Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Early identification of ICU patients at risk of complications: Regularization based on robustness and stability of explanations

Tiago Amador [a,d], Saulo Saturnino [a,b], Adriano Veloso [a,c,*], Nivio Ziviani [a,c]

[a] Universidade Federal de Minas Gerais, Brazil
[b] Lifecenter Hospital, Brazil
[c] CIIA–Health, Innovation Center on Artificial Intelligence for Health, Brazil
[d] Universidade Estadual de Feira de Santana, Brazil

ARTICLE INFO

Keywords:
Machine learning
Medical data science
Model explainability
Regularization

ABSTRACT

The aim of this study is to build machine learning models to predict severe complications using administrative and clinical elements that are collected immediately after patient admission to the intensive care unit (ICU). Risk models are of increasing importance in the ICU setting. However, they generally present the black-box issue because they do not provide meaningful information about the logic involved in patient-specific predictions. Fortunately, effective algorithms exist for explaining black-box models, and in practice, they offer valuable explanations for model predictions. These explanations are becoming essential to engender trust and accreditation to the model. However, once the model is implemented, a major issue is whether it will continue to employ the same prediction logic as originally intended to. To build our models, features are obtained from patient administrative data, laboratory results and vital signs available within the first hour after ICU admission. This enables our models to provide great anticipation because complications can occur at any moment during ICU stay. To build models that continue to work as originally designed we first propose to measure (i) how the provided explanations vary for different inputs (that is, robustness), and (ii) how the provided explanations change with models built from different patient sub-populations (that is, stability). Second, we employ these measures as regularization terms that are coupled with a feature selection procedure such that the final model provides predictions with more robust and stable explanations. Experiments were conducted on a dataset containing 6000 ICU admissions of 5474 patients. Results obtained on an external validation cohort of 1069 patients with 1086 ICU admissions showed that selecting features based on robustness led to gains in terms of predictive power that varied from 6.8% to 9.4%, whereas selecting features based on stability led to gains that varied from 7.2% to 11.5%, depending on the target complication. Our results are of practical importance as our models predict complications with great anticipation, thus facilitating timely and protective interventions.

1. Introduction

Intensive care units (ICUs) provide the facilities, resources and specialized personnel to the comprehensive management of patients having or at risk of developing life-threatening organ dysfunctions [1]. Treating the evident physiologic disorder is only the first step in the care of ICU patients because they are subject to many complications arising from the etiology of the disorder, as well from the advanced therapy that is required [2]. Important complications during ICU stay, such as infections, delirium, myopathies, neuropathies and nutritional disturbances have consistently been shown to be associated with long-term quality of life impairments [3] and death [4]. Thus, the first motivation for this study is to predict the occurrence of complications at the ICU as early as possible, providing the opportunity for early intervention, and to potentially achieving better outcomes.

As ICUs are data-rich environments where multiple signals are continuously monitored, machine learning models have already been developed to identify patients who are at increased risk of complications [5–11]. Whereas these models typically excel at capturing complex relationships between patient signals [12,13], they suffer from the black-box issue [14], that is, the mechanism by which signals are combined to convert inputs into outputs is opaque [15]. Complex black-box models for an early prediction of complications raise certain concerns because medical decisions in the ICU may have life or death consequences [16].

* Corresponding author at: Universidade Federal de Minas Gerais, Brazil.
E-mail address: adrianov@dcc.ufmg.br (A. Veloso).

https://doi.org/10.1016/j.artmed.2022.102283
Received 27 July 2021; Received in revised form 14 March 2022; Accepted 17 March 2022
Available online 22 March 2022

Thus, the prediction mechanism of these models must be shown to be in accordance with real clinical knowledge before the model is implemented [17–19]. Here we have an additional motivation: provide reliable and explainable predictions to optimize clinician's confidence in using these insights into their decision-making process. As we know, lack of confidence is a major issue in deploying AI-based solutions in the intensive care field.

Algorithms that aim at explaining black-box models are being increasingly used to understand the prediction mechanism [20,21]. These algorithms provide a mechanistic understanding of the functioning of the model by revealing the way features are jointly related to form the final prediction [22]. In ICU data modeling, the importance of a signal is calculated by considering many forms of interactions involving the signal. This type of explanation can mitigate the issue with black-box models because one can design or select a model that associates appropriate importance to the signals, thus providing a trustful reason for model predictions [15,23].

However, once the model is implemented, a major concern is whether it will continue to work as originally intended to [24]–that is, does the prediction mechanism remain in accordance with clinical knowledge? Trusting the provided explanations implies trusting both the model's prediction mechanism and the data that it was built from [25]. ICU data is complex and the resulting model may have potential flaws because the training data is incomplete, that is, it has not exhausted all the ways in which signals can interact.

It is important to mention that in the early prediction setting, required therapies and interventions cannot be included in the model as features because they will occur only after the prediction is provided, but they may change the likely outcome of complications. Consequently, explanations may differ greatly with small variations in the training set [26]. Moreover, the implications of over-trusting an ICU risk model can be disastrous because a flawed (but trusted) model, may falsely reassure a clinician, possibly leading to a systematic change in clinical practice [27,28].

Thus, the most important motivation for our study is to assure the model's consistency and maximizing predictive performance, providing means to verify whether the model will continue to employ the same prediction mechanism as originally intended, assuming the early prediction setting. For that, we investigate robustness and stability of explanations over different patient sub-populations as follows:

- First, to quantify the robustness of explanations we compute the similarity matrices for the inputs and their corresponding explanations, and then calculate the spatial auto-correlation between these two similarity matrices. Thus, correlation captures the key aspect of robustness, namely the degree to which variations in inputs and explanations are related. Similar to [29], we argue that explanations should be robust, that is, similar inputs should not lead to substantially different explanations.
- Second, we quantify the stability of explanations by comparing the provided explanations when the model is trained in different bootstrapped patient sub-populations. We argue that explanations should be stable, that is, small variations in the training data should not lead to substantially different explanations.
- Third, we show the existence of high performance models with high values of robustness and stability of explanations. Thus, we propose to use measures of robustness and stability as regularization terms to select more effective ICU risk models. Regularization is the process of adding information to solve an ill-posed problem or to prevent overfitting. Specifically, we use regularization terms to select features to compose the model iteratively such that the final model provides robust and stable explanations without hurting its prediction performance. We show that regularized models are not only effective in terms of prediction performance, but they also maintain a similar prediction mechanism on future data.

The remainder of this paper is organized as follows. In Section 2, we discuss relevant related work. In Section 3, we present the data, features used to build models, and model development. In Section 4, we discuss the robustness and stability of explanations. In Section 5, we present experimental results. In Section 6, we present concluding remarks as well as directions for future research.

2. Related work

There is vast literature on the use of machine learning methods for predicting events during ICU stay. These methods have been designed primarily to support medical experts in their clinical decision-making process. Systematic reviews and discussions are present in [6,9,11,18,30].

Studies differ mainly on the machine learning method employed and on the targeted complications being considered. In [31], the authors developed a decision tree model to early identify COVID-19 patients at risk for thromboembolic complications. In [10], the authors employed machine learning methods to predict renal failure, bleeding and mortality during critical care in real-time after cardiothoracic surgery. In [32], the authors used machine learning methods to early identify patients at risk of circulatory failure with a considerably lower false-alarm rate than that in conventional threshold-based systems. In [33], the authors evaluated the performance of machine learning methods in the prediction of 30-day post-surgical mortality. In [34], the authors developed a machine learning method to predict hypotension during ICU stay. In [35], the authors evaluated the performance of machine learning models in predicting one-year mortality at hospital admission. In [36], the authors employed machine learning methods to predict morbid conditions.

The black-box issue has been recently considered as an important aspect of machine learning methods that predict complications in ICU patients [37]. The lack of interpretability of machine learning models is being increasingly associated with insufficient clinical confidence on the models and little approval from the physicians [38]. Authors in [39] employed a deep learning method that produced an interpretable model for an early prediction of sepsis. In general, using machine learning models in a more realistic scenario and adding techniques that deal with the black-box issue to a certain extent have led to stronger and individualized risk prediction [40].

However, there is room for improvement when it comes to tools to excel model's performance, properly addressing the black-box issues, expanding the scope to a general approach, and improving physician's confidence. These are the main goals of this study, which are complementary to the recent literature. Typically, machine learning methods are designed to predict a specific complication. Thus, our study significantly differs from the aforementioned studies because we exploit the link between explanation and prediction to develop a more general regularization concept that can improve the prediction of diverse complications.

3. Materials and methods

Our models were trained on data from patient data, laboratory results and vital signs available within the first hour after ICU admission. This implies that predictions of labels (that is, complications) are available within the first hour after ICU admission, but they may occur anytime during ICU stay. We randomly built one million models that employed a different subset of features, as explained in Section 4.3. A Shapley additive explanations algorithm [41] was applied to the models to obtain the importance of the features driving patient-specific predictions. We considered a diverse set of possible complications to predict, including delirium, central line-associated bloodstream infection (CLABSI), ventilator-associated pneumonia (VAP) and mortality. Finally, for each model, we employed robustness and stability of explanations as regularization terms to estimate the best performing

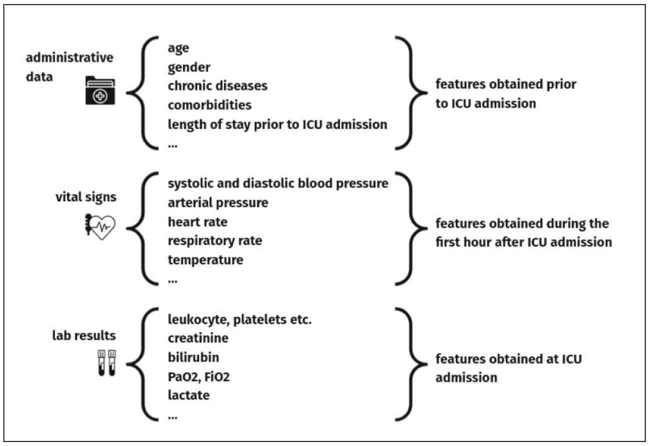
administrative data
age
gender
chronic diseases
comorbidities
length of stay prior to ICU admission
...
features obtained prior to ICU admission
vital signs
systolic and diastolic blood pressure
arterial pressure
heart rate
respiratory rate
temperature
...
features obtained during the first hour after ICU admission
lab results
leukocyte, platelets etc.
creatinine
bilirubin
PaO2, FiO2
lactate
...
features obtained at ICU admission

Fig. 1. Subset of the features used to learn the models.

Table 1
Complications during ICU stay. Number between parentheses indicates the fraction of cases with patient death.

	Development	Validation
	(n = 6,000)	(n = 1,086)
Delirium	147 (10%)	89 (11%)
Ventilator-associated Pneumonia	39 (31%)	8 (50%)
Central Line Blood Infection	11 (27%)	4 (50%)
Mortality	414	63

models. Prediction performance was measured by the area under the receiver operating characteristic curve (AUROC).

The remainder of this section is organized as follows. In Section 3.1, we present the data used for model development. In Section 3.2, we present the features and the labels used to build the models. In Section 3.3, we present the algorithm to predict the occurrence of complications in ICU admissions. In Section 3.4, we present details on the Shapley algorithm.

3.1. Data

For model development, we retrospectively collected data from patients admitted to three mixed medical and surgical ICUs of the Life-Center Hospital in Belo Horizonte, Brazil, covering the period from July-31-2016 to April-30-2019. All data records were anonymized. The study was approved by the appropriate local institutional review boards (LifeCenter CAAE: 89812318.5.0000.5126). To adhere to the SAPS III model, both the development and validation cohorts were composed of patients older than 16 years with an ICU stay lasting more than 24 h.

The models were internally validated using multiple rounds of cross-validation on 6000 admissions and 5474 patients covering the period from July-31-2016 to December-31-2018. For model validation, we employed data collected from patients admitted to the same three ICUs, but covering the period from January-01-2019 to April-30-2019 (thus simulating future data). The resulting validation cohort contained 1086 ICU admissions and 1069 patients. The main objective of validation was to demonstrate that regularized models maintained a similar prediction mechanism on future data as it was during model training.

3.2. Features and labels

Fig. 1 shows a subset of the features used to build our models. The features are a mix of static information extracted from data comprising demographics and diagnoses (obtained prior to ICU admission), daily laboratory results (obtained prior to ICU admission), and vital high-frequency signals collected from ICU equipment (during the first hour after ICU admission).

The features used to build the models were available during the first hour of ICU admission. Briefly, our working data were a subset of the SAPS III model; therefore, we could link our model to current clinical practice.

Table 1 lists the labels for training our models that includes complications that can occur at any time during ICU stay.

These complications are detailed as follows:

- Delirium is defined as a rapid change in consciousness (hours to days) characterized by reduced environmental awareness, decreased attention and altered cognition. These clinical features can manifest themselves as memory deficits, disorientation, hallucinations, fluctuating levels of alertness, and motor abnormalities.

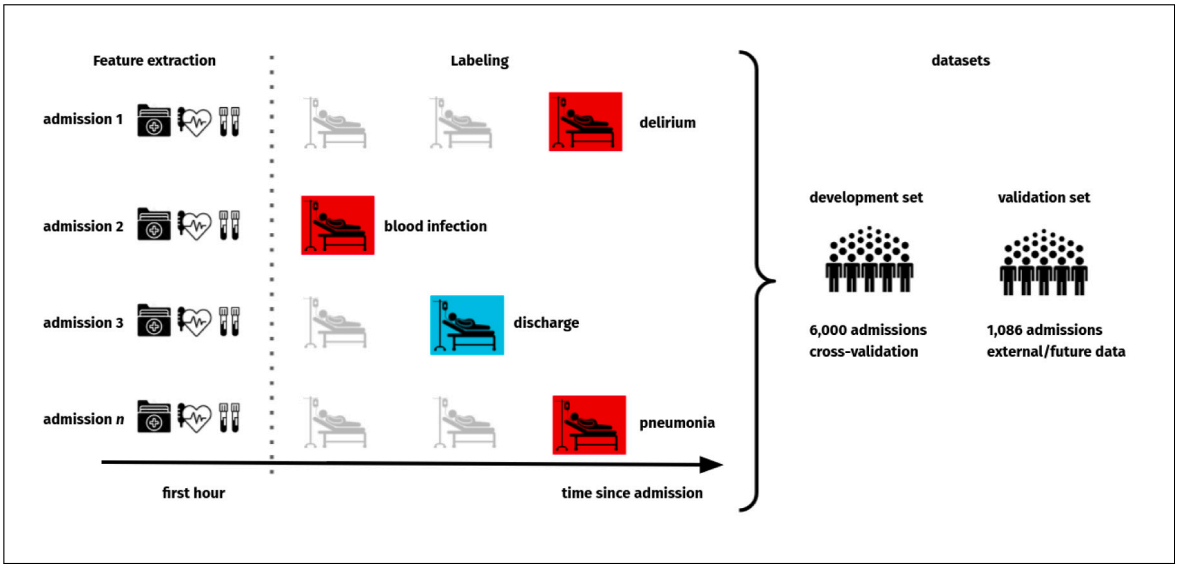

Fig. 2. Early prediction setting: features are available within the first hour after ICU admission, but labels (that is, complications) may occur anytime during ICU stay. Data is divided into development and validation sets to build and evaluate the prediction performance of the models.

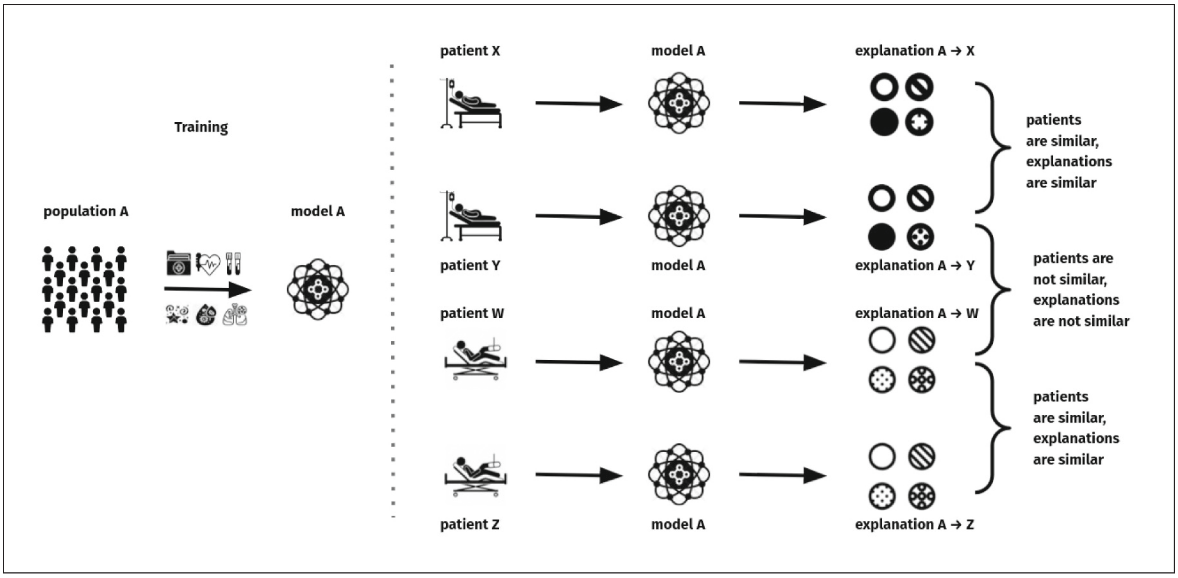

Fig. 3. Similar inputs (i.e., patients) are given to the model, and if the corresponding explanations are similar, the robustness of the set of features is high.

- A central line is a catheter that is placed into a patient's large vein, typically in the neck, chest, arms or groin. The central line is often used to draw blood, or to give fluids and medications more easily. The line can be left in place for several days if required. Sometimes, bacteria or other germs can enter the patient's central line and into their bloodstream. This can cause an infection called central line-associated bloodstream infection (CLABSI).
- Ventilator-associated pneumonia (VAP) is a type of lung infection that occurs in patients who are on mechanical ventilation breathing machines in the ICU. It is a major source of increased illness and death.

Fig. 2 depicts feature extraction and labeling. Our focus was on the early prediction setting, that is, predictions are performed almost immediately after ICU admission (more precisely, during the first hour of admission). Thus, we employed only vital signals measured during the first hour after ICU admission. These signals were aggregated (down-sampling), resulting in minimum and maximum values for each signal.

3.3. Model development

We applied an additive boosting algorithm to predict the occurrence of the aforementioned complications in ICU admissions. We used the LightGBM implementation [42], which follows the gradient boosting technique that fits boosted decision trees by minimizing a gradient error. Trees are added iteratively to the ensemble and are fit to correct the prediction errors made by prior decision trees. The ensemble minimizes the cross-entropy loss function using gradient descent. LightGBM provides hyperparameters that must be tuned, including the number of trees to compose the ensemble (T), learning rate (γ), and maximum tree depth (θ).

We sampled several models by randomly selecting k features from the set of available features with $2 \geq k \geq 25$. For each set of features we built models by using combinations of T (10, 50, 100), γ (0.05, 0.1, 0.2), and θ (5,36). Therefore, a specific subset of features resulted in 18 different models. We repeated the random exploration of feature-sets such that roughly 1,000,000 models were produced for each target complication.

3.4. Explanations

We applied a Shapley additive explanations (SHAP) algorithm to each model to obtain the importance of the features that drive patient-specific predictions [41]. SHAP is a model-agnostic representation of the importance of each feature on a particular prediction that is represented using Shapley values inspired by cooperative game theory. Given the current set of feature values, a Shapley value quantifies how much a single feature in the context of its interaction with other features contributes to the difference between the actual prediction and the mean prediction. That is, the sum of the Shapley values for all features plus the mean prediction equals the actual prediction.

Importantly, the Shapley value for a feature should not be considered as its direct and isolated effect, but as its compound effect when interacting with the other features. Shapley values consider all possible predictions for an instance using all possible combinations of inputs, and because of this exhaustive approach, SHAP can guarantee properties like consistency and local accuracy [41]. In summary, SHAP provides a vector of approximated Shapley values for each input, also known as an explanation vector. SHAP explanation vectors have the same dimension of the inputs and each value in an explanation vector indicates the importance of the corresponding feature in a particular prediction.

4. Regularization based on robustness and stability of explanations

In this section, we define the concepts of robustness and stability of explanations and the approaches we proposed to measure them. We also discuss the method of employing these concepts for model regularization. Finally, we present a method for feature selection that employs the concepts of stability and robustness of explanations. The main objective of using stability and robustness as regularization terms when learning models is to avoid models that change their mechanism of prediction in future data.

The remainder of this section is organized as follows. In Section 4.1, we introduce the robustness of explanations. In Section 4.2, we introduce the stability of explanations. In Section 4.3, we discuss ways to sample the model space to select the features that compose a model regularized in terms of robustness or stability.

4.1. Robustness of explanations (α)

Robustness measures the extent to which similar inputs have similar explanation vectors. To measure the robustness of explanations we first create a similarity matrix from the inputs (that is, patients). Then, we

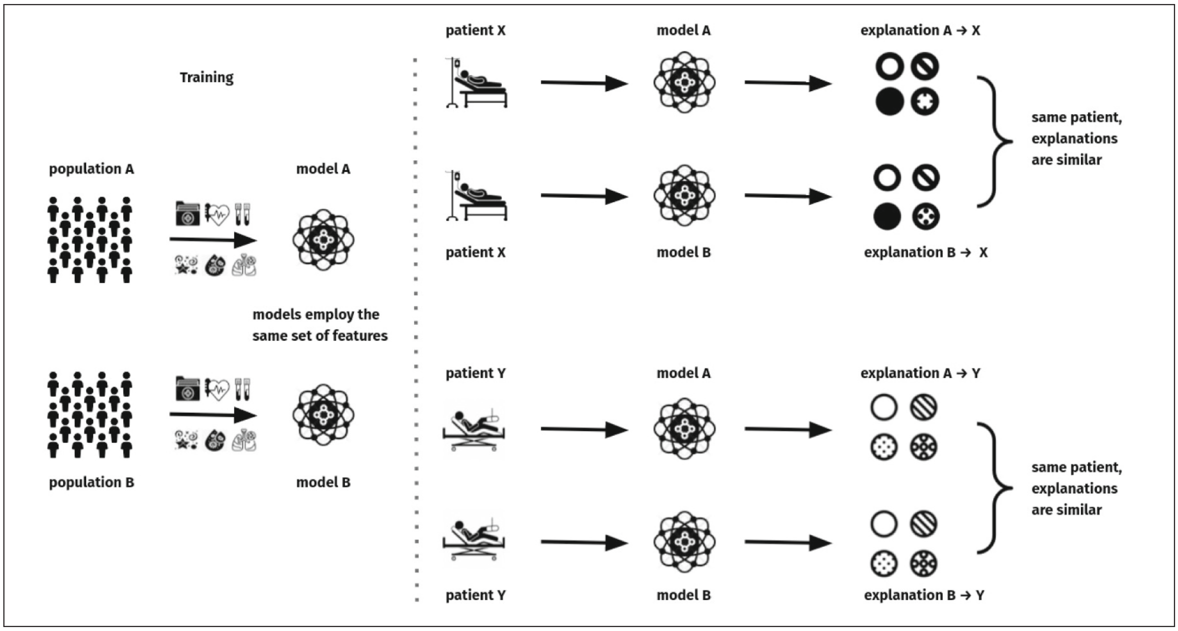

Fig. 4. Models *A* and *B* employ the same set of features but are trained on two slightly different populations *A* and *B*. Then, the same inputs (i.e., patients) are given to models *A* and *B*, and if the corresponding explanations are similar, then the stability of the set of features is high.

create a parallel similarity matrix from the corresponding explanation vectors. For both matrices, similarity is defined in terms of the Euclidean distance. Once the two parallel similarity matrices are created, we employ the Mantel *r* coefficient [43], which gives the spatial autocorrelation between two similarity matrices. It is non-parametric and computes the significance of the correlation through permutations of the rows and columns of one similarity matrix. The statistic test is the Pearson product−moment correlation coefficient *r* that falls in the range of −1 to +1, where close to −1 indicates a strong negative correlation (that is, explanations are not robust) and close to +1 indicates strong positive correlation. An *r* value of 0 indicates no correlation between inputs and explanation vectors. Fig. 3 illustrates the assessment of robustness.

4.2. Stability of explanations (β)

Stability measures the extent to which the same input leads to similar explanation vectors when it is given to models built from training sets with small variations. To measure the stability of explanations we perform multiple bootstrap rounds (that is, sample the training data uniformly with replacement), and in each round we create a similarity matrix from the obtained explanation vectors. The bootstrapping process results in multiple similarity matrices with one matrix per round. They are then compared with the similarity matrix from the explanation vectors obtained from the original training set (that is, the baseline matrix). Comparison between bootstrapped matrices and the baseline matrix is defined in terms of the Mantel *r* coefficient. Because we have multiple matrices, we simply calculate the average *r* value. High average values of *r* indicate that explanations are stable with variations in the training set, whereas low average values of *r* indicate that explanations vary greatly with small variations in the training set. Fig. 4 illustrates the assessment if stability.

4.3. Sampling the model space by selecting features that compose a model

We sample the model space by selecting the features that compose a model. We begin by enumerating all possible models composed of a single feature. Next, we select the feature within the best performing

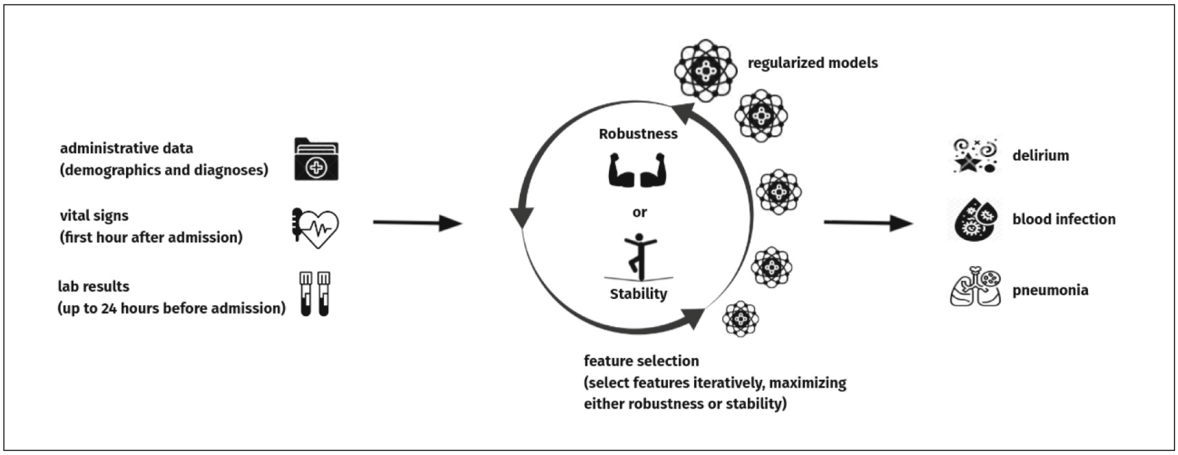
administrative data
(demographics and diagnoses)
vital signs
(first hour after admission)
lab results
(up to 24 hours before admission)
regularized models
Robustness
or
Stability
feature selection
(select features iteratively, maximizing either robustness or stability)
delirium
blood infection
pneumonia

Fig. 5. The proposed regularization process. Features are selected iteratively by maximizing either model robustness or stability. The final model is used to predict the complications.

Table 2
Some characteristics of the ICU patients in the development and validation datasets.

	Development	Validation
	(n=6,000)	(n=1,086)
Age, years	65 (54 – 80)	65 (53 – 80)
Sex		
Female	3074 (51.2%)	562 (51.7%)
Male	2926 (48.8%)	524 (48.3%)
Comorbidities		
Aids	32 (0.005%)	7 (0.007%)
Hypertension	3543 (59.0%)	532 (52.1%)
Diabetes	919 (15.3%)	196 (19.5%)
Cardiac arrhythmia	304 (5.1%)	27 (2.6%)
Dementia	528 (8.8%)	145 (14.2%)
Morbid obesity	373 (6.2%)	59 (5.6%)
Cancer therapy	212 (3.5%)	90 (8.6%)
Admission category		
Medical	3275 (54.6%)	664 (61.1%)
Scheduled surgery	2158 (36.0%)	272 (25.0%)
Unscheduled surgery	553 (9.4%)	103 (9.5%)
Type of surgery		
Cardiac surgery	53 (0.9%)	1 (0.1%)
Trauma	42 (0.7%)	3 (0.3%)
Neuro surgery	24 (0.4%)	1 (0.1%)
Length of stay before ICU, days	2.76 (0–1)	2.00 (0–1)
Length of ICU stay, days	3.87 (1–4)	4.08 (2–5)
Number of ICU admissions		
1	5034 (92.1%)	928 (96.5%)
2	361 (6.6%)	31 (3.2%)
3	51 (0.9%)	2 (0.02%)
≥4	19 (0.4%)	1 (0.01%)

model according to a specific utility criterion. Then, we enumerate all possible models composed of two features. The model enumeration process continues by including one feature after each iteration, until the cross-validation performance of the model starts to degrade. We ensure that no feature appears more than once within the same model.

When sampling the model space, we want to discourage learning a model f, for which explanations are neither robust nor stable, to avoid the risk of over-trusting the model's mechanism of prediction. Therefore we define utilities $U_\alpha(f)$ and $U_\beta(f)$ of a model f as the following functions:

$$U_\alpha(f) = P + C \times \alpha, \text{with } C \geq 0, \alpha > 0 \tag{1}$$

$$U_\beta(f) = P + C \times \beta, \text{with } C \geq 0, \beta > 0 \tag{2}$$

where P represents prediction performance measure, C denotes the regularization coefficient, α denotes the robustness of explanations and β represents the stability of explanations. Lower values of C will encourage sampling models with apparently high prediction performance irrespective of the robustness and stability of the provided explanations. In contrast, higher values of C may lead to models with low prediction performance. Appropriate values of C may lead to highly performing models with robust and stable explanations, and therefore these models are more likely to employ a prediction mechanism that is similar to the expected one. We disregard any model with negative robustness or negative stability.

Fig. 5 presents an overview of the proposed regularization process. As explained earlier, features obtained from patient data, vital signs and laboratory results are available within the first hour after ICU admission. Therefore, predictions of labels (that is, complications) are available within the first hour after ICU admission, but they may occur anytime during ICU stay. Next, features are selected iteratively by maximizing either model robustness or stability until the best performing model is selected. Finally, the final model is used for the early identification of patients at risk of complications.

5. Experimental results

In this section, we present our evaluation procedure and then report results for the early identification of patients at risk of complications. In particular, our experiments aim to answer the following research questions:

RQ1: Is there a trade-off between robustness and stability of explanations and predictive performance?

RQ2: Can we build improved models using robustness and stability of explanations as regularization terms?

The remainder of this section is organized as follows. In Section 5.1, we present the patient data available at the time of ICU admission. In Section 5.2, we present the metric used to assess the prediction performance. In Section 5.3, we discuss how we validate the development of the models. In Section 5.3, we present experimental results and discussion.

5.1. Datasets

Table 2 presents the characteristics of the patients in the development and validation datasets using data from their ICU admission. In the development dataset, the median age was 65 years with interquartile range (IQR) of 54 – 80 and 3074 (51.2%) were female.

5.2. Model performance

We assessed the prediction performance in terms of AUROC. AUROC is considered a balanced performance measure, rendering it useful for imbalanced datasets [44], and it measures the ability of the model to correctly rank positive from negative cases. It takes values between 0.5 (random predictions) and 1 (all positive cases are ranked higher than negative cases).

5.3. Setup

We conducted five-fold cross-validation using the development dataset containing 6,000 inputs to build the models. The dataset was arranged into five folds, each containing 1,200 inputs. At each run, four folds were used as training set (4,800 inputs) and the remaining fold was used as the test set (1,200 inputs). We reported the average AUROC value for the five runs. Robustness and stability of explanations were also averaged over the five folds. This entire process was executed separately for each label (that is, complications), namely delirium, VAP, CLABSI, and mortality.

We also evaluated the performance of our models on a separated validation cohort dataset (1,086 inputs) as an independent dataset (that is, future data), in which we evaluated models built from the development cohort. All experiments were run on an Intel i7-3770K 3.50 GHz processor.

We provided comparison with popular algorithms, namely XGBoost [45], random forest [46], and logistic regression [47]. To ensure a fair comparison, for all algorithms we used the SKLearn implementation,[1] and the parameter settings used were in line with typical values used for them. Specifically, for random forest we set the number of decision trees to 100, minimum samples to split on at an internal node to 2, and maximum number of leaf nodes is unlimited. For XGBoost we set the learning rate to 0.1, number of trees to 100, and minimum samples to split on at an internal node to 2. For logistic regression we set the maximum number of iterations taken for the solvers to converge to 100. We performed an initial evaluation, which showed that these parameter settings usually lead to the best results for each algorithm. All the experiments were conducted on a commodity PC with Intel(R) Core(TM) i5–4460 CPU, running GNU/Linux at 3.40Ghz and holding 8GB of RAM.

[1] https://scikit-learn.org/.

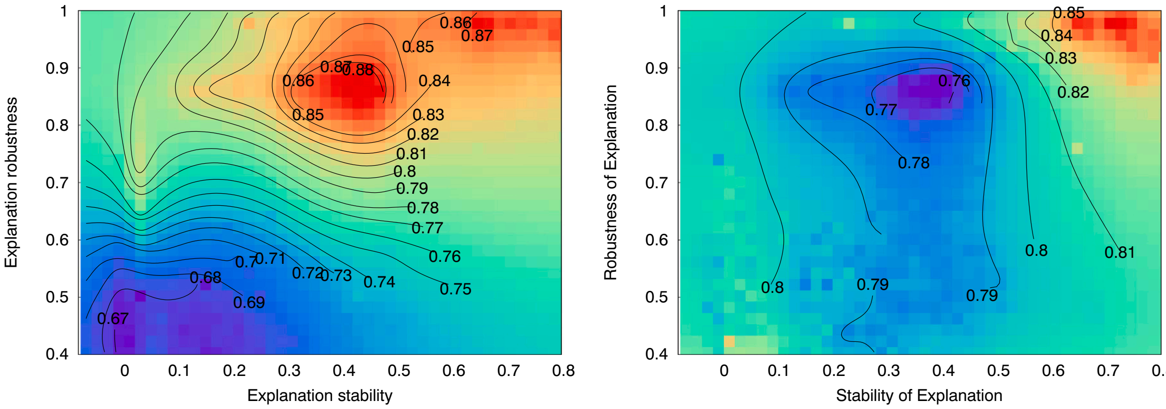

Fig. 6. Heatmap for models predicting Delirium. The color indicates the distribution of AUROC values for these models: blue is associated with relatively lower AUROC values, while red is associated with relatively higher AUROC values. Left – Cross-validation performance. Right – Performance on the validation dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

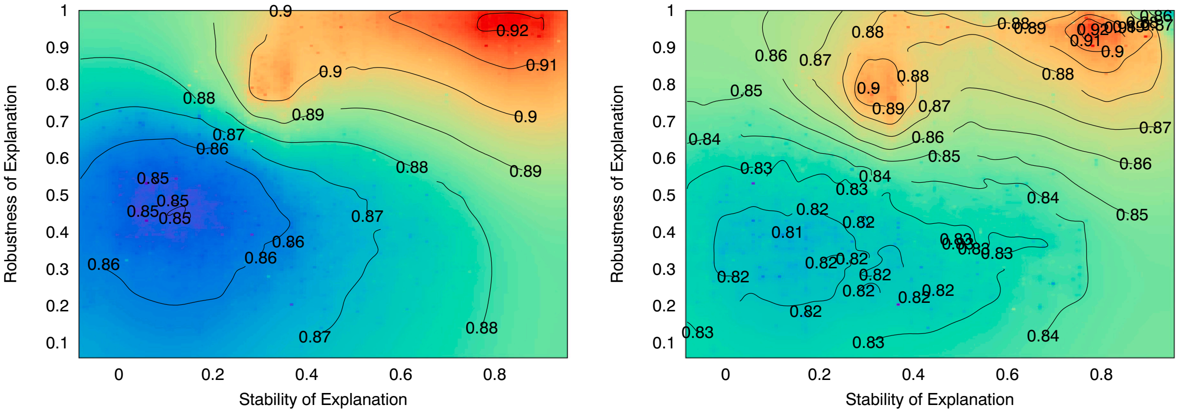

Fig. 7. Heatmap for models predicting VAP. The color indicates the distribution of AUROC values for these models: blue is associated with relatively lower AUROC values, while red is associated with relatively higher AUROC values. Left – Cross-validation performance. Right – Performance on the validation dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

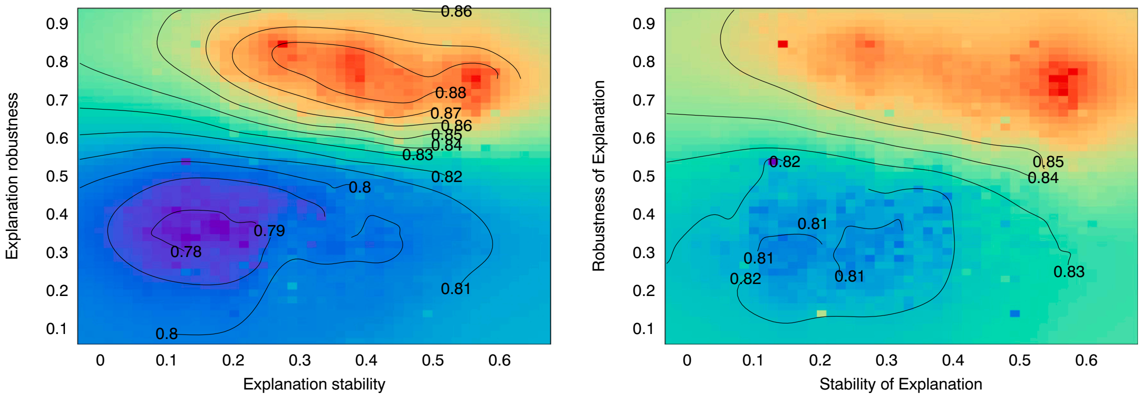

Fig. 8. Heatmap for models predicting CLABSI. The color indicates the distribution of AUROC values for these models: blue is associated with relatively lower AUROC values, while red is associated with relatively higher AUROC values. Left – Cross-validation performance. Right – Performance on the validation dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

With these settings, all considered algorithms perform the training procedures within few seconds. More precisely, the longest training time for XGBoost was 41 s, for Random Forest it was 71 s, and for Logistic Regression it was only 29 s.

5.4. Results and discussion

Our experiments evaluated the relationship between stability and robustness and the employment of these measures to provide improved prediction performance. To answer RQ1 we sampled the model space (one million models for each targeted complication) to grasp the

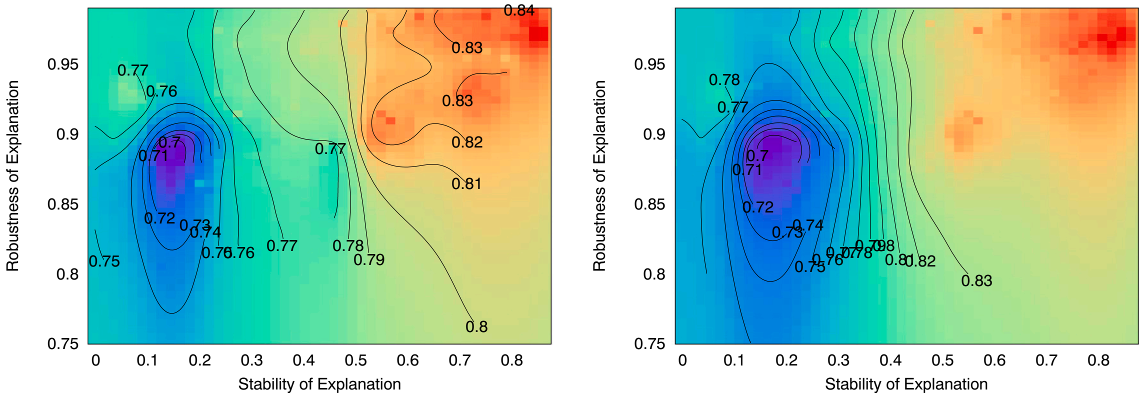

Fig. 9. Heatmap for models predicting mortality. The color indicates the distribution of AUROC values for these models: blue is associated with relatively lower AUROC values, while red is associated with relatively higher AUROC values. Left – Cross-validation performance. Right – Performance on the validation dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Models predicting delirium – Predictive performance with varying values of *C*.

C	Stability		Robustness	
	Development	Validation	Development	Validation
0.0	0.79 ↓ (20)	0.77 ↓ (20)	0.79 ↓ (20)	0.77 ↓ (20)
0.1	0.82 ↑ (29)	0.82 ↑ (30)	0.80 ↕ (0)	0.79 ↑ (20)
0.2	0.84 ↑ (30)	0.84 ↑ (27)	0.82 ↑ (29)	0.80 ↑ (29)
0.3	0.86 ↑ (27)	0.85 ↑ (35)	0.84 ↑ (30)	0.82 ↑ (30)
0.4	0.82 ↑ (29)	0.83 ↑ (5)	0.83 ↑ (23)	0.81 ↑ (23)
0.5	0.82 ↑ (29)	0.80 ↑ (29)	0.81 ↑ (20)	0.80 ↑ (29)

Table 4
Models predicting VAP – Predictive performance with varying values of *C*.

C	Stability		Robustness	
	Development	Validation	Development	Validation
0.0	0.86 ↓ (29)	0.84 ↓ (29)	0.86 ↓ (29)	↓ 0.84 (2)
0.1	0.88 ↕ (0)	0.86 ↕ (0)	0.86 ↓ (29)	↓ 0.84 (2)
0.2	0.90 ↑ (29)	0.90 ↑ (30)	0.88 ↕ (0)	↑ 0.87 (1)
0.3	0.92 ↑ (30)	0.90 ↑ (30)	0.90 ↑ (29)	↑ 0.88 (2)
0.4	0.90 ↑ (29)	0.90 ↑ (30)	0.89 ↑ (20)	↑ 0.88 (2)
0.5	0.88 ↕ (0)	0.88 ↑ (29)	0.88 ↕ (0)	↕ 0.86 (0)

relationship between robustness and stability of explanations and predictive performance. Fig. 6 shows the distribution of these quantities among models trained to predict delirium. Clearly, predictive performance increased with both robustness and stability of explanations. The best performing models are the ones located in the top right corner in each fig.

A comparison of the two heatmaps, as shown in Fig. 6, indicated differences in AUROC values when the stability of explanations ranged from 0.3 to 0.5 and robustness of explanations ranged from 0.8 to 0.9. Still, predictive performance achieved the highest values when robustness and stability of explanations were higher. Specifically, AUROC values were as high as 0.88 with cross-validation using the development dataset, and as high as 0.85 on the validation dataset (that is, future data).

Figs. 7, 8 and 9 present similar trends when analyzing models trained to predict the other targeted complications, namely VAP, CLABSI and mortality, respectively. In all cases, predictive performance increased with both robustness and stability of explanations. In summary, models predicting VAP had AUROC values as high as 0.92 with cross-validation and also on the external validation dataset. Models predicting CLABSI had AUROC values as high as 0.88 with cross-validation and 0.85 on the external validation dataset. Finally, models predicting mortality had AUROC values as high as 0.84 with cross-validation and 0.83 on the external validation dataset.

To answer RQ2, we sampled the model space as described in Section 3.3. Specifically, we iteratively selected the features to compose the model. At each iteration, the selected feature is the one that provides maximum utility. By varying regularization coefficient *C* we control the importance that the regularization terms will have in the feature selection process. For *C*=0, utility is given solely in terms of a predictive performance measure (namely, AUROC), that is, features are incorporated into the model by simply maximizing AUROC. As *C* increases, the features being selected tend to increase robustness and stability of explanations of the resulting model. We varied regularization coefficient *C* from 0 to 0.5, and we reported the results in terms of AUROC. Further, we considered the method introduced in [48] to provide baseline comparison. It employs logistic regression for model fitting using all features available. Before fitting, continuous features are standardized (zero mean, one s.d.).

Table 5
Models predicting CLABSI – Predictive performance with varying values of *C*.

C	Stability		Robustness	
	Development	Validation	Development	Validation
0.0	0.82 ↓ (29)	0.79 ↓ (23)	0.82 ↓ (29)	0.79 ↓ (23)
0.1	0.83 ↓ (20)	0.81 ↓ (20)	0.84 ↕ (0)	0.82 ↕ (0)
0.2	0.86 ↑ (29)	0.84 ↑ (29)	0.86 ↑ (29)	0.84 ↑ (29)
0.3	0.88 ↑ (30)	0.86 ↑ (30)	0.88 ↑ (30)	0.85 ↑ (23)
0.4	0.87 ↑ (23)	0.85 ↑ (23)	0.88 ↑ (30)	0.85 ↑ (23)
0.5	0.85 ↑ (20)	0.85 ↑ (23)	0.85 ↑ (20)	0.83 ↑ (20)

Table 3 shows performance values of models predicting delirium. Up arrows indicate an improvement in performance when compared with that in [48], whereas down arrows indicate the opposite trend. In both cases, the number within parentheses shows the absolute difference in terms of AUROC. Clearly, selecting features based solely on AUROC (that is, *C*=0.0) does not lead to the best models. Instead, the proposed

Table 6
Models predicting mortality – Predictive performance with varying values of *C*.

C	Stability		Robustness	
	Development	Validation	Development	Validation
0.0	0.81 ↓ (20)	0.77 ↓ (23)	0.81 ↓ (20)	0.77 ↓ (23)
0.1	0.82 ↕ (0)	0.80 ↕ (0)	0.82 ↕ (0)	0.80 ↕ (0)
0.2	0.83 ↑ (20)	0.82 ↑ (29)	0.84 ↑ (29)	0.82 ↑ (29)
0.3	0.86 ↑ (30)	0.85 ↑ (5)	0.86 ↑ (30)	0.84 ↑ (30)
0.4	0.84 ↑ (29)	0.82 ↑ (29)	0.84 ↑ (29)	0.84 ↑ (30)
0.5	0.81 ↓ (20)	0.80 ↕ (0)	0.80 ↓ (29)	0.78 ↓ (29)

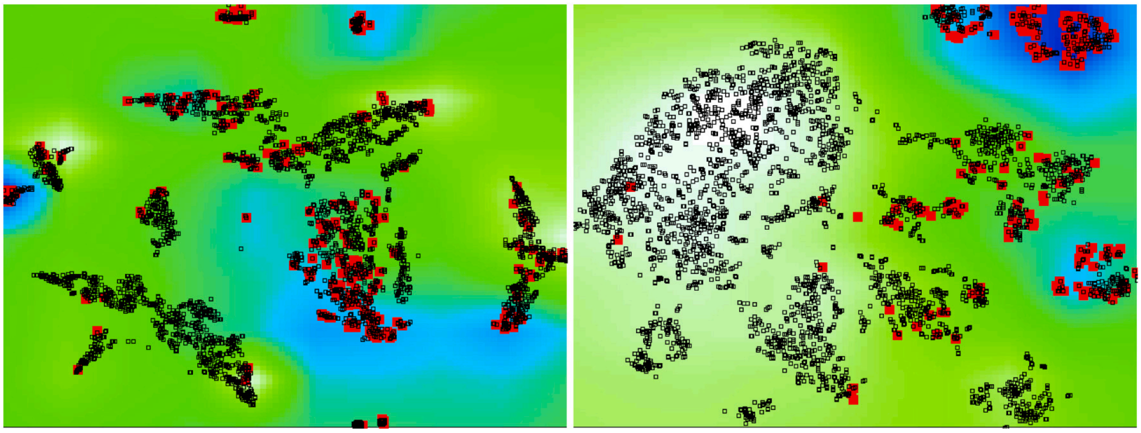

Fig. 10. Two models trained to predict mortality. Left – Utility of the model is given solely by AUROC. Right – Utility of the model is given by Eq. 1 (stability of explanations).

Table 7
Best regularized models versus baselines XGBoost, Random Forest and Logistic Regression. α and β values indicate the best models in terms of robustness and stability, respectively.

Target complication		Gains over XGBoost	Rand Forest	LogReg
Robustness				
Delirium	(α=0.3)	6.8%	8.2%	10.3%
VAP	(α=0.3)	4.9%	5.0%	5.5%
CLABSI	(α=0.3)	7.9%	8.2%	8.8%
Mortality	(α=0.3)	9.4%	9.2%	9.5%
Stability				
Delirium	(β=0.3)	11.1%	11.5%	12.1%
VAP	(β=0.2)	7.4%	7.2%	7.6%
CLABSI	(β=0.3)	8.8%	8.8%	9.2%
Mortality	(β=0.3)	10.2%	10.0%	10.7%

regularization terms play an important role in feature selection. Best results in terms of AUROC were obtained with C=0.3. Notably, performance on the development and validation datasets became closer as C increased, suggesting that both robustness and stability of explanations were useful for model generalization. Further, the performance obtained with C=0.3 greatly surpassed the baseline prediction performance in terms of AUROC.

Tables 4, 5 and 6 present the same analyses for models trained to predict VAP, CLABSI, and mortality, respectively. The same trend was observed when predicting these complications, that is, the best results in terms of AUROC were obtained using moderate values of C. Higher values may select features that improve explanation stability and robustness, but are weak in terms of increasing AUROC values.

Clearly, the proposed regularization terms were highly effective in selecting features that produce models with high generalization for all complications considered. Finally, it is important to note that explanation stability and robustness are highly effective regularizers even when being used within gradient boosting trees that employ other types of regularization, such as L1 and L2.

To better understand the impact of using robustness and stability of explanations as regularizers when selecting features, we plotted a heatmap showing the decision boundaries for models predicting mortality. Basically, we represented each point (that is, a patient) using the corresponding feature values, and then we used t-SNE [49] to visualize the data in two dimensions. Fig. 10 (Left) shows the decision boundary for the best model obtained with C=0, and Fig. 10 (Right) shows the best models obtained with C=0.3. Points in red correspond to patients who died during ICU stay. Interestingly, features selected using C=0.2 produced a model that was substantially more homogeneous, that is, it improved the separability of the different outcomes.

Finally, Table 7 presents the comparison between regularized models (that is, stability or robustness) and popular machine learning methods. All models were trained using the development set and evaluated using the validation set. The choice for α and β was performed using cross-validation on the development set. Generally, regularization based on stability results in higher gains over the baselines. Interestingly, a recent systematic review has shown that, in a general population, no performance benefit of complex machine learning algorithms over logistic regression is observed for clinical predictions [50] however, in an ICU population, a significant performance difference is observed.

6. Conclusions and future research

Complications in the ICU increase mortality and cost and are associated with long-term consequences. Therefore, the ability to predict, as early as possible, the risk for major complications and patient deterioration is of paramount importance. For instance, there are bundles to prevent VAP and CLABSI that can be strengthened in patients identified as at high risk for these conditions; environment modifications are the best actions to prevent delirium, and together with presence of family members are effective measures that can be implemented early. All of these three conditions are associated with mortality, which can be reduced if they are prevented.

For a large cohort of patients in critical care, we developed and validated a machine-learning algorithm that uses clinical and demographics data to forecast the risk for major complications and death within the first hour after admission in the ICU. Improved predictive performance was obtained by using novel regularization terms that were proposed to be used within a feature selection procedure. Features selected using the proposed regularization terms produced models with better separability and improved prediction performance when compared against popular baselines. The proposed regularization terms forced the selection of models based on the stability and robustness of prediction explanations, and thus regularized models had the additional advantage of continuing to employ the same prediction logic on future data.

In addition to the theoretical contributions of our models, practical advantages may arise from avoiding further suffering, because VAP, CLABSI, and delirium are major factors that can increase ICU length of stay, and effective early prediction system would certainly enable improved actions to prevent complications in the ICU, as stated above. Future research goals include the development of models for dynamic prediction of ICU length of stay and mortality, as well as models to indicate the necessary resources to treat specific subgroups of patients. It is true that the use of AI-based solutions in practice nowadays is scarce.

In this way, improving model's performance based on regularization methods – whilst we include novel features coming from the patient evolution in ICU – and enhancing clinical confidence could offer an evolution in further steps to implement these tools in the process of clinical decision-making.

CRediT authorship contribution statement

Tiago Amador: Conceptualization of this study, Methodology, Software. **Saulo Saturnino:** Data curation, Writing - Original draft preparation. **Adriano Veloso:** Methodology, Writing - Original draft preparation. **Nivio Ziviani:** Writing - Original draft preparation.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

References

[1] Marshall J, Bosco L, Adhikari N, Connolly B, Diaz J, Dorman T, Fowler R, Meyfroidt G, Nakagawa S, Pelosi P, Vincent J, Vollman K, Zimmerman J. What is an intensive care unit? A report of the task force of the world federation of societies of intensive and critical care medicine. J Crit Care 2017:270–6.

[2] To K, Napolitano L. Common complications in the critically ill patient. SurgClinNAm 2012;92:1519–57.

[3] Pandharipande P, Girard T, Jackson J, Morandi A, Thompson J, Pun B, Brummel N, Hughes C, Vasilevskis E, Shintani A, Moons K, Geevarghese S, Canonico A, Hopkins R, Bernard G, Dittus R, <collab>for the BRAIN-ICU Study Investigators EE. Long-term cognitive impairment after critical illness. New Engl J Med 2013;369:1306–16.

[4] Wollschlager C, Conrad A. Common complications in critically ill patients. Dis Mon 1988;34:225–93.

[5] Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, Jorm L. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. Sci Rep 2020;10:1–10.

[6] Gutierrez G. Artificial intelligence in the intensive care unit. Crit Care 2020;24. 101-101.

[7] Huddar V, Desiraju B, Rajani V, Bhattacharya S, Roy S, Reddy C. Predicting complications in critical care using heterogeneous clinical data. IEEE Access 2016; 4:7988–8001.

[8] Johnson A, Ghassemi M, Nemati S, Niehaus K, Clifton D, Clifford G. Machine learning and decision support in critical care. In: Proc. of IEEE; 2016. p. 444–66.

[9] Kamio T, Van T, Masamune K. Use of machine-learning approaches to predict clinical deterioration in critically ill patients: a systematic review. IntJMedResHealth Sci 2017;6:1–7.

[10] Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann S, Stamm C, Hofmann T, Falk V, Eickhoff C. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018;6: 905–14.

[11] Shillan D, Sterne J, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care 2019;23: 284.

[12] Docherty A, Lone N. Exploiting big data for critical care research. Curr Opin Crit Care 2015;21:467–72.

[13] Murdoch T, Detsky A. The inevitable application of big data to health care. JAMA 2013;309:1351–2.

[14] Editorial. Opening the black box of machine learning. Lancet RespirMed 2018;6: 801.

[15] Thorsen-Meyer H, Nielsen A, Nielsen A, Kaas-Hansen B, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2020;2(4):e179–91.

[16] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): towards medical XAI. CoRR 2019;abs/1907.07374.

[17] Delahanty R, Kaufman D, Jones S. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. Crit Care Med 2018;29:6.

[18] Hanson C, Marshall B. Artificial intelligence applications in the intensive care unit. Crit Care Med 2001;29:427–35.

[19] Nigri E, Ziviani N, Cappabianco F, Antunes A, Veloso A. Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease. In: Proc. of International Joint Conference on Neural Networks; 2020. p. 1–8.

[20] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60.

[21] Samek W, Müller K. Towards explainable artificial intelligence. In: Explainable AI: Interpreting, Explaining And Visualizing Deep Learning. of Lecture Notes in Computer Science, 11700. Springer; 2019. p. 5–22.

[22] Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proc. of the Conference on Fairness, Accountability, and Transparency; 2019. p. 279–88.

[23] Alves T, Laender A, Veloso A, Ziviani N. Dynamic prediction of ICU mortality risk using domain adaptation. In: Proc.of IEEE International Conference on Big Data; 2018. p. 1328–36.

[24] Pirracchio R, Cohen M, Malenica I, Cohen J, Chambaz A, Cannesson M, Lee C, Resche-Rigon M, Hubbard A, Group AR. Big data and targeted machine learning in action to assist medical decision in the ICU. Anaesth Crit Care Pain Med 2019;38: 377–84.

[25] Valle D, Pimentel T, Veloso A. Assessing the reliability of visual explanations of deep models with adversarial perturbations. In: Proc. of International Joint Conference on Neural Networks; 2020. p. 1–8.

[26] Ghorbani A, Zou J. Data shapley: equitable valuation of data for machine learning. In: Proc.of International Conference on Machine Learning; 2019. p. 2242–51.

[27] Beam A, Kohane I. Big data and machine learning in health care. JAMA 2018;319: 1317–8.

[28] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New EnglJMed 2019;380:1347–58.

[29] Alvarez-Melis D, Jaakkola T. On the robustness of interpretability methods. CoRR 2018;abs/1806.08049.

[30] Bailly S, Meyfroidt G, Timsit F. What is new in ICU in 2050: big data and machine learning. Intensive Care Med 2018;44:1524–7.

[31] de Sande DV, Genderen MV, Rosman B, Diether M, Endeman H, den Akker JV, Ludwig M, Huiskens J, Gommers D, Bommel JV. Predicting thromboembolic complications in COVID-19 ICU patients using machine learning. JClinTransRes 2020;6:179–86.

[32] Hyland S, Faltys M, Huser M, Lyu X, Gumbsch T, Esteban C, Bock C, Horn M, Moor M, Rieck B, Zimmermann M, Bodenham D, Borgwardt K, Ratsch G, Merz T. Early prediction of circulatory failure in the intensive care unit using machine learning. NatMed 2020:364–73.

[33] Chiew C, Liu N, Wong T, Sim Y, Abdullah H. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg 2020;272:1133–9.

[34] Cherifa M, Interian Y, Blet A, Resche-Rigon M, Pirracchio R. The physiological deep learner: first application of multitask deep learning to predict hypotension in critically ill patients. ArtifIntellMed 2021;118:102118.

[35] Blanes-Selva V, Ruiz-García V, Tortajada S, Benedí J, Valdivieso B, García-Gómez J. Design of 1-year mortality forecast at hospital admission: a machine learning approach. Health InformJ 2021;27:146045822098758.

[36] Carrasco-Ribelles L, Pardo-Mas JR, Tortajada S, Sáez C, Valdivieso B, García-Gómez J. Predicting morbidity by local similarities in multi-scale patient trajectories. JBiomedInform 2021;120:103837.

[37] Cruz H, Pfahringer B, Martensen T, Schneider F, Meyer A, Böttinger E, Schapranow M. Using interpretability approaches to update "black-box" clinical prediction models: an external validation study in nephrology. ArtifIntellMed 2021;111:101982.

[38] Valente F, Henriques J, Paredes S, Rocha T, de Carvalho P, Morais J. A new approach for interpretability and reliability in clinical risk prediction: acute coronary syndrome scenario. ArtifIntellMed 2021;117:102113.

[39] Shashikumar S, Josef C, Sharma A, Nemati S. DeepAISE - an interpretable and recurrent neural survival model for early prediction of sepsis. ArtifIntellMed 2021; 113:102036.

[40] Delgado R, Núñez-González JD, Yébenes JC, Lavado A. Survival in the intensive care unit: a prognosis model based on Bayesian classifiers. ArtifIntellMed 2021; 115:102054.

[41] Lundberg S, Lee S. A unified approach to interpreting model predictions. In: Proc.of Annual Conference on Neural Information Processing Systems; 2017. p. 4765–74.

[42] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. Lightgbm: a highly efficient gradient boosting decision tree. In: Proc. of Annual Conference on Neural Information Processing Systems; 2017. p. 3146–54.

[43] Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27:209–20.

[44] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PloS One 2017;12:e0177678.

[45] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. of KDD; 2016. p. 785–94.

[46] Breiman L. Random forests. MachLearn 2001;45:5–32.

[47] Cox D. The regression analysis of binary sequences. J R Stat Soc B Methodol 1958; 20:215–32.

[48] Hyland S, Faltys M, Huser M, Lyu X, Gumbsch T, Esteban C, Bock C, Horn M, Moor M, Rieck B, Zimmermann M, Bodenham D, Borgwardt K, Ratsch G, Merz T. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020;26:364–73.

[49] van der Maaten L, Hinton G. Visualizing data using t-SNE. JMachLearnRes 2008;9: 2579–605.

[50] Christodoulou E, Ma J, Collins G, Steyerberg E, Verbakela J, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110: 12–22.