Artificial Intelligence In Medicine 151 (2024) 102862

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Two-step interpretable modeling of ICU-AIs

G. Lancia [a,*], M.R.J. Varkila [b], O.L. Cremer [b], C. Spitoni [a]

[a] Mathematics Department, Utrecht University, Budapestlaan, 6, Utrecht, 3584CD, The Netherlands
[b] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, Utrecht, 3584 CG, The Netherlands

ARTICLE INFO

Keywords:
Landmarking approach
Convolutional neural networks
Dynamic prediction
ICU acquired infections
Saliency maps

ABSTRACT

We present a novel methodology for integrating *high resolution* longitudinal data with the dynamic prediction capabilities of survival models. The aim is two-fold: to improve the predictive power while maintaining the interpretability of the models. To go beyond the *black box* paradigm of artificial neural networks, we propose a parsimonious and robust semi-parametric approach (i.e., a landmarking competing risks model) that combines routinely collected *low-resolution* data with predictive features extracted from a convolutional neural network, that was trained on *high resolution* time-dependent information. We then use *saliency maps* to analyze and explain the extra predictive power of this model. To illustrate our methodology, we focus on healthcare-associated infections in patients admitted to an intensive care unit.

1. Introduction

Artificial Neural Networks (ANNs) are very accurate predicting tools when compared to more conventional survival models [1–3]. However, they are often seen as *black boxes*, since often it is not possible to express the connection between ANN predictions and input data in a closed form. ANN models are indeed complicated to interpret and it is challenging to identify which predictors are the most relevant [4]. In contrast, semi-parametric hazard-based survival models [5] are examples of interpretable models, whose hazards can measure (directly or indirectly) the effect of each covariate on the outcome of interest.

In order to properly model the temporal evolution of the survival process, including longitudinal information (e.g., biomarkers, health status, clinical measurements) as time-dependent covariates is often informative. These covariates are usually *internal* and they require extra modeling to predict survival functions accurately [6]. The use of Joint Modeling (JM), which attempts to jointly model the longitudinal covariates and the event time, might be then a natural choice [7–9]. Although JMs can efficiently estimate the underlying parameters when the model is correctly specified, they are sensitive to misspecification of the longitudinal trajectory [10] and they are complex to estimate.

For these reasons, we consider a Landmarking (LM) approach for the dynamic prediction of the outcome of interest (e.g., intensive care unit acquired infections). LM is indeed a pragmatic approach that avoids specifying a model for the longitudinal covariates and it is robust under misspecification of the longitudinal processes [11,12]. The main idea behind LM is to select a point in time s known as a landmark. By selecting subjects at risk at s (i.e., left-truncation at time s) and by imposing administrative right-censoring at time $s + w$ (*horizon time*), a landmark dataset is then constructed. Thus, for a time-dependent covariate $Z(t)$, only the value $Z(s)$ at s is considered so that the resulting LM dataset can be analyzed by using standard methods: $Z(s)$ is indeed treated as a time constant covariate. In case of competing events, the LM approach can be generalized to the Competing Risks model (LM-CR), see [13].

The novelty of the manuscript is the inclusion in the LM-CR model of time-dependent information coming from *high-resolution* Electronic Health Record (EHR) data: vital signals recorded in the Intensive Care Unit (ICU) monitors and sampled every minute (i.e., heart rate, mean arterial blood pressure, pulse pressure, arterial oxygen saturation, and respiratory rate). A type of deep neural network, a Convolutional Neural Network (CNN), that looks for predicting patterns present in the signals prior to the landmark time s, is used as a features' extractor to be included in the main LM-CR model. We hypothesize indeed that these patterns represent additional information, not contained in the *lower-resolution* covariates.

Although the LM-CR is in itself an interpretable model, we would like to interpret the additional predicting power of the CNN score in terms of the medical conditions of the patients. Thus, we studied the pattern recognition performed by the CNN and made it interpretable via a Saliency Map Order Equivalent (SMOE) scale [14]; an algorithm that describes the statistics of the activated feature maps of the hidden layers of the network. By the SMOE scale, we could visualize the regions of the input data with the highest *saliency* for the prediction. Hence, we extracted subsets of the signal with the highest cumulative

* Correspondence to: Mathematics Department, University of Genoa, Via Dodecaneso 35, 16146, Genoa, Italy.
E-mail address: giacomo.lancia@gmail.com (G. Lancia).

https://doi.org/10.1016/j.artmed.2024.102862
Received 2 October 2023; Received in revised form 25 March 2024; Accepted 25 March 2024
Available online 28 March 2024

saliency, to perform a data-driven clustering of patients who are more likely to experience the outcome in the fore-coming prediction window. This approach represents a proof of concept for future applications of our method.

In order to illustrate the methodology, we focused on healthcare-associated infections in patients admitted to an ICU, where they were a major cause of morbidity and mortality [15–18]. Therefore, early identification of infectious events could help physicians in the prevention and management of infectious complications in the ICU [19,20]. Moreover, the dynamic prediction of nosocomial infections is a modeling challenging task. The establishment of the presence of infection is not straightforward, and the exact time of infection onset cannot be directly observed. Hence, a method that can predict an approaching infection, might give the partitioners valuable lead time to intervene.

The structure of the paper is the following. In Section 2 we describe the data and define the outcome we want to predict; in Section 3 we introduce the two-step modeling approach; in Section 4 we explain the design of the CNN, its training, and the *risk score*'s extraction. In Section 5 we define and fit the LM-CR model with the inclusion of the *risk score* extracted by the CNN. Finally, in Section 6 we perform a data-driven clustering based on the SMOE scale analysis of the EHR instances. The *Supplementary material* file contains further information about the data, the selection of the design of the CNN, and a more detailed explanation of the SMOE scale used in the paper.

2. The data

We analyzed data from the Molecular Diagnosis and Risk Stratification of Sepsis (MARS)-cohort [21]. We selected patients >18 years of age having a length of stay >48 h, who had been admitted to the ICU of one of the participating study centers between 2011 and 2018. In addition, we also used high-resolution data streams from vital signs monitors which had been recorded in the hospital information system at a 1-min resolution.

As the outcome parameter for our primary modeling attempt, we used the onset of the first occurrence of a suspected Intensive Care Unit Acquired Infection (ICU-AI) within a 24-h time window prior to the moment of prediction. The time of infection onset was determined by either the start of new empirical antimicrobial treatment or the sampling of blood culture (subsequently also followed by antibiotic therapy), whichever occurred first. The dataset thus consisted of 5075 ICU admissions in which 871 first cases of ICU-AIs occurred. Importantly, the incidence of ICU-AI remained relatively constant across ICU stay at a mean rate of 0.04 (SE 0.01) events per day during the first 10 days in ICU. Median time of onset was 5.25 (IQR 3.80–9.45) days following admission.

We selected candidate predictors among several variables based on literature review, a priori consensus of clinical importance, and prevalence in the study population. These covariates include both time-fixed variables reflecting the baseline risk of infection, as well as time-dependent data representing the dynamics of the clinical evolution of patients over time, e.g., laboratory values and physiological response and organ function parameters; see Table 1 and Table 2 in Section 1 of the *Supplementary Material*.

3. Two-step modeling strategy

This section offers a concise introduction to the methodology we have proposed. To take advantage of all longitudinal clinical data and to include observations with different temporal resolutions, we designed our model by means of a *two-step* modeling approach. Specifically:

Step 1: We utilize a CNN to investigate the longitudinal evolution of EHR data. The specific EHR data under examination are the high-frequency vital signs recorded by the monitors in the ICU. These vital signs are sampled at a frequency of 1 min. The CNN is finalized to provide a *risk score of infection* (or more simply the *risk score* or CNN score). The risk score of infection is designed to prospect the occurrence of an infectious episode at any time during the therapy, given the EHR. The higher the risk score, the more the clinical risk of an infectious episode to occur in the near future. For ease of use, the risk score ranges from 0 to 1. Despite achieving values from 0 to 1, the risk score does not represent the probability of infection. From a theoretical perspective, the CNN output is not a probability. More details about this step are discussed throughout Section 4.

Step 2: The LM-CR model is fitted, including all explanatory variables, i.e., *baseline covariates* (e.g., sex, age, ICU admission type, and admission comorbidities), the *low-frequency predictors* (e.g., consciousness score, laboratory measurements, and bacterial colonization) and the *risk score* derived in Step 1. This model combines two models: the Landmark approach and the Competing Risk model. The Landmark model allows us to predict the onset of a suspected infectious episode at any moment of the therapy, based on the data at one previous moment of the ICU stay. The Competing Risk is based on the implementation of a Cox proportional hazard model with two failure causes: the onset of an acquired infection and the occurrence of one of two exclusive events, namely patient death or discharge from the ICU. Additional mathematical details and further insight into this step are elaborated in Section 5.

In summary, we trained the CNN using EHR data and evaluated the risk score of infection throughout each patient's ICU stay. As we will discuss in Section 4, we evaluated the risk score for each patient at 8-h intervals starting from ICU admission. The evaluation of the risk score at some generic time, provided information regarding the chance of an infectious episode in the forthcoming 24 h. Subsequently, the risk scores were integrated with both the low-frequency and the baseline predictors. This comprehensive and massive set of predictors was then employed to train the LM-CR model, serving as the primary tool for making dynamic predictions concerning the onset of infectious episodes.

4. Step 1: CNN at work

This section gives insight into the CNN model, including its structure, the data it uses, and how it was trained and tested. It also explains how the CNN model has generated the risks scored.

4.1. Selection of high-frequency instances and imputation

With the term *high-frequency* covariates, we refer to the five high-frequency vital signs available to us, namely Heart Rate (HR), mean Arterial Blood Pressure (ABP), pulse pressure (PP), functional oxygen saturation (SaO_2), and Respiratory Rate (RR). As mentioned, these predictors are sampled with a sampling frequency equal to one minute. These data were arranged in various 24-h time series (i.e., each time series contains 1440 records, one record per minute).

Thus, we selected and extracted the *time series instances* as follows:

1. We excluded the final 24 h of data for patients who passed away during their ICU stay. These time windows might indeed contain unrepresentative information. In fact, medical decisions to withhold treatment in the last 24 h before death could result in extreme or abnormal records. Hence, the use of these records could affect the learning phase of CNN.

2. Starting from admission time τ_0^i of the ith patient, we partitioned all physiological vital signs into time windows of width $w = 24$ h until achieving the final time T_ℓ^i of the patient record (defined as in point 1 for the patients who died during the stay). Therefore, we obtained the set of intervals $\mathcal{P}^i$ for the patient i:

$$\mathcal{P}_i := \bigcup_{k\geq 1} \left\{ [\tau_0^i + (k-1)w, \min(\tau_0^i + kw, T_\ell^i)] \right\}$$

Likewise, we defined the set of time windows *shifted by* δ as:

$$\mathcal{P}_i^\delta := \bigcup_{k\geq 1} \left\{ [\tau_0^i + \delta + (k-1)w, \min(\tau_0^i + \delta + kw, T_\ell^i)] \right\},$$

provided that $T_\ell^i \geq \tau_0^i + \delta$. Hence, the time windows selected for the patient i are the ones belonging to the set $\mathcal{P}_i^{\text{total}} := \mathcal{P}_i \cup \mathcal{P}_i^{8\text{ h}} \cup \mathcal{P}_i^{16\text{ h}}$; see Fig. 1. The collection of the time windows in $\mathcal{P}_i^{\text{total}}$ (i.e., consecutive windows of 24 h and their translations of 8 and 16 h), allows chunk the longitudinal evolution of the signals coherently with the way we extracted the low-frequency time-dependent covariates of Step 2. We shall refer to the portion of the five vital signs signals corresponding to an interval in $\mathcal{P}_i^{\text{total}}$ with the term *time series instance*.
3. Per each patient i who has acquired no infection during his/her stay in the ICU, we termed his/her time series instances as the *not-infected* instances. For such a patient we considered all time series instances whose time windows are in $\mathcal{P}_i^{\text{total}}$.
4. For each patient i who acquired an infection during the stay in the ICU, we first divided his/her ICU stay as in point 2 ($\mathcal{P}_i^{\text{total}}$). We then labeled as an event all the time windows where an ICU-AI event has occurred (i.e., those time windows including the time-stamp at which the ICU-AI episode was recorded). Likewise, we also labeled all the time windows preceding the time window containing the onset of an ICU-AI event as outcome events. By doing this, we tagged as an event those time series instances anticipating at the most the 24 h prior to the moment when the ICU-AI episode was reported. This choice comes naturally with the necessity of modeling the high volatility of the outcome variable under the exam. Unlike other events, such as death in the ICU, the onset of acquired infections cannot be detected at one precise moment unless the worsening of clinical conditions has become overt. All remaining time windows associated with no infectious episode, are therefore treated as not-infected.
5. We considered the first ICU-AI episode while discarding all the other recurrent episodes from the same patient. More precisely, all the instances following the first infection were discarded.
6. We equipped each *time series instance* with an extra time series monitoring the presence of missing values: This strategy allowed us to track the percentage of missing records at each time stamp.

Hence, each *time series instance* was described by a 6×1440 matrix, whose rows represent the type of *time series features* (i.e., HR, ABP, pulse pressure, SaO_2, BR and missing records) and the columns the time domain (note that 1440 corresponds to the total number of records in a day; calculated as 24 h multiplied by 60 min per hour). The illustration of one sample *time series instance* is shown in Fig. 2.

Missing values of EHR have been imputed by using a zero-order spline, i.e., the Last Occurrence Carried Forward (LOCF) method. Despite being a very simplistic approach, we noted that it has already been applied in some other similar contexts; for example [22,23]. In our case, however, the simplicity of this imputation method is mitigated by the inclusion, per each time series instance, of an extra time series reporting the intervals and the number of vital signs that were missing. This strategy helps the CNN model to better recognize the correct informativeness of patterns, that are transmitted through the first layers. By construction, when the first convolutional layer processes the features of complete vital signs, the extra time series of missing values is not involved in the convolutional operator, assuming zero values. However, when vital signs are not complete, the processing of ancillary time series works as an additional term, readjusting the argument of the activation function. This adjustment is modulated by specific weights that are refined during the learning phase.

In addition to theoretical considerations, the choice of the LOCF imputation method was also motivated by a comparative analysis involving two alternative methods. The first method was the multivariate kNN (k-Nearest Neighbours) [24]. The second one was constructed to better align with the inherent nature of missing values in the ICU context. We referred to such a strategy as *ICUAI-Imputation method*. In essence, missing intervals with an amplitude exceeding 4 h were substituted with a constant out-of-range value (e.g., 100), while intervals shorter than 4 h were imputed with a null constant value. The rationale behind the ICUAI-imputation method drew inspiration from the practical medical perspective in managing Electronic Health Records (EHR). Intervals of approximately 4 h or longer typically correspond to the duration of surgical operations. Shorter intervals were often associated with the temporary interruption of ICU monitoring, resulting from the unintentional detachment of devices, either by a patient or due to device malpositioning. With this imputation strategy, we systematically filled specific types of missing intervals with designated placeholder values; this way, the imputed patterns of missing intervals could also distinguish clinical events of interest occurring during the ICU stay. A comparison between all these imputation methods revealed that the LOCF ensures the highest performance for the CNN model. Deeper insights into this are available in Section 2 of the *Supplementary Material*.

Before feeding the vital signs into the CNN model, we preprocessed the vital signs. In particular, we applied a single linear transformation to all time-series features to map them into the range $[-1, 1]$. We devised and applied a linear transformation to all the time series features of the same kind. Thus, considering the overall statistics of available vital signs, we crafted linear mappings for each time-series feature to rescale them within the $[-1, 1]$ domain. More insight into the pre-processing is available in Section 2 of the *Supplementary Material*.

We remark that in order to illustrate our methodology, we opted to concentrate on a 24-h time window primarily. For completeness, the analysis was also repeated with a 48-h window, as reported in Section 2 of the *Supplementary material*.

4.2. Design of the CNN

The last decade has shown how the predicting skill of CNN turned out to be highly successful in solving various tasks in many different contexts, e.g. image recognition [25–28], anomaly detection [29–31], and time series forecasting [32–35] among others. In fact, this class of Artificial Neural Networks (ANNs) is specifically designed to work with grid-structured data. Its great ability in processing complex multi-level data is mostly due to the combination of both convolutional and max-pooling operators which enable the encoding of sequential patterns along the multi-dimensional domains of input data. In order to derive the risk score of infection from the EHR, we have therefore chosen to utilize a convolutional network: its architecture is composed of convolutional, pooling, and dense layers only. Such a choice seems natural since the CNN is translational invariant; this allows us to search for relevant clinical patterns that might be informative about the early occurrence of the first acquired infection.

In order to give quantitative grounds to this choice, we compared CNN's accuracy with other traditional Machine Learning and ANN-based models, namely the Logistic Regression (LR), the linear Supported Vector Machine (SVM), the Multi-Layer Perceptron (MLP), and the CNN-LSTM networks (where LSTM stands for Long Short-Term Memory). Thus, we trained, validated, and compared the predictive power of the mentioned models over a fine grid of hyperparameters. We introduce here that the performances of the models were evaluated using the area Under the Receiver Operating Characteristic curve

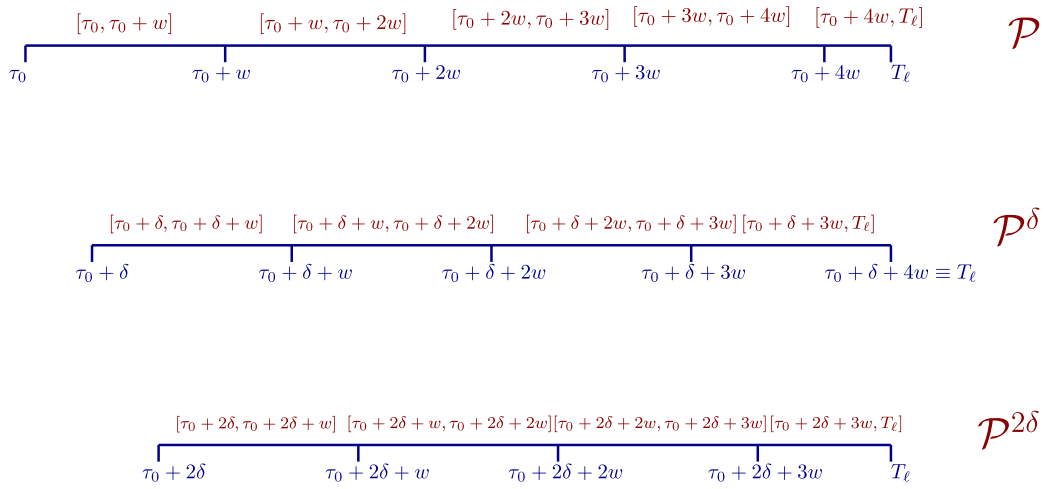

Fig. 1. Example of time windows selected for one patient. τ_0 denotes the admission time of the patient, while w the amplitude of the prediction window. The set with all these prediction windows is denoted with $\mathcal{P}$. The picture below shows the selection of windows shifted by a quantity δ; the set of windows is denoted with $\mathcal{P}^\delta$. Similarly, the selection of windows shifted by a quantity 2δ is also shown.

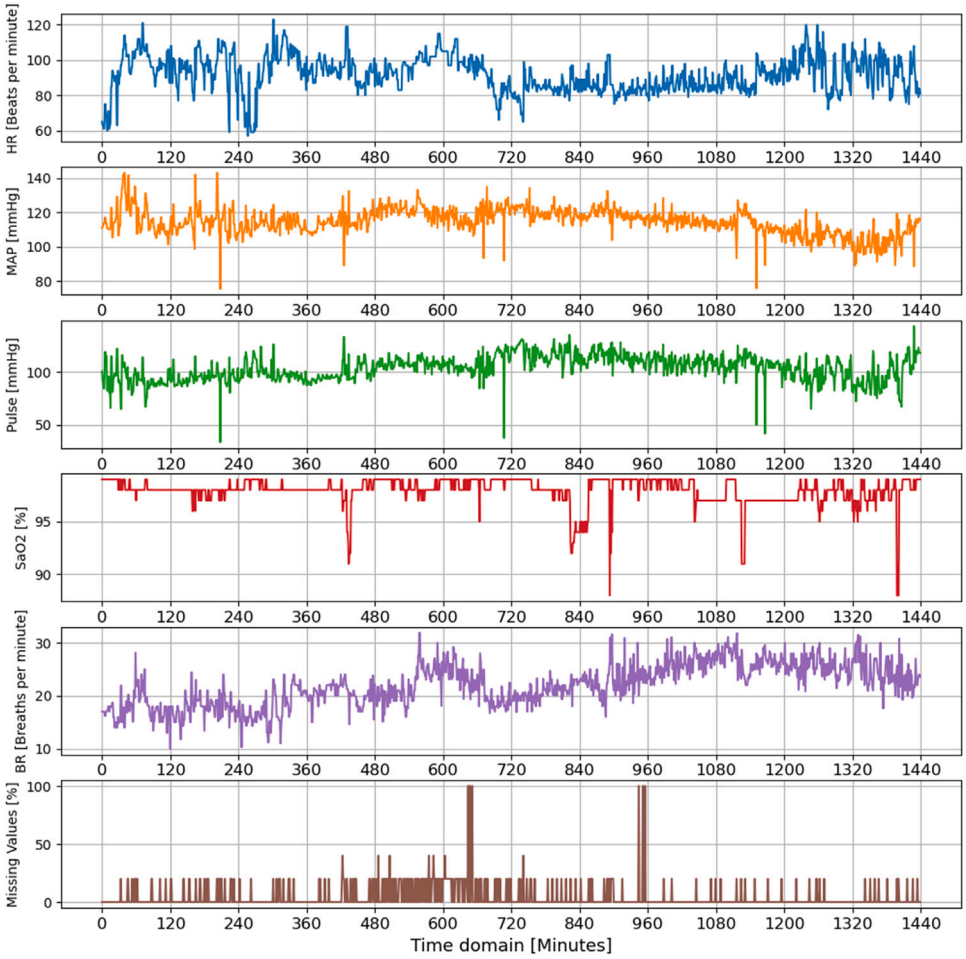

Fig. 2. Example of *time series instance*. x-axis: time-domain (24 h). y-axis: the values taken by each time series feature. In specific, HR in blue, ABP in orange, Pulse Pressure in green, SaO_2 in red, BR in purple, and the auxiliary time series (with the missing values incidence) in brown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(AUROC), or more simply AUROC score. The AUROC score of each model has been listed in Table 1. In order to ensure that the candidate models were able to investigate longitudinal evolution across different time scales of interest, we examined their performance using both 24-h and 48-h prediction windows (i.e., the 24-h instance and 48-h instance models of Table 1). We stress that the selection of the 48-h instances was made by readapting the strategy of Section 4.1. The results of Table 1 revealed that the 24-h CNN model did not emerge

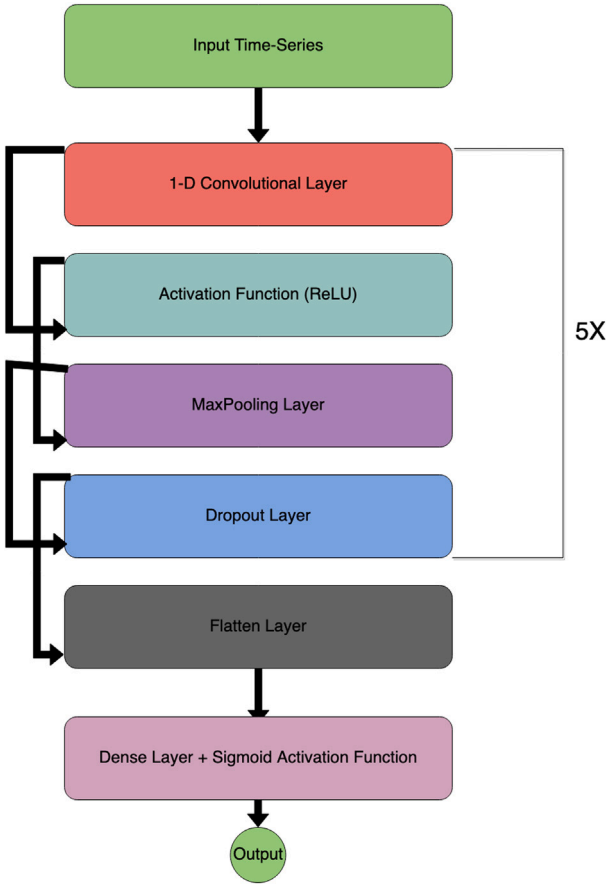

Fig. 3. Schematic illustration of the CNN model. The input signal is processed by a convolutional layer (128 filters of size 3). The *ReLU* function is applied before a *max-pooling* operator which reduces the size of the features. After each *max-pooling* layer follows a *dropout layer* whose dropout rate is 0.25. This sequence of hidden layers is repeated five times. The feature maps are then flattened into an array (flatten layer) and then propagated through a *fully-connected* layer (dense layer) with a sigmoid activation function.

Table 1
Model selection summary: The highest performance achieved during the validation phase, measured by AUROC, is reported for each investigated model. The columns displaying AUROC scores represent either the 24-h instance model or the 48-h instance model. AUROC scores have been rounded to the nearest second decimal. Errors were assessed using the Standard Error Mean, and if too short, they were substituted with the minimum error, i.e., 0.01.

Model	AUROC (24-h model)	AUROC (48-h model)
LR	0.59 ± 0.01	0.59 ± 0.01
SVM	0.57 ± 0.01	0.57 ±0.01
MLP	0.63 ± 0.01	0.61 ± 0.01
CNN	0.72 ± 0.01	0.68 ± 0.02
CNN-LSTM	0.74 ± 0.01	0.59 ± 0.01

as the absolute top performer. It achieved an AUC of 0.72, while the 24-h CNN-LSTM got an AUC of 0.74. However, we were motivated to select the CNN model because of its skill in modeling the risk score of infection when considering both the 24-h and 48-h instances. In fact, the 48-h CNN-LSTM showed a decrease in its predictive power, yielding an AUC of 0.59, whereas the 48-h CNN achieved an AUC of 0.68. The CNN model demonstrated a more robust skill in capturing relevant patterns with both time scales. Also, we opted for a CNN design, due to the benefit of explaining its pattern recognition activity thought the saliency map analysis, as will be presented in Section 6. Further details regarding the model selection strategy are available in Section 2 of the *Supplementary Material.*

The final architecture chosen for the CNN is the following:

1. *Convolutional Layers*: The number of filters on each layer is 128, and each filter has a size of 3 (pixels). We call a *feature map* the output of a filter applied to the previous layer.
2. *Activation Layer*: The ReLU function (i.e. $\mathrm{ReLU}(x) := \max(0, x)$) is applied after each convolution operator. This application of a non-linear activation function on the feature maps gives rise to the *activated feature maps*.
3. *Max-pooling layer*: The activated feature maps are resampled via a max-pooling operator with a pooling size of 2 (sub-sampling).

Also, a *dropout layer* with a dropout rate of 0.25 is included after each max-pooling layer. This sequence of hidden layers is repeated five times. The last feature map is flattened into an array and then propagated into a *fully-connected* layer (dense layer) with a sigmoid activation function. The activation function returns a positive output between 0 and 1, that is, the risk score. The architecture of the chosen CNN is sketched out in Fig. 3.

4.3. Training and overall evaluation of the CNN

Before training the model with the EHR data, we made a selection among the time series instances available to us. Specifically, we opted for under-sampling the total amount of time series instances. This choice has a double reason. Firstly, we aimed to ensure that the risk score which will be employed in the Landmark model was evaluated from time series instances not previously processed by the CNN during the training or validation phase. The finality of this step was to incorporate into the Landmark model a risk score that originated from time series instances that were never propagated through the CNN before. Secondly, we aimed to address the inherent challenge of training CNNs effectively on very imbalanced datasets. The number of time series instances in the case group (i.e., those instances representing the ICU-AI episodes) was less than one-twentieth of the total amount of time series instances in the control group (i.e., those instances not representing the ICU-AI episodes). Thus, we opted for fitting the CNN model on a population of *time series instances* with a control-case ratio of 8:1 (i.e., the number of time series instances in the control group is 8 times larger than the case group). It is important to remark we applied a random under-sampling on the control group only.

The fit of the model was designed to optimize the binary cross-entropy loss function through the ADAM algorithm [36]. Therefore, we trained the CNN to solve a binary classification task. We anticipate that the difference of AUROC between the Deep-LM-CR and the LM-CR model (i.e., both Landmark models with and without the CNN score, respectively; see Section 5.2) was employed to evaluate the relative goodness between the two models. Moreover, we considered the Brier score [37] as an alternative metric for assessing the prediction power of the models. The Brier Skill [38] was utilized to assess the relative increase in predictive performance of Deep-LM-CR with respect to the LM-CR.

Although our main interest is not in the prediction formulated by the CNN itself, we also needed to guarantee that the CNN model was able to classify the *time series instances*. Internal validation was performed using the *5-fold cross-validation* method. During the validation of the CNN model as a binary classifier, the data were split into 5 different folds; one at a time each fold was employed to validate the model, while the remaining data were utilized during the training phase. The overall AUROC is then the average over the 5 folds. We recall that the search for the optimal configuration was conducted through the validation of the models over a fine grid of hyperparameters. Additional details are available in Section 2 of *Supplementary Material* in which we delved into the CNN model's AUROC behavior for three key hyperparameters

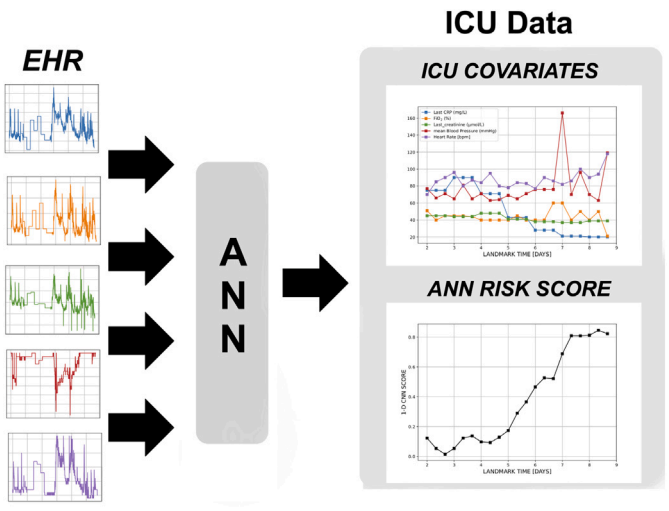

Fig. 4. Schematic representation of the inclusion of the CNN-based risk score $Z_{\text{CNN}}(t_{LM})$ in the ICU cohort data.

4.4. CNN risk score

The extraction of the CNN score and its inclusion in the LM-CR model represent the novel ideas of the manuscript. The risk score of infection is evaluated by means of the CNN, whose architecture was discussed in Section 4.2 together with its training phase in Section 4.3.

The procedure we designed for evaluating the risk score is the following:

1. Consider the vital signs of patient i (HR, ABP, pulse pressure, SaO_2, and RR) and the time series flagging the missing records.
2. Starting from the ICU admission time, extract the 24-h *time series instances* by means of an 8-h sliding time window (see Section 4.1), corresponding to the intervals in $\mathcal{P}_i$.
3. Propagate the *time series instances* through the hidden layers of the fitted CNN model and evaluate the risk score.
4. Assign the risk score to the corresponding time-stamp (i.e., day-month-hour-minute).

A scheme of how we incorporated the risk score into the ICU predictors is illustrated in Fig. 4: for each patient, the risk score was calculated for a set of subsequent times, named *Landmark Times* (LM) and denoted with t_{LM}. Also, at each t_{LM} the values of other time-dependent covariates are reported as well (e.g., CRP, FiO_2, creatinine level, mean blood pressure, mean heart rate). Incidentally, t_{LM} denotes the generic LM time; the LM times represent the ensemble of times at which the Landmark model was fitted, as we shall discuss in Section 5.

5. Step 2: *Deep* LM-CR model

5.1. Notations and LM-CR model

In this Section, we shall present the LM model following the notation used in [13].

We consider a cohort consisting of N subjects, and we denote with $\tilde{T}$ the time of failure, C the censoring time, D the cause of failure, and $\mathbf{Z}(\cdot)$ and an array of covariates. In a general framework, a subject can only experience one of J mutually excluding competing causes of failure; when it occurs D takes a value in {1, …, J} corresponding to the cause under the exam. Alternatively, when no cause has been experienced yet, D always takes value 0. For the ith subject, the tuple $(T_i, \Delta_i, \mathbf{Z}_i(\cdot))$ represents respectively the observed time $T_i = \min(\tilde{T}_i, C_i)$ (i.e., the earliest of failure and censoring time), the cause of failure $\Delta_i = \mathbf{1}(\tilde{T}_i < C_i)D_i$ (with $\mathbf{1}(\cdot)$ the indicator function), and $\mathbf{Z}_i(\cdot)$ the covariates up to time T_i. Note that $\Delta_i = 0$ denotes that the patient has experienced no failing causes; its clinical history has been censored. We shall adopt the subscript j to refer to the competing causes of failure, with $j \in \{1, \dots, J\}$.

We would like to derive a dynamic prediction of the probability distribution function of the failure time of cause j at some time horizon (t_{hor}), conditional on surviving event-free and on the information available at a fixed time t_{LM} (*landmark time*). In other words, given a prediction window w (such that $t_{hor} = t_{LM} + w$), we would like to estimate the survival probability and the Cumulative Incidence Function (CIF) of cause j, namely

$$S_{LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM}) := \mathbb{P}(T > t_{hor}|\mathbf{Z}(t_{LM}), t_{LM}), \tag{1}$$

$$F_{j,LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM}) := \mathbb{P}(T \le t_{hor}, \Delta = j|\mathbf{Z}(t_{LM}), t_{LM}). \tag{2}$$

Thus, the LM approach consists of two steps:

1. We first divide the time domain of our observations $[s_0, s_1]$ into n equi-spaced landmark points denoted with $\{t^k_{LM}\}_{k=1}^n$, where $t^1_{LM} \equiv s_0$ and $t^n_{LM} \equiv s_1$. We fix the width of the prediction window w (i.e., the *lead time*), and then for each LM time t^k_{LM} we create a dataset by selecting all the subjects at risk at time t^k_{LM} and by imposing *administrative* right-censoring at the time $t^k_{LM} + w$ (*horizon time*). Thus, for a vector of time-dependent covariates $\mathbf{Z}(t)$, only the values $\mathbf{Z}(t^k_{LM})$ at t^k_{LM} are considered in the kth dataset. Finally, we create an extensive dataset by stacking all the datasets extracted at each landmark time t^k_{LM} (LM *super-dateset*).
2. The second step consist of fitting the *LM-CR super-model* on the stacked LM *super-dateset* [13]. Since at each t^k_{LM}, the vector $\mathbf{Z}(t^k_{LM})$ is treated as a time constant vector of covariates, the dataset can be analyzed by using standard survival analysis methods.

In the *LM-CR super-model* we fit indeed a Cox proportional hazard model for the cause-specific hazard λ_j:

$$\lambda_j(t|t_{LM}, \mathbf{Z}(t_{LM})) = \lambda_{0j}(t|t_{LM}) \exp[\beta_j^T(t_{LM})\mathbf{Z}(t_{LM})], \tag{3}$$

where $\lambda_{0j}(t|t_{LM})$ denotes the (unspecified) baseline hazards and $\beta_j(t_{LM})$ the set of regressors specific for the jth cause in within the interval $[t_{LM}, t_{LM} + w]$. We assume that the coefficients β depend on t_{LM} in a smooth way, i.e., $\beta_j(t_{LM}) = f_j(t_{LM}, \beta_j^{(0)})$ with $\beta_j^{(0)}$ a vector of regression parameter and $f_\beta(\cdot)$ a parametric function on time, e.g., a spline. Our choice has been a quadratic function:

$$\beta_j(t_{LM}) := \beta_j^{(0)} + \beta_j^{(1)} t_{LM} + \beta_j^{(2)} t^2_{LM}.$$

The estimation of $\lambda_{0j}(t|t_{LM})$ can be made through Breslow-type estimator; we can model such a dependence as

$$\lambda(t|t_{LM})_{0j} = \lambda_{0j}(t) \exp(\gamma_j(t_{LM})). \tag{4}$$

As for the coefficients β, we assume the coefficients γ of (4) to be parametrically dependent on the landmark times, e.g. by means of a quadratic spline

$$\gamma_j(t_{LM}) := \gamma_j^{(0)} + \gamma_j^{(1)} t_{LM} + \gamma_j^{(2)} t^2_{LM}.$$

Fitting this model with the Breslow partial likelihood for tied observations is equivalent to maximizing the pseudo-partial log-likelihood, as shown in [13]. The landmark supermodel can be then fitted directly by applying a simple Cox model to the stacked data set. Hence, after estimating the coefficients and the baseline cause-specific hazards, we get the *plug-in* estimators for the survival probabilities (i.e., $\hat{S}_{LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM})$) and of the CIF of cause j (i.e., $\hat{F}_{j,LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM})$). The explicit form of these estimators is the following:

$$\hat{S}_{LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM})$$

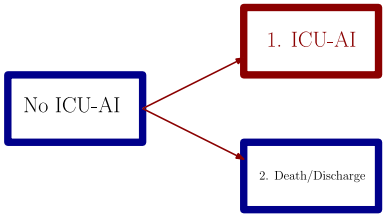

Fig. 5. Competing risks model for ICU-AI.

$$= \exp\left(- \sum_{j=1}^{J} \exp\left(\boldsymbol{Z}(t_{LM})\hat{\beta}_j(t_{LM}) + \hat{\gamma}_j(t_{LM})\right) \left[\hat{\Lambda}_{0j}(t_{hor}) - \hat{\Lambda}_{0j}(t_{LM})\right] \right), \tag{5}$$

and

$$\hat{F}_{j,LM}(t_{hor}|\boldsymbol{Z}(t_{LM}), t_{LM}) = \sum_{t_{LM} < t_i \le t_{hor}} \hat{\lambda}_{0j}(t_i|\boldsymbol{Z}(t_{LM}))\hat{S}_{LM}(t_{hor}|\boldsymbol{Z}(t_{LM}), t_{LM}). \tag{6}$$

The *estimated cause-specific baseline* of (6) is given by

$$\hat{\lambda}_{0j}(t_i) = \frac{\#(t_{LM} \le t_i \le t_{hor}, \Delta_i = j)}{\sum_{t_{LM}: t_{LM} \le t_i \le t_{hor}} \sum_{t_k: t_{LM} \le t_i \le t_k \le t_{hor}} \exp\left[\boldsymbol{Z}_k(t_{LM})^T \hat{\beta}_j(t_{LM}) + \gamma_j(t_{LM})\right]}, \tag{7}$$

while the *estimated cause-specific cumulative baseline* is simply

$$\hat{\Lambda}_{0j}(t) = \sum_{t_i \le t} \hat{\lambda}_{0j}(t_i).$$

5.2. LM-CR for ICU-AI

In the context of dynamic predictions for ICU-AIs, we adopted a CR model and considered three causes of failure: *ICU-AI*, *death in the ICU* and *discharge*; see Fig. 5. No right censoring is present in the data, since no patient left the ICU before discharge or death.

Following the notation used in Section 5.1, we denote with $\tilde{T}$ the time of failure, D the cause of failure (i.e., $D = 1$ denotes an ICU-AI, while $D = 2$ discharge or death), and $\mathbf{Z}(\cdot)$ the array of covariates. For the ith subject the triple $(T_i, \Delta_i, \mathbf{Z}_i(\cdot))$ denotes the observed time $T_i \equiv \tilde{T}_i$, the cause of failure $\Delta_i \equiv D_i$, and $\mathbf{Z}_i(\cdot)$ the vector of covariates.

In this article, we consider the prediction window was set to $w = 24$ h. The time domain is $[s_0, s_1]$, with $s_0 = 48$ h and $s_1 = 240$ h, and we consider $n = 25$ LM times t_{LM}, i.e., two subsequent LM times are at a distance of 8 h.

If we denote with $Z_{\text{CNN}}(t)$ the CNN risk score at time t (see Fig. 4) and with $\mathbf{Z}(t)$ the vector of all the other covariates in the LM-CR model at time t, we are interested at the dynamic predictions of the following two models:

1. $\pi_1 := F_{1,LM}(t_{hor}|\mathbf{Z}(t_{LM}), t_{LM})$: i.e., the CIF of infection conditioned on the survival up to time t_{LM} and on the *low frequency* covariates (LM-CR model);
2. $\pi_2 := F_{1,LM}(t_{hor}|\mathbf{Z}(t_{LM}), Z_{\text{CNN}}(t_{LM}), t_{LM})$: the CIF of infection conditioned on the survival up to time t_{LM} on both the *low frequency* covariates and Z_{CNN} (Deep-LM-CR model).

By comparing the accuracies of π_1 and π_2, we can measure the added predictive power of the CNN score. We shall refer to the first model as the LM-CR and the second as the Deep-LM-CR.

5.3. Evaluation of LM-CR model

We utilized the AUROC metric to evaluate the prediction made at each single landmark time. When considering an overall measure, the evaluation of a global AUROC needed to consider the time-dependent character of the dynamic. Similarly to the estimator of the prediction error proposed in [39], we evaluated the overall AUROC by taking into account the change in time of the size of the risk-set. The absence of censoring allowed us to estimate the overall AUROC score simply through

$$\text{AUROC}_{\text{global}} = \frac{\sum_{k=1}^{n} R(t_{LM}^k)\,\text{AUROC}(t_{LM}^k)}{\sum_{k=1}^{n} R(t_{LM}^k)}, \tag{8}$$

with t_{LM}^k the kth landmark time, n the total number of landmark times, and $R(t_{LM}^k)$ the size of the risk-set at time t_{LM}^k. Likewise, we estimated the overall Brier score as

$$\text{BS}_{\text{global}} = \frac{\sum_{k=1}^{n} R(t_{LM}^k)\,\text{BS}(t_{LM}^k)}{\sum_{l=1}^{n} R(t_{LM}^l)}. \tag{9}$$

The individual impact of predictors in formulating the final prediction has been visualized through heat maps. We also looked at the relative variation of the overall AUROC between the model including all predictors and the one where one single predictor is removed. Thus, we summarized the results with a heat map to illustrate the relative change in AUROC resulting from the removal of a single predictor at specific landmarking times.

Finally, we remark that internal validation was performed using a *10-fold cross-validation* method. The overall $\text{AUROC}_{\text{global}}$ and the $\text{AUROC}(t_{LM}^k)$, evaluated at each time t_{LM}^k, were obtained by taking the average over the 10 folds. For both CR-LM and Deep-CR-LM models, we reported 95% bootstrap confidence intervals.

5.4. Results

In this Section we shall illustrate how the CNN risk score Z_{CNN} adds extra predictive information to the model, not present in the standard covariates.

In Fig. 6 we plotted the empirical distribution of $Z_{\text{CNN}}(t_{LM})$ for three landmark points (i.e., $t_{LM} \in \{3, 6, 8\}$) and stratified by the cause of failure. As expected, the distribution of Z_{CNN} for infected patients is more skewed on the right: while on day three this phenomenon is mild, on days 6 and 8 the skewness of the density distribution is much more evident.

In Fig. 7, we reported the Pearson correlations between the CNN risk score and the vital signals averaged per 24-h time windows prior to the landmark time (i.e., the time-dependent covariates included in the LM-CR). Although the risk score is evaluated relative to these signals, only mild correlations are present. Our main hypothesis is indeed that $Z_{CNN}(t_{LM})$ has added predictive information, not contained in the other covariates $\mathbf{Z}(t_{LM})$.

Moreover, with regards to the cause-specific hazards for infection, the CNN risk score turned out to be the most important predictor: $\beta^{(0)}_{1;\text{CNN}} = 4.8$ (95%CI 3.05–6.72). A complete list with all cause-specific hazards for ICU-AI is reported in Table 3 of the *Supplementary Material*.

The LM approach provides a *plug-in* estimator for the dynamic prediction (2) of the CIFs of ICU-AI. To give an example of the dynamic prediction allowed by the model, we reported in Fig. 8 the CIFs for the LM-CR and the Deep-LM-CR models as a function of both the landmark time and the quantile groups of the fitted linear predictors. Given the value of the covariates at the landmark time t_{LM}, the CIF at any s, with $s \in [t_{TM}, t_{LM} + w]$ is given indeed by the *plug-in* estimator $\hat{F}_{1,LM}(s|\boldsymbol{Z}(t_{LM}), t_{LM})$ of (2).

The dashed red line in Fig. 8 denotes an arbitrary warning level for the CIF of infection (e.g., 8%). We can see that, for the fourth quantile Q_4 and at LM time $t_{LM} = 4$ days, the Deep-LM-CR model has a *lead time* of circa 3 h in reaching the warning threshold before the LM-CR model.

The overall measure for the LM-CR model is $\text{AUROC}_{\text{global}} = 0.69$ (95%CI 0.68–0.70), while for the Deep-LM-CR is $\text{AUROC}_{\text{global}} = 0.75$ (95%CI 0.73–0.76). The $\text{AUROC}(t_{LM}^k)$ scores evaluated at each time

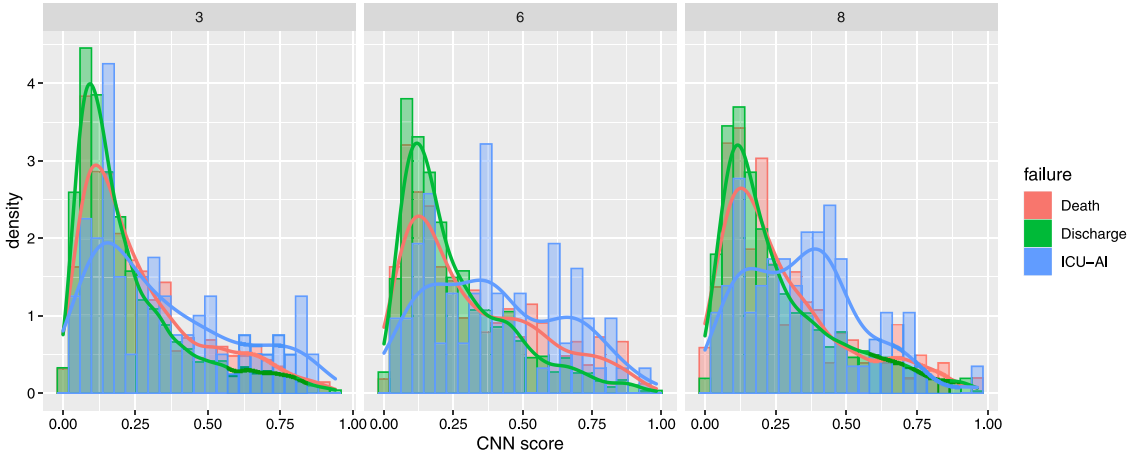

Fig. 6. Distribution of the CNN risk score at three different landmark points ($t^k_{LM} \in \{3, 6, 8\}$ days), stratified for the cause of failure.

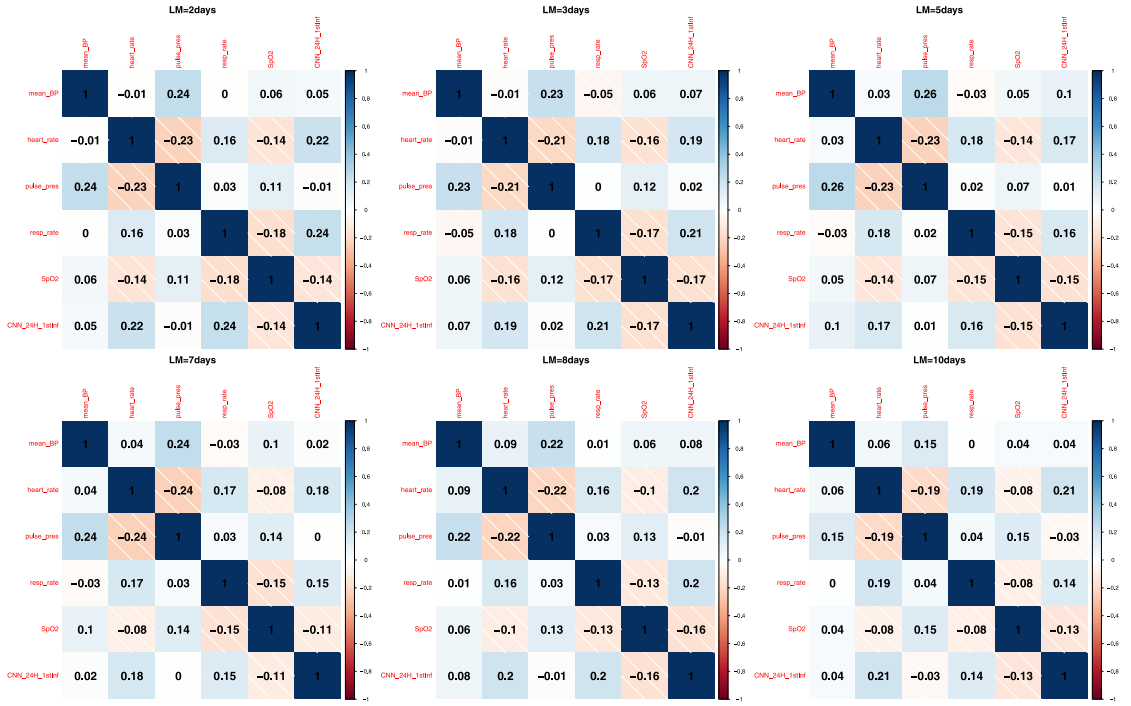

Fig. 7. Correlation plot: CNN risk score vs. the vital signals (averaged in the 24 h before the landmark).

t^k_{LM}, with $k \in \{1, \ldots, n\}$ are shown in Fig. 9. The LM-CR model always shows lower predictive performance than the Deep-LM-CR. We noticed that at the early stage of the ICU stay (e.g., days 3.66) and around day 7, the CNN can improve the prediction of the traditional ICU clinical covariates of about 8–9%, see Fig. 10.

The evaluation of the Brier Score revealed, for the LM-CR model, an overall measure $BS_{global} = 0.037$ (95% CI 0.036–0.039), while for the Deep-LM-CR we had $BS_{global} = 0.036$ (95% CI 0.035–0.037). The scores $BS(t^k_{LM})$ of each landmark time, are shown in Fig. 11. For the majority of landmark times, we observed that the Deep-LM-CR was slightly more accurate than the other one; in contrast, the LM-CR turned out to be somewhat more precise in a few landmark times on days 2 and 4 and around days 9 and 10. Such a result is also reported in Fig. 12; where the Brier Skill is shown. In the end, we observed an overall Brier Skill of 0.03 with 95% CI equal to (0.01, 0.07). The evaluation of all CI was accomplished via bootstrap resampling (bootstrap population equal to 1000 samples).

The impact of each explanatory variable Z_j involved in the Deep-LM-CR model is shown in Fig. 13, in which we reported the heat-map of the relative increase in AUROC between the Deep-LM-CR without the covariate Z_j and the full model (with $\mathbf{Z}(t_{LM})$ and $Z_{CNN}(t_{LM})$). When $Z_j = Z_{CNN}$, we see that we observe a relative increase in AUROC of at least 4%.

In conclusion, an examination of the global performance of both the LM-CR and Deep-LM-CR models was also conducted, taking into account larger amplitude sliding windows of 16 and 24 h. In the case of the 16-h sliding window, it was observed that the LM-CR exhibited an overall $\text{AUROC}_{\text{global}}$ of 0.70 (95% CI 0.69–0.71), whereas the Deep-LM-CR demonstrated a higher $\text{AUROC}_{\text{global}}$ of 0.73 (95% CI 0.72–0.74). Similarly, for the 24-h sliding window, the LM-CR displayed an overall $\text{AUROC}_{\text{global}}$ of 0.69 (95% CI 0.68–0.70), while the Deep-LM-CR exhibited a comparable $\text{AUROC}_{\text{global}}$ of 0.73 (95% CI 0.72–0.74). A comparison between these findings and those obtained for models employing an 8-h sliding window suggests a preference for the latter, as it attains the highest AUC when integrating the risk scores. It is noteworthy that the 8-h sampling frequency stands as the maximum among low-frequency ICU covariates. Consequently, this analysis was

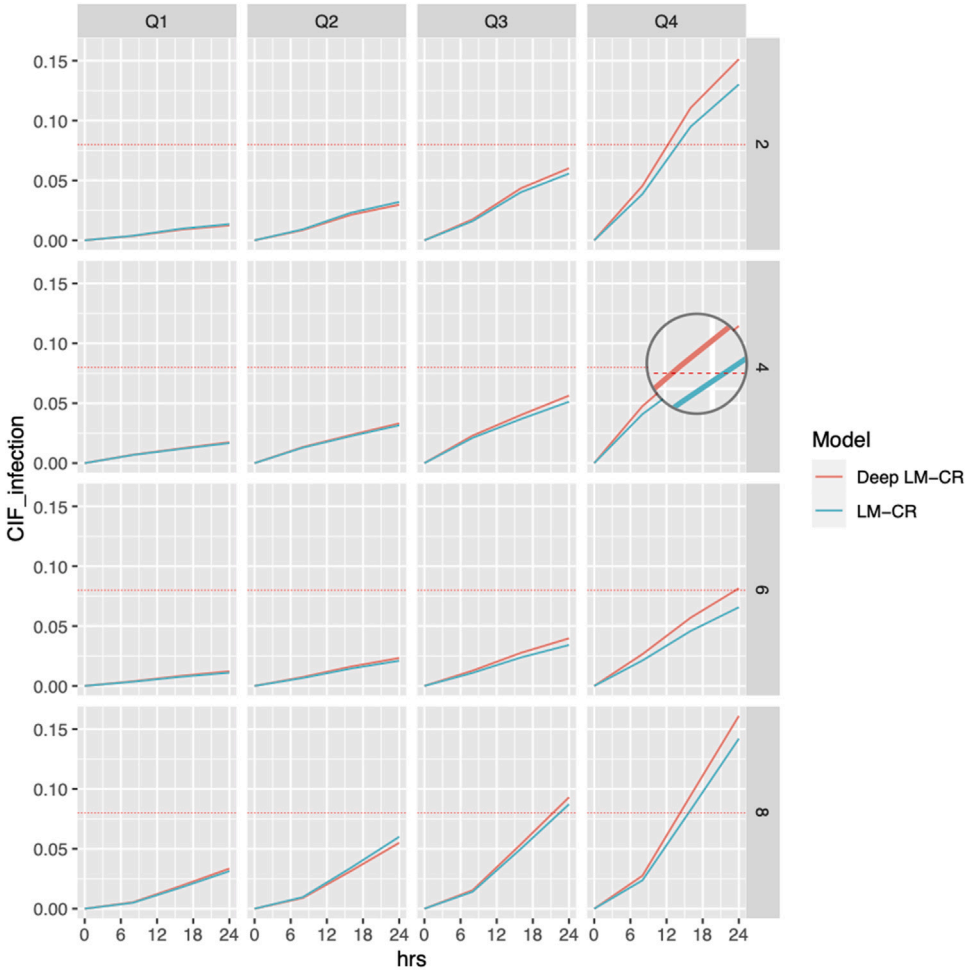

Fig. 8. Comparison of the CIFs at different landmark times (i.e., $t_{LM}^{k} \in \{2, 4, 6, 8\}$ days) of the models LM-CR and Deep-LM-CR.

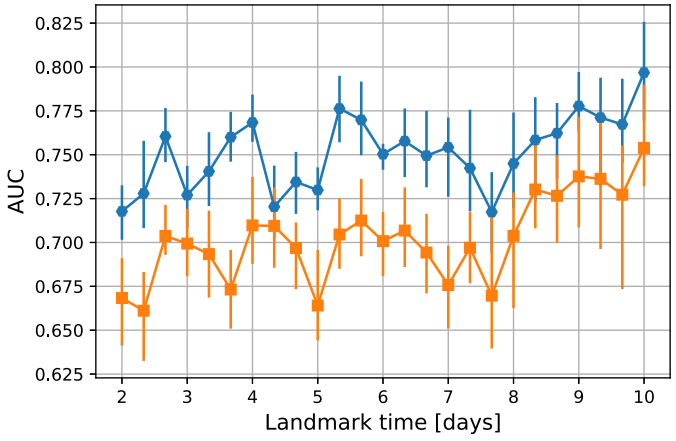

Fig. 9. AUROC score (y-axis) as a function of the landmark times(x-axis). The two curves represent the predictive performance of the basic CR-LM model (orange) and the Deep-CR-LM model (blue). The error bars denote the 95% bootstrap confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

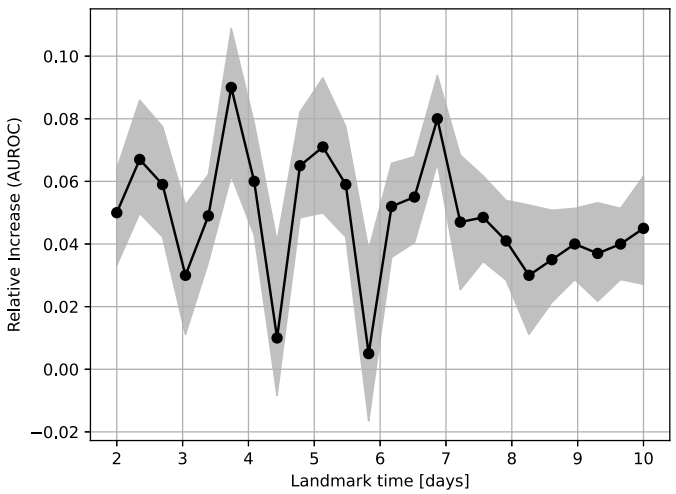

Fig. 10. Overall relative increase of AUROC score (y-axis) as a function of the landmark times (x-axis) when including CNN-based risk score. The 95% CI is represented as the light gray area.

constrained to datasets with sampling periods equal to or larger than 8 h.

Summing up, we have shown that the two-step modeling can effectively lead to an increase in the accuracy of the predictions. The extra predicting power comes from the inclusion of the CNN-based risk score, which is a summary measure of the predicting patterns found by the CNN model trained on only five vital signs signals (sample frequency of 1 min).

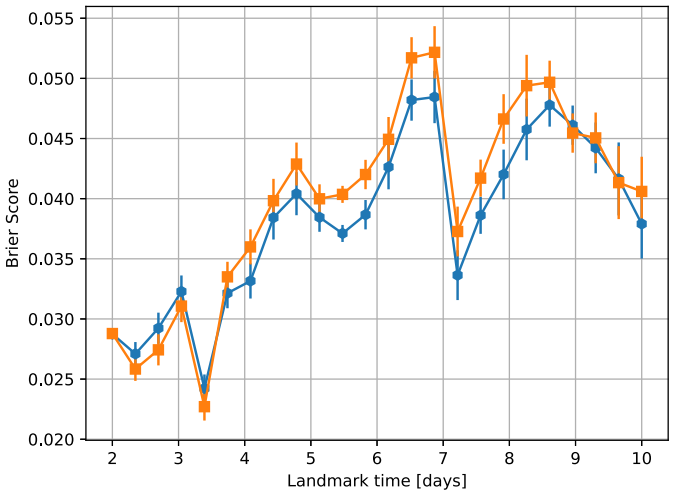

Fig. 11. Brier score (y-axis) as a function of the landmark times (x-axis). In orange, the predictive power of the CR-LM model is shown, while the Deep-CR-LM is in blue. The error bars represent the 95% bootstrap confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

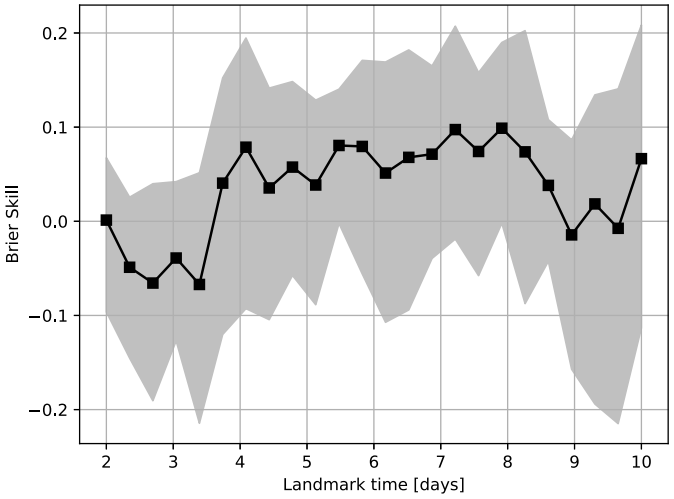

Fig. 12. Brier skill (y-axis) as a function of the landmark times (x-axis). The 95% CI is represented as the light gray area.

We remark that in our analysis we did not consider recurrent infections, but we limited the attention to the first episode of ICU-AI.

5.5. Comparison with a full ANN-model

For the sake of completeness, we would like to compare the predictions from our two-step modeling strategy with a full-ANN model. More specifically, we considered a Two-Branch Artificial Neural Network (TB-ANN) in order to process simultaneously high-frequency, low-frequency, and fixed-time covariates within a unique ANN model. The TB-ANN consists of two distinct branches; the first has the scope of analyzing the high-frequency data only, while the other analyses low-frequency and fixed covariates simultaneously. The two branches are then connected and propagated through a prediction layer (i.e., a dense layer with a sigmoid activation function) returning an output score similar to the CNN risk score. More details about the TB-ANN's architecture are discussed in the Supplementary Material. Therefore, we are interested in comparing a model based on the estimation of interpretable quantities as hazard ratios (i.e., Deep LM-CR) and a completely ANN-based model (i.e., TB-ANN) for predicting first infectious episodes.

Similarly to the Deep LM-CR, we have trained the TB-ANN at equi-spaced landmarking times within the time domain $[s_0, s_1]$, with s_0 and s_1 equal to 48 and 240, respectively; two generic subsequent landmarking times are 8-h distant. When training the TB-ANN models, we only considered the data available at each landmarking time to forecast the presence of an infectious episode in the next 24 h. To assess the TB-ANN predictive skill we primarily referred to the AUROC metric; we observed an overall AUROC (in the sense of Section 5.3) equal to 0.72 (95% CI 0.55–0.9). As a secondary metric, we also considered the Brier Score; we obtained an overall Brier Score of 0.08 (95% CI 0.06–0.11). The average values of AUROC and BS at each landmark time are reported, respectively, in Figs. 14 and 15. Error bars denote the 95% confidence intervals

For most landmark times, we see that the TB-ANN's AUROC scores lay around values 0.7 and 0.8. However, large fluctuations are present on days 2.67, 5, 8.33, and 10. Especially for the last two mentioned, we have to remark that the reduction of the number of events at late landmark days might overestimate the AUROC scores. For the Brier Score, we observed different profiles at both early and late landmark days. In fact, in the region 2–6 days the Brier Score presented important fluctuations around the global value of 0.08. In particular, on days 3.33 and 3.66, we observed a score of 0.03, while higher scores larger than 0.10 were observed on days 2.66, and 4.66. In contrast with this, the region 6–10 days appeared more stable, with much lower fluctuations, around the value of 0.10.

The Deep LM-CR revealed a more stable prediction than the TB-ANN, while the accuracy was similar. However, our two-step strategy allows an immediate interpretation of the impact of each low-frequency covariate on the prediction and offers the possibility of using methods for interpreting the activity of the CNN, as explained in Section 6.

6. Explainability of CNN-based prediction of ICU-AI

In this section, we present our attempt to make interpretable the activity of the CNN. As shown in Section 5.4, the CNN-based risk score has added predicting power to the LM-CR model. However, for the moment, we do not have any information about the saliency of the vital signs selected by CNN during the training. This knowledge might be crucial for shedding some light on the relation between the activity of pattern recognition of the network and the medical conditions of a patient when an ICU-AI is approaching.

To investigate which characteristics of the pattern selected by the CNN, we use the so-called *Explainable Artificial Intelligence* (XAI), namely a class of methods designed to understand the decisions and the predictions formulated by ANN techniques [40–42]. In the last decade, XAI has turned out a fundamental tool to make various ANN applications more reliable and transparent [43–48]. The scope of XAI is to contrast indeed the widespread *black box* attitude that many users have when applying ANN techniques.

6.1. Explainability via SMOE scale

In the context of ANN, a saliency map is a technique to delve into the activated features of the hidden layers with the scope of revealing which parts of the input domain are most captured for formulating the network's decisions.

The Saliency Map Order Equivalent scale (SMOE) used in the present paper is based on the algorithm developed by [14]: an efficient and non-gradient method based on the statistical analysis of the activated feature maps. For a more detailed description of the SMOE scale, we refer the reader to Section 3 of the *Supplementary Material*.

We would like to use the saliency maps for selecting, in the original 24-h time series, the most relevant 8-h patterns. We stress that the

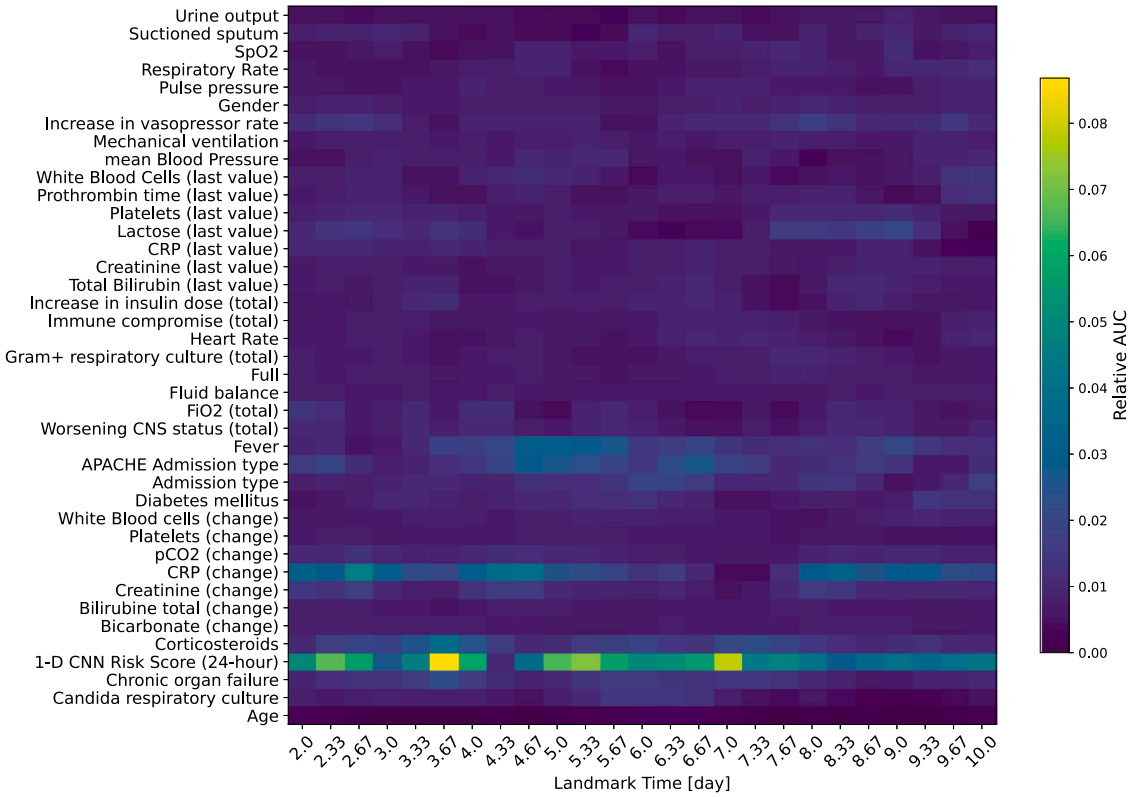

Fig. 13. AUC heat-maps evaluating the impact of each predictor in the Deep-LM-CR model when predicting ICU-AI. The color of each pixel denotes the magnitude of the impact (relative AUROC increase) of one covariate (y-axis) for the LM time (x-axis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

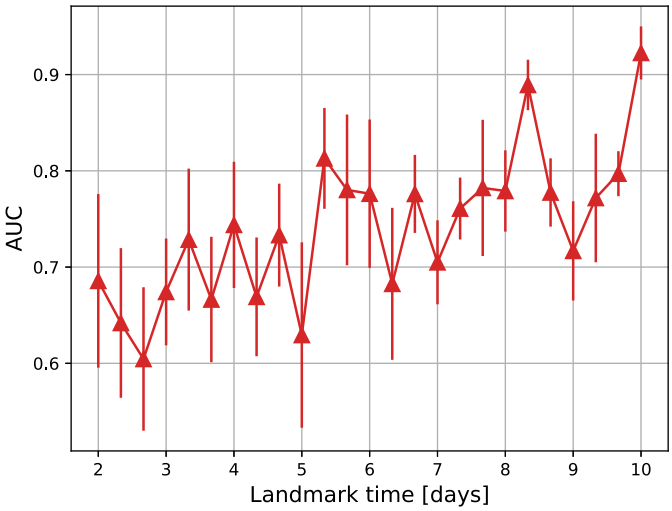

Fig. 14. Mean value of AUROC of the TB-ANN at each landmark time. Error bars are the 95% CI.

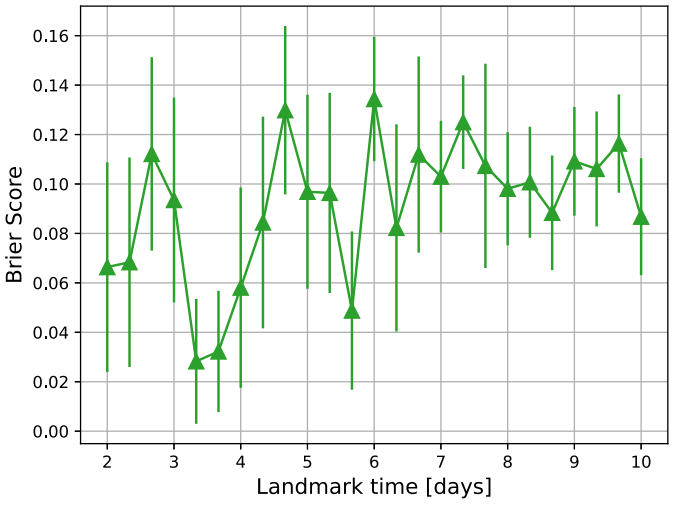

Fig. 15. Mean value of Brier Score of the TB-ANN at each landmark time. Error bars are the 95% CI.

choice of an 8-h pattern was purposefully made to provide readers with a clear illustration of the method. As shown in the previous section, the time scale of 8 h turned out to be suitable to investigate the clinical dynamics of patients. Such a choice therefore aims for maximum alignment with the results demonstrated for the Deep LR-CR model.

Thus, the approach we proposed is the following:

1. We fit three different CNNs, one for each of $t^k_{LM} \in \{3, 7, 10\}$. We consider three distinct CNNs because the predicting patterns found by the network might differ among different periods of the ICU stay (see for instance the discussion in Section 6.3). The LM point *3 days* is a proxy for an early time of the stay, *7 days* for an intermediate time, and finally *10 days* for a later moment. The design of the networks is the same as described in Section 4.2. All these models are validated via 5-fold cross-validation.
2. We study the pattern recognition performed by the hidden layer, and we make it interpretable via the *SMOE scale*. Through this method, we can visualize the regions of the input data with the highest saliency. Specifically, for each model developed at every

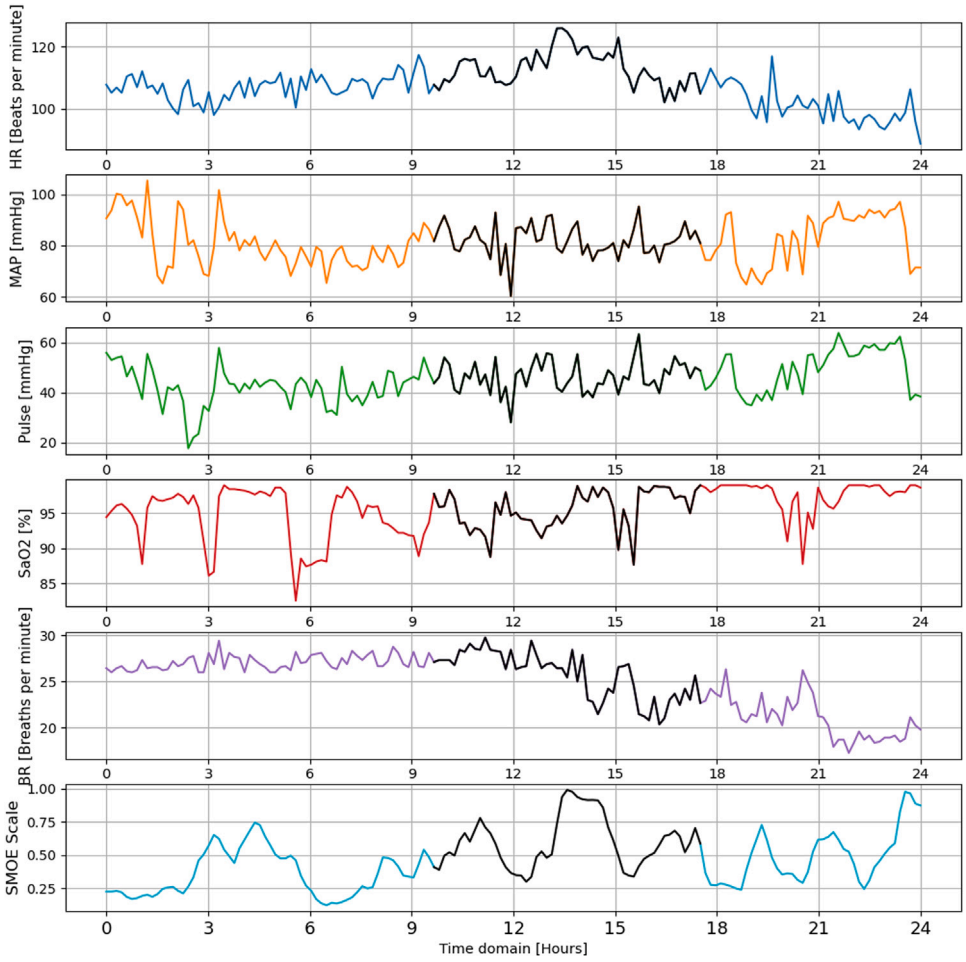

Fig. 16. Schematic visualization of the 8-h most salient patterns within a 24-h time series sample. Starting from the top and descending, the first signals represent the vital signs considered. The cyan line represents the corresponding SMOE map. In black, the 8-h chunk with the highest averaged SMOE value is outlined. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

LM time t^k_{LM}, we construct and visualize the saliency maps of the test set only. We repeat this action for each test set of each cross-validation fold.

3. From each saliency map, we extract the *8-h interval* with the highest *cumulative saliency value*. After having extracted the most relevant 8-h patterns from each *time series instance*, we can focus on their interpretation and their clustering. An example of the extraction of the 8-h most salient pattern is shown in Fig. 16.

6.2. Data-driven clustering of salient patterns

We focus now our attention on the clustering of the most salient patterns extracted in Section 6.1. We would like indeed to answer the question: *how can we link the activity of pattern recognition to some medical conditions, appearing when an ICU-AI is approaching?* Our strategy for answering the question is the following:

1. We collect the set of the most predictive patterns with an amplitude of 8 h, obtained by applying the SMOE scale to the time series instances, as explained in Section 6.1.
2. We consider four clinical critical conditions, i.e., *tachycardia*, *hypotension*, *desaturation*, and *hyperventilation* (see Table 2), which could predict the approaching of one ICU-AI episode. These medical conditions reflect the main symptoms of the Systemic Inflammatory Response Syndrome (SIRS), see [49]. Tachycardia, hypotension, and hyperventilation are quite spread in the ICU, and they are usually mentioned in general guidelines for the ascertainment of SIRS [50]. For the criteria reported in Table 2 we refer to [50]; in specific for Desaturation, we refer to [51].
3. We evaluate the mean values of HR, ABP, SaO_2 and BR for each of the most salient *8-h pattern* extracted via the SMOE scale. Depending on the values obtained (see the criteria in Table 2), we check the presence of the four clinical critical conditions. Thus, the combination of these conditions produces 16 different possible clinical situations of interest, as shown in Table 3: they represent the classes of the proposed data-driven clustering. In Fig. 17 the 16 distinct classes are represented as nodes of a graph (i.e., a four-dimensional hypercube (Q_4)).

Table 2
Critical conditions and their criteria.

Critical condition	Criterion
Tachycardia	Hearth rate $\geq$ 90 beats per minute
Hypotension	Arterial blood pressure (mean) $\leq$ 80 mmHg
Desaturation	$SaO_2 \leq 95\%$
Hyperventilation	Breath rate $\geq$ 24 breaths per minute

6.3. Results of the data-driven clustering

Histograms with the relative frequencies of the 16 data-driven clusters are shown in Fig. 18. For day 3 (see Fig. 18(a) and (b)),

Table 3
List of the 16 clinical conditions (classes of the clustering).

Class	Data driven cluster (Clinical conditions)
0	None
1	Tachycardia
2	Hypotension
3	Hypotension, tachycardia
4	Desaturation
5	Desaturation, tachycardia
6	Desaturation, hypotension
7	Desaturation, hypotension, tachycardia
8	Hyperventilation
9	Hyperventilation, tachycardia
10	Hyperventilation, hypotension
11	Hyperventilation, hypotension, tachycardia
12	Hyperventilation, desaturation
13	Hyperventilation, desaturation, tachycardia
14	Hyperventilation, desaturation, hypotension
15	Hyperventilation, desaturation, hypotension, tachycardia

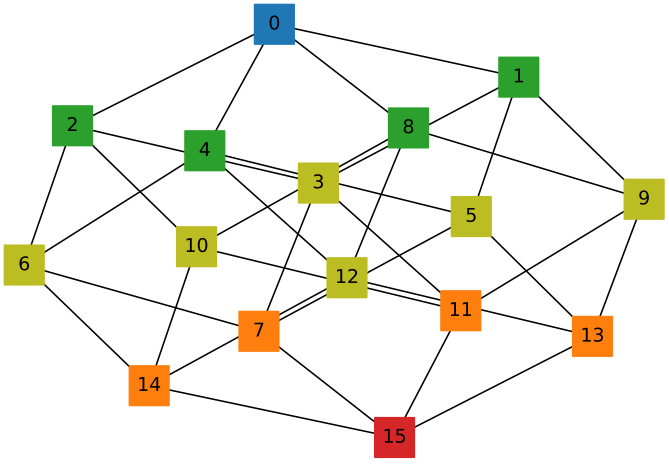

Fig. 17. Illustration of the hypercube graph (Q_4) with the 16 classes of the clustering. The numbers on the nodes denote the classes as stated in Table 3. The coloring of each node reflects the gravity of each clinical condition; (blue) No criticality, (green) one critical condition, (yellow) two critical conditions, (orange) three critical conditions, and (red) all four critical conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

two-sample Kolmogorov–Smirnov test [52] reveals that the sample distributions of the classes between *not-infected* and *infected* instances are not significantly different (p-value = 0.21). However, we can observe a completely different scenario on both days 7 and 10 (see Fig. 18(d)–(f)), where the null hypothesis of the two-samples Kolmorogov-Smirnov test is rejected (p-value = 0.0003 and p-value = 10^{-10} respectively). Hence, this analysis shows that different clinical conditions could represent an essential feature of the patterns that the CNN model captures during the learning phase. For instance, for *infected* instances, at day 10, the prevalence of at least one of these 16 conditions is around 94%, while 79% at day 7; see Fig. 18(d)–(f). Precisely, on day 10, events with hyperventilation correspond at 70% of samples, and in combination with tachycardia 23%. While a day 7 tachycardia is much more relevant and occurs in 50% of infectious samples. Therefore, the most salient 8-h subinterval of our *time series instance* can be linked to precise medical conditions, which are known to be related to the presence of an ICU-AI.

7. Conclusions

We have shown that the proposed two-step modeling of ICU-AI is simultaneously an accurate predicting tool and an interpretable model. As we have discussed, predicting an infection with our adopted definition is a challenging problem: the time to infection is determined by the start of an antibiotic treatment. Hence, the impossibility of determining the actual time of infection represents an intrinsic obstacle to building a performative prediction model based on high-frequency data.

However, the CNN can capture potentially predicting patterns by analyzing the time series of five vital sign signals. These patterns contain extra predictive information and they are only mildly correlated with the averaged quantities of the vital signals, routinely included in the traditional survival models. Moreover, we have shown as well that the SMOE scale might help physicians in clustering patients with an approaching infection.

In this work, we have considered a survival model without censoring, since ICU patients are fully monitored during their stay. However, methods based on the pseudo-observations [53] represent a solid strategy to contrast the biasing of the desired dynamic prediction due to the censoring data. In the context of LM-based survival dynamic predictions, such an approach has already been proposed; e.g., the work of [13], and in a similar way [54], presents a well-founded generalization of landmark models able to estimate how baseline and covariate effects lead to the desired dynamic predictions with left and right censoring. Likewise, a first attempt to conjugate ANN and survival predictions have recently been proposed by [55]. Despite considering only a simple MLP architecture to solve a generalized model with a logit link, this work represents a promising approach for developing new methodologies for increasing the accuracy of the survival predictions obtained by a multiplex ANN architecture fed with censored data. In comparison with the TB-ANN, we showed that an LM approach can lead to slightly more accurate and well-calibrated predictions. Despite showing almost the same overall accuracy level, the TB-ANN predictions tend to be much more sensitive on different landmark days. To our knowledge, this fact reflects the intrinsic difficulty of well-calibrating an ANN classifier when analyzing a vast amount of information coming from different data structures.

We have illustrated the methodology in a competing risks framework. However, the LM approach has recently been extended to *multi-state* models, even without the Markov assumption [56,57]. Therefore, as a further extension, we could model recurrent infections as new states in a non-Markov multi-state model, with transition hazards that might depend on the previous infections' sequence. Moreover, another future challenging direction of investigation is a sort of *inversion* of the CNN, in order to identify and classify the patterns in the signal with higher predicting power. This analysis might help in performing a more precise clustering of the patients with fore-coming ICU-AI.

Code availability

Python codes and modules are available on GitHub: https://github.com/glancia93/ICUAI-dynamic-prediction/blob/main/ICUAI_module.py.

CRediT authorship contribution statement

G. Lancia: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

None

Declaration of generative AI in scientific writing

The authors declare that no generative AI and AI-assisted technologies have been utilized during the writing process of this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102862.

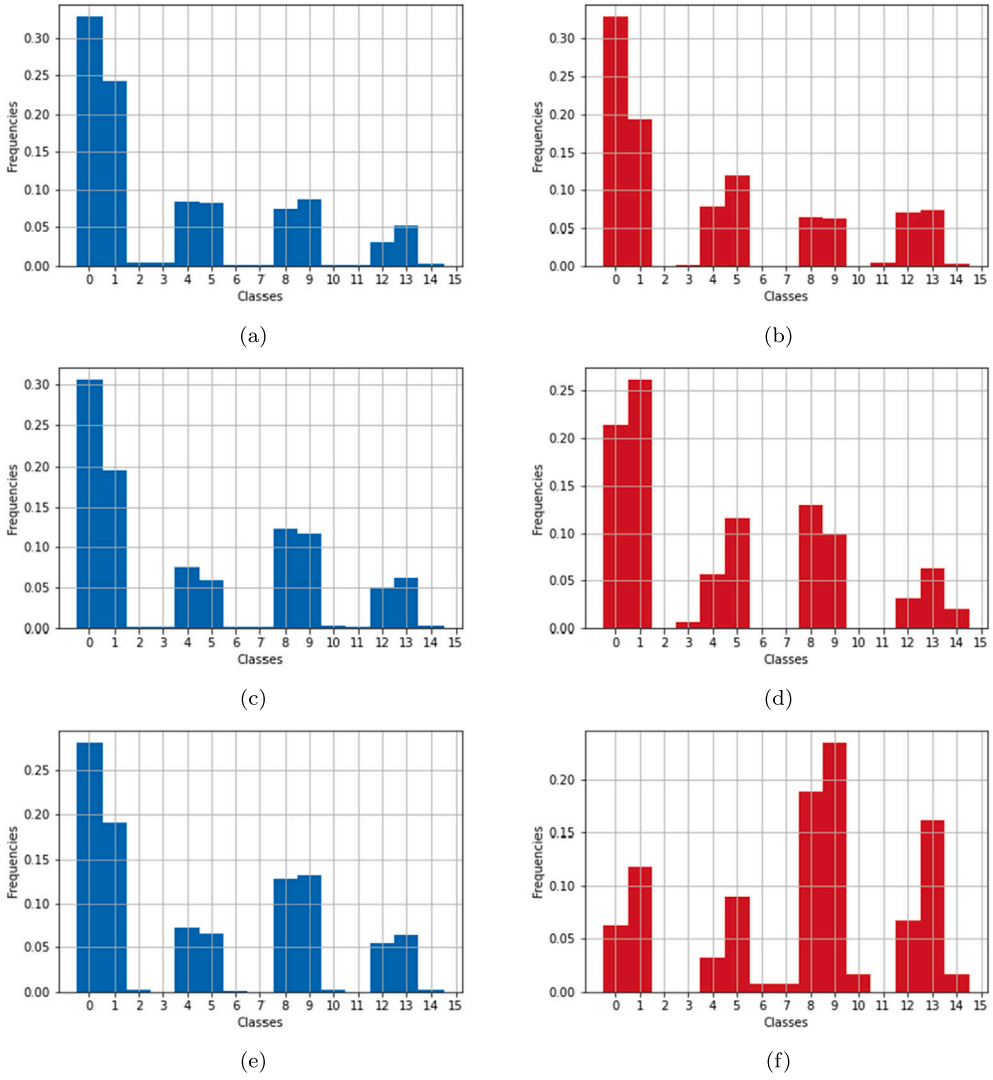

Fig. 18. Histograms the data-driven clustering approach. Bins on the x-axis represent the 16 classes. Blue histograms concern the non-infected instances, whereas the red ones the infected instances. CNN trained on day 3 is described by (a) and (b), on day 7 by (c) and (d), and on day 10 by (e) and (f). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

[1] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.

[2] Zeng Z, Hou Z, Li T, Deng L, Hou J, Huang X, Li J, Sun M, Wang Y, Wu Q, et al. A deep learning approach to predicting ventilator parameters for mechanically ventilated septic patients. 2022, arXiv preprint arXiv:2202.10921.

[3] Ivanov O, Molander K, Dunne R, Liu S, Masek K, Lewis E, Wolf L, Travers D, Brecher D, Delaney D, et al. Accurate detection of sepsis at ED triage using machine learning with clinical natural language processing. 2022, arXiv preprint arXiv:2204.07657.

[4] May R, Dandy G, Maier H. Review of input variable selection methods for artificial neural networks. In: Suzuki K, editor. Artificial neural networks. Rijeka: IntechOpen; 2011, http://dx.doi.org/10.5772/16004, chapter 2.

[5] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Springer; 1993.

[6] Cortese G, Andersen PK. Competing risks and time-dependent covariates. Biom J 2010;52(1):138–58.

[7] Proust-Lima C, Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics 2009;10(3):535–49.

[8] Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 2011;67(3):819–29.

[9] Rizopoulos D. Joint models for longitudinal and time-to-event data: With applications in R. CRC Press; 2012.

[10] Ferrer L, Putter H, Proust-Lima C. Individual dynamic predictions using landmarking and joint modeling: Validation of estimators and robustness assessment. Stat Methods Med Res 2019;28(12):3649–66. http://dx.doi.org/10.1177/0962280218811837.

[11] Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat 2007;34(1):70–85.

[12] van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. CRC Press; 2011.

[13] Nicolaie M, Van Houwelingen J, De Witte T, Putter H. Dynamic prediction by landmarking in competing risks. Stat Med 2013;32(12):2031–47.

[14] Mundhenk TN, Chen BY, Friedland G. Efficient saliency maps for explainable AI. 2019, arXiv preprint arXiv:1911.11293.

[15] Vincent J, Rello J, Marshall J. International study of the prevalence and outcomes of infection in intensive care units. JAMA 2009;302(21):2323–9.

[16] Maki DG, Crnich CJ, Safdar N. Nosocomial infection in the intensive care unit. Crit Care Med 2008;1003.

[17] Blot S, é E, Harbarth S, Asehnoune K, Poulakou G, Luyt CE, Rello J, Klompas M, Depuydt P, Eckmann C, Martin-Loeches I, Povoa P, Bouadma L, Timsit JF, Zahar JR. Healthcare-associated infections in adult intensive care unit patients: Changes in epidemiology, diagnosis, prevention and contributions of new technologies. Intensive Crit Care Nurs 2022;70:103227.

[18] Spagnolo AM. Bacterial infections: Surveillance, prevention and control. 2024, p. 181.

[19] Dantes RB, Epstein L. Combatting sepsis: a public health perspective. Clin Infect Dis 2018;67(8):1300–2.

[20] Zwerwer LR, Luz CF, Soudis D, Giudice N, Nijsten MW, Glasner C, Renes MH, Sinha B. Identifying the need for infection-related consultations in intensive care patients using machine learning models. Sci Rep 2024;14(1):2317.

[21] Klouwenberg PMK, Ong DS, Bos LD, de Beer FM, van Hooijdonk RT, Huson MA, Straat M, van Vught LA, Wieske L, Horn J, et al. Interobserver agreement of centers for disease control and prevention criteria for classifying infections in critically ill patients. Crit Care Med 2013;41(10):2373–8.
[22] Gandin I, Scagnetto A, Romani S, Barbati G. Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit. J Biomed Inform 2021;121:103876.
[23] Deng Y, Ma Y, Fu J, Wang X, Yu C, Lv J, Man S, Wang B, Li L. A dynamic machine learning model for prediction of NAFLD in a health checkup population: A longitudinal study. Heliyon 2023;9(8).
[24] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5.
[25] Liu YH. Feature extraction and image recognition with convolutional neural networks. In: Journal of physics: conference series. Vol. 1087, IOP Publishing; 2018, 062032.
[26] Zheng H, Fu J, Mei T, Luo J. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 5209–17.
[27] Lou G, Shi H. Face image recognition based on convolutional neural network. China Commun 2020;17(2):117–24.
[28] Kagaya H, Aizawa K, Ogawa M. Food detection and recognition using convolutional neural network. In: Proceedings of the 22nd ACM international conference on multimedia. 2014, p. 1085–8.
[29] Kwon D, Natarajan K, Suh SC, Kim H, Kim J. An empirical study on network anomaly detection using convolutional neural networks. In: ICDCS. 2018, p. 1595–8.
[30] Naseer S, Saleem Y, Khalid S, Bashir MK, Han J, Iqbal MM, Han K. Enhanced network anomaly detection based on deep neural networks. IEEE Access 2018;6:48231–46.
[31] Staar B, Lütjen M, Freitag M. Anomaly detection with convolutional neural networks for industrial surface inspection. Procedia CIRP 2019;79:484–9.
[32] Borovykh A, Bohte S, Oosterlee K. Conditional time series forecasting with convolutional neural networks. In: Lecture notes in computer science/lecture notes in artificial intelligence. 2017, p. 729–30.
[33] Selvin S, Vinayakumar R, Gopalakrishnan E, Menon VK, Soman K. Stock price prediction using LSTM, RNN and CNN-sliding window model. In: 2017 international conference on advances in computing, communications and informatics. icacci, IEEE; 2017, p. 1643–7.
[34] Livieris IE, Pintelas E, Pintelas P. A CNN–LSTM model for gold price time-series forecasting. Neural Comput Appl 2020;32(23):17351–60.
[35] Guo-yan X, Jin Z, Cun-you S, Wen-bin H, Fan L. Combined hydrological time series forecasting model based on CNN and MC. Comput Mod 2019;(11):23.
[36] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: ICLR. 2015, URL: http://arxiv.org/abs/1412.6980.
[37] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78(1):1–3.
[38] Wilks DS. Statistical methods in the atmospheric sciences, vol. 100, Academic Press; 2011.
[39] Spitoni C, Lammens V, Putter H. Prediction errors for state occupation and transition probabilities in multi-state models. Biom J 2018;60(1):34–48.
[40] Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence. Gaithersburg, Maryland; 2020.
[41] Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 2021;76:89–106.
[42] Castelvecchi D. Can we open the black box of AI? Nat News 2016;538(7623):20.
[43] Al-Najjar HA, Pradhan B, Beydoun G, Sarkar R, Park H-J, Alamri A. A novel method using explainable artificial intelligence (XAI)-based Shapley additive explanations for spatial landslide prediction using time-series SAR dataset. Gondwana Res 2022.
[44] Neubauer MS, Roy A. Explainable AI for high energy physics. 2022, arXiv preprint arXiv:2206.06632.
[45] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell 2020;2(10):573–84.
[46] Dave D, Naik H, Singhal S, Patel P. Explainable ai meets healthcare: A study on heart disease dataset. 2020, arXiv preprint arXiv:2011.03195.
[47] Lancia G, Goede I, Spitoni C, Dijkstra H. Physics captured by data-based methods in El Niño prediction. Chaos 2022;32(10).
[48] Lancia G, Durastanti C, Spitoni C, De Benedictis I, Sciortino A, Cirillo EN, Ledda M, Lisi A, Convertino A, Mussi V. Learning models for classifying Raman spectra of genomic DNA from tumor subtypes. 2023, arXiv preprint arXiv:2302.08918.
[49] Chakraborty RK, Burns B. Systemic inflammatory response syndrome. 2019.
[50] Comstedt P, Storgaard M, Lassen AT. The systemic inflammatory response syndrome (SIRS) in acutely hospitalised medical patients: a cohort study. Scand J Trauma Resusc Emerg Med 2009;17(1):1–6.
[51] Hafen BB, Sharma S. Oxygen saturation. StatPearls, StatPearls Publishing; 2022.
[52] Hodges JL. The significance probability of the Smirnov two-sample test. Ark Mat 1958;3(5):469–86.
[53] Andersen PK, Pohar Perme M. Pseudo-observations in survival analysis. Statist Methods Med Res 2010;19(1):71–99.
[54] Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. Stat Med 2013;32(18):3089–101.
[55] Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. IEEE J Biomed Health Inform 2020;24(11):3308–14.
[56] Putter H, Spitoni C. Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator. Stat Methods Med Res 2018;27(7):2081–92.
[57] Hoff R, Putter H, Mehlum IS, Gran JM. Landmark estimation of transition probabilities in non-Markov multi-state models with covariates. Lifetime Data Anal 2019;25(4):660–80.