. To reduce the computational complexity, we estimated the SHAP feature importance from 500 training data instances randomly and repeated the process 30 times. We plot the SHAP feature importance values for MLP in Fig. 2.

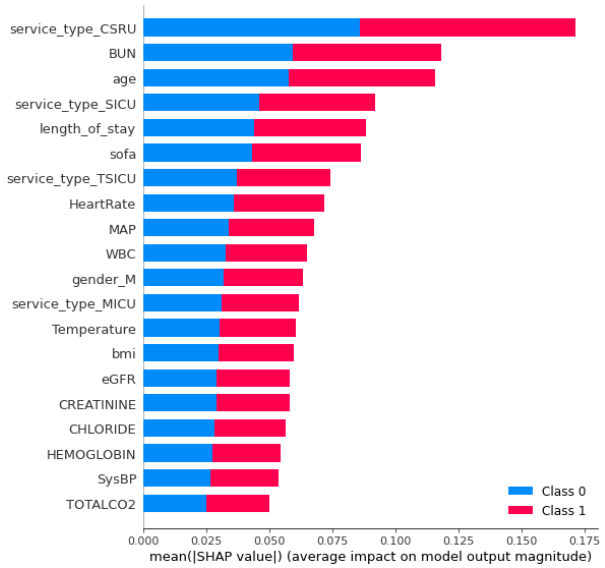

Fig. 2. SHAP feature importance for MLP

We observed from Fig. 2 that service unit, CSRU had the largest SHAP feature importance and was responsible for changing the predicted absolute ICU mortality risk prediction on average by around 8% regardless of classes. Similar to the analysis of the RF model, clinicians could accurately deduce the final predicted mortality risk by looking at the 5 most influential features, namely service unit CSRU, blood urea nitrogen level, age, service unit SICU and length of stay while choosing the remaining features randomly.

F. Comparison of Influential Features

We compare the top few most influential features extracted from the various models, and we tabulate these in Table V.

TABLE V

TOP 5 RANKED INFLUENTIAL FEATURES ACROSS DIFFERENT MODELS

	LR	DT	RF	MLP
1	CSRU	Length of stay	Length of stay	CSRU
2	Age	CSRU	CSRU	BUN
3	TSICU	BUN	BUN	Age
4	BUN	Age	Age	SICU
5	Gender (M)	eGFR	Creatinine	Length of stay

We observed a high degree of agreement in the most important features across all models; in particular, we observed that service unit CSRU, age, and blood urea nitrogen levels ranked among the top 5 most influential features across all models. These features are possible indicators for poor renal or cardiovascular functions. Despite employing different methods to make predictions, these models arrive at similar predictions and a set of common variables.

V. INTERPRETING RISK PREDICTIONS

Next, we apply the influential factors obtained in Section IV to interpret the predictions derived from our risk calculators. We focus on 4 specific test cases, namely, patients with icustay_id #223086, #271230, #237482, and #271203, from the test data set as the former two patients survived past 28 days after discharge, while the latter two patients did not survive past 28 days after discharge. These profiles represent the spectrum of patients in the dataset and serve to illustrate the utility as well as the limitation of our methods.

Logistic Regression. We recall that LR predicts the mortality risk of a patient based on the linear function $\boldsymbol{\theta}^T\mathbf{x}+\theta_0$. Hence the final prediction is primarily influenced by how far these feature values deviate from the population sample, with its importance weighted by the coefficients in $(\boldsymbol{\theta}, \theta_0)$. As such, a clinician who wishes to understand whether the LR's prediction is sound can examine how far the patient's feature values deviate from the population, with special attention on features whose coefficients have larger magnitude.

As an illustration, we show the numerical values of each patient corresponding to the influential features, as listed in Table III in our fitted LR model in Table VI.

TABLE VI

RAW FEATURE VALUES OF THE 4 EXAMPLES

Features	Range	Test Examples			
		#237482	#271203	#271230	#223086
Service Unit		CCU	SICU	MICU	CSRU
Age (years)	63.2 ± 16.2	73.7	52.2	70.53	57.04
BUN (mg/dL)	22.4 ± 15.6	28	19	**54**	13
SOFA	4.1 ± 3	7	2	**8**	5
Heart Rate (bpm)	83.9 ± 13.8	**108**	82	82	81.5
SysBP (mmHg)	119.9 ± 15.9	**100**	132.5	**138**	106
Temperature (Celsius)	36.9 ± 0.5	36.4	**37.8**	37.1	36.8
Hemoglobin (g/dL)	10.5 ± 1.7	**8.3**	12	10	9.2
BMI (kg m^{-2})	28.1 ± 6.4	32.7	27.6	**38.5**	**43.49**

We observed that Patient #237482 had a high heart rate and a low systolic blood pressure and hemoglobin level, relative to the population. These features contributed towards predicting a high mortality risk, and in this instance, the patient did not survive. We observed that Patient #271230 had a relatively high BUN, SOFA score, systolic blood pressure and BMI. The elevated BUN and SOFA score, which have positive coefficients, outweighed the elevated systolic pressure and BMI numbers, which have negative coefficients. As a result, our model predicted an overall elevated mortality risk, even though the patient did survive past the 28 days.

Decision Tree. A DT model makes a prediction by answering a sequence of queries, which checks if a particular

feature value is above or below a learned threshold. Hence, a clinician who wishes to understand if the learned DT model makes decisions in a fashion that is clinically sound can study these queries, and in particular, check if these queries can be backed by clinical expertise. The respective branches of all examples are shown in Table VII.

TABLE VII
DT BRANCHES OF PATIENTS #237482, #271203, #271230, #223086

Nodes	#237482	#271203	#271230	#223086
1	BUN $>$ 19	BUN $\leq$ 19	BUN $>$ 19	BUN $\leq$ 19
2	TSICU $\leq$ 0.5	Length of stay $\leq$ 2	TSICU $\leq$ 0.5	Length of stay $\leq$ 2
3	CSRU $\leq$ 0.5	CSRU $\leq$ 0.5	CSRU $\leq$ 0.5	CSRU $>$ 0.5
4	SpO2 $\leq$ 99.9	Age $\leq$ 52.6	SpO2 $\leq$ 99.9	-
5	Hemoglobin $\leq$ 11.9	Age $>$ 43.7	Hemoglobin $\leq$ 11.9	-

We observed from Table VII that Patients #237482 and #271230 undertook the same sequence of branches under our learned model even though they had different outcomes. One possible reason for the incorrect classification may be that the learned DT model has limited classification power because it only has 5 levels. Another reason may be that DT models are inherently limited in that they on hard decision boundaries, and cannot distinguish between features that are near or far away from a decision boundary. In some cases, the severity of a certain feature value may sometimes convey clinically relevant information.

Random Forest. As we noted in Section IV-C, RF is an ensemble method, and hence its results are less amenable to direction interpretation compared to LR and DT. As such, we rely on the SHAP explanation forces to interpret the predictions provided by the RF model. Broadly speaking, Shapley values can be viewed as "forces", which either increase or decrease the mortality risk. Each Shapley value is represented as an arrow that pushes to increase or decrease the prediction probability and the magnitude of the contribution is showed by the size of the arrow. These forces would balance each other out at the actual prediction of the data instance. The SHAP explanation force plots for the examples #237482 and #223086 are shown in Figs. 3, 4 respectively.

Fig. 3. SHAP explanation force plot for Patient #237482

We observed from Figs. 3 and 4 that the baseline probability, which was the average predicted probability, was 0.5. We observe from Fig. 3 that patient #237482 had an elevated predicted risk of 0.63. Also, the marginal risk-decreasing contribution of being in the service unit CCU is the largest, according to the size of the arrow. Although each of the influential risk-increasing factors such as not being in service unit, CSRU, 3 days in the ICU and an estimated GFR of 19.28 mL/min/1.73m^2 has less absolute marginal contribution than the service unit CCU, the combined effects of these risk-increasing factors outweighed the risk-decreasing factor and thus, leads to an overall higher risk prediction.

We observed from Fig. 4 that patient #223086 had a low predicted risk of 0.05. Also, the marginal risk-decreasing contribution of being in the service unit CSRU is the largest, according to the size of the arrow. The risk-increasing factors had negligible marginal contribution to the overall risk prediction, whereas risk-decreasing factors such as being in the service unit CSRU, 1 day in the ICU, blood urea nitrogen level of 12 mg/dL have large absolute marginal contribution, as observed from the size of the arrows. This leads to an exceptionally low overall mortality risk.

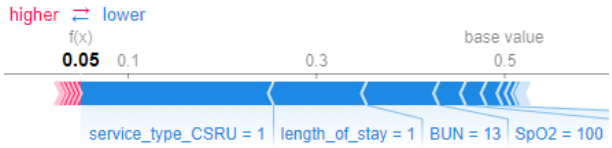

Fig. 4. SHAP explanation force plot for Patient #223086

k-Nearest Neighbors. The k-Nearest Neighbors model makes prediction based on the outcomes of the nearest neighbors. As the best estimator used 16 neighbors to make a prediction, we showed the feature values of the patient #237482 and the range of its 16 neighbors in Table VIII.

TABLE VIII
FEATURE VALUES OF PATIENT #237482 AND ITS 16 NEIGHBORS

Features	#237482	Range of 16 neighbors' feature values
Age	73.7	73.7 ± 5.67
Length of stay	3	3.96 ± 1.39
SOFA	7	5.79 ± 1.68
DiasBP	51	53.5 ± 3.38
HeartRate	108	99.3 ± 5.89
MAP	62	65.9 ± 3.41
SpO2	96	96.8 ± 0.84
SysBP	100	96.3 ± 6.58
Temperature	36.4	36.7 ± 0.18
BUN	28	37.4 ± 13.9
Chloride	92	94.5 ± 2.6
Creatinine	2.4	2.22 ± 0.447
Hemoglobin	8.3	9.37 ± 0.459
Platelet	266	223 ± 121
Potassium	4	4.27 ± 0.42
Sodium	133	134 ± 2
TotalCO2	31	31.1 ± 1.28
WBC	8.3	8.41 ± 1.66
BMI	32.7	29.8 ± 3.8
eGFR	19.3	29.8 ± 7.62
Gender	F	14 male, 2 female
Service Unit	CCU	14 CCU, 2 MICU
Label	Positive	15 positive, 1 negative

We observed from Table VIII the majority of the neighbors had a positive label, resulting in a positive prediction for test example #237482. Also, we observed that the feature values of the test sample were generally close to the range of its neighbors despite a few anomalies. This was likely due to a large range in the raw values of certain features such as platelet count. This resulted in a large disparity even though the standardized values were similar.

Multilayer Perceptron. Our analysis for MLP is based on examining SHAP explanation force plots, and is similar to our discussion regarding the RF model. The SHAP ex-

planation force plots for the example #271203 is shown in Figure 5.

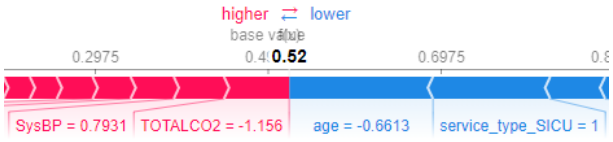

Fig. 5. SHAP explanation force plot for Patient #271203

We observed from Fig. 5 that the base value, which was the average predicted probability, was 0.5. We observe from Fig. 5 that patient #271203 had a predicted risk of 0.52. Also, the marginal risk-decreasing contribution of age, with a raw value of 52.2 is the largest, according to the size of the arrow. Although risk-decreasing factors such as age and being in the service unit SICU have large marginal contribution, these contributions are counteracted by several less influential risk-increasing factors such as the systolic blood pressure and total carbon dioxide level. Hence, the MLP model may not be able to provide a definitive prediction on this patient's mortality risk and clinicians could consider using other models such as DT or k-NN to estimate the mortality risk.

VI. Conclusion

In this study, based on patients' profiles extracted from the MIMIC-III clinical database, we developed various risk calculators to predict 28-day mortality risk of ICU patients at discharge. We have determined the common influential features that all calculators relied on to make predictions. In addition, we have developed various methods to interpret the risk predictions. On top of achieving high performance on the test data, these calculators shared a high degree of commonality among the set of influential features. This showed the effectiveness of these risk calculators in clinical settings. However, clinicians must still exercise caution due to the possible existence of confounding variables and hidden latent factors that may not be present in the features we used. As this study can be helpful in providing interpretable yet accurate predictions for clinicians, future research can be done on exploring *causal* models. This has the potential to provide clinicians with insights regarding the causal relationships between these features and thus improving the patients' health conditions by targeting the independent variables.

Acknowledgements

The authors would like to thank Dr. Pavitra Krishnaswamy for her comments on the manuscript.

References

[1] J. R. Le Gall, S. Lemeshow, and F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, Jama, vol. 270,24, 1993. pp. 2957-2963. doi:10.1001/jama.270.24.2957

[2] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining and L.G. Thijs, 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, vol. 22(7), Jul. 1996, pp 707-710. doi: 10.1007/BF01709751.

[3] S. Purushotham, C. Meng, Z. Che and Y. Liu, Benchmarking deep learning models on large healthcare datasets, Journal of Biomedical Informatics, vol. 83, 2018, pp. 112-134. doi: 10.1016/j.jbi.2018.04.007.

[4] I. Silva, G. Moody, D. J. Scott, L. A. Celi and R. G. Mark, Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012, Sept. 2012 Comput Cardiol, vol. 39, 2010, pp. 245-248.

[5] A. E. Johnson, N. Dunkley, L. Mayaud, A. Tsanas, A. A. Kramer and G. D. Clifford, Patient specific predictions in the intensive care unit using a Bayesian ensemble, 2012 Computing in Cardiology, Krakow, 2012, pp. 249-252.

[6] Y. W. Lin, Y. Zhou, F. Faghri, M. J. Shaw and R. H. Campbell, Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory, PloS one, vol. 14(7), Jul. 2019, e0218942, doi: 10.1371/journal.pone.0218942.

[7] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak and Z. Huang, Interpretable clinical prediction via attention-based neural network, BMC Medical Informatics and Decision Making, vol. 20(3), 2020, pp. 1-9.

[8] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz and W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in neural information processing systems, vol. 29, 2015, pp. 3504-3512.

[9] W. Caicedo-Torres and J. Gutierrez, ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU, Journal of Biomedical Informatics, vol. 98, 2019, doi: 10.1016/j.jbi.2019.103269.

[10] X. Lu, P. Hu, Z. Mao, P. Kuo, et al., Interpretable Machine Learning Model for Early Prediction of Mortality in Elderly Patients with Multiple Organ Dysfunction Syndrome (MODS): a Multicenter Retrospective Study and Cross Validation, arXiv:2001.10977 [physics.med-ph], Jan. 2020.

[11] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.

[12] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11): 2278—2324, 1998.

[13] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2008.

[14] A. E. W. Johnson, T. J. Pollard, L. Shen et al. MIMIC-III, a freely accessible critical care database. Scientific Data, 2016, vol. 3(1)16-35, doi: 10.1038/sdata.2016.35

[15] MIT Critical Data, Secondary Analysis of Electronic Health Records. Massachusetts Institute of Technology, Cambridge, MA, USA, 2016.

[16] A. S. Levey, L. A. Stevens, C. H. Schmid et al., A new equation to estimate glomerular filtration rate [published correction appears in Ann Intern Med. 2011 Sep 20;155(6):408], Ann Intern Med, vol. 150(9), 2009, pp. 604-612, doi: 10.7326/0003-4819-150-9-200905050-00006

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, vol. 16, Jun. 2002, pp 321-357, doi: 10.1613/jair.953

[18] N. Meinshausen, P. Buhlmann, Stability Selection, Statistical Methodology, vol. 72, Sep. 2010, pp. 417-473, doi: https://doi.org/10.1111/j.1467-9868.2010.00740.

[19] B. H. Menze, B. M. Kelm, R. Masuchm, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics, vol. 10, Jul. 2009, pp. 213, doi: 10.1186/1471-2105-10-213

[20] L. S. Shapley, A Value for n-Person Games, Contributions to the Theory of Games (AM-28), vol. 2, 1953, pp. 307-317

[21] M. A. Ahmad, C. Eckert, A. Teredesai, G. McKelvey, Interpretable Machine Learning in Healthcare, The IEEE Intelligent Informatics Bulletin, vol. 19, Aug. 2018, pp. 1-7, doi: 10.1145/3233547.3233667

[22] R. Yang, Who dies from COVID-19? Post-hoc explanations of mortality prediction models using coalitional game theory, surrogate trees, and partial dependence plots, medRxiv. Preprint at https://doi.org/10.1101/2020.06.07.20124933 (2020).

[23] S. M., Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, In Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774, arXiv: 1705.07874