Journal of Biomedical Informatics 121 (2021) 103876

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to Intensive care unit

Ilaria Gandin [a,*], Arjuna Scagnetto [a], Simona Romani [b,c], Giulia Barbati [a]

[a] Department of Medical Sciences, Biostatistics Unit, University of Trieste, Trieste, Italy
[b] Department of Medical Sciences, University of Trieste, Trieste, Italy
[c] Cardiothoracovascular Department, Azienda Sanitaria Universitaria Giuliano Isontina, Trieste, Italy

ARTICLE INFO

Keywords:
Deep learning
Interpretability
Electronic health records

ABSTRACT

Interpretability is fundamental in healthcare problems and the lack of it in deep learning models is currently the major barrier in the usage of such powerful algorithms in the field. The study describes the implementation of an attention layer for Long Short-Term Memory (LSTM) neural network that provides a useful picture on the influence of the several input variables included in the model.

A cohort of 10,616 patients with cardiovascular diseases is selected from the MIMIC III dataset, an openly available database of electronic health records (EHRs) including all patients admitted to an ICU at Boston's Medical Centre. For each patient, we consider a 10-length sequence of 1-hour windows in which 48 clinical parameters are extracted to predict the occurrence of death in the next 7 days. Inspired from the recent developments in the field of attention mechanisms for sequential data, we implement a recurrent neural network with LSTM cells incorporating an attention mechanism to identify features driving model's decisions over time.

The performance of the LSTM model, measured in terms of AUC, is 0.790 (SD = 0.015). Regard our primary objective, i.e. model interpretability, we investigate the role of attention weights. We find good correspondence with driving predictors of a transparent model (r = 0.611, 95% CI [0.395, 0.763]). Moreover, most influential features identified at the cohort-level emerge as known risk factors in the clinical context.

Despite the limitations of study dataset, this work brings further evidence of the potential of attention mechanisms in making deep learning model more interpretable and suggests the application of this strategy for the sequential analysis of EHRs.

1. Introduction

Automated decision support systems are currently one of the most challenging goals in medical research. Given the great steps forward made by Artificial Intelligence (AI) in several fields for complex tasks, as automated driving, algorithmic trading, plant production management, customer behaviour, fraud detection, financial loan approval, great expectation has been put on the development of automated systems for medical applications.

One of the most common complex tasks that a clinician is going to face is the prognostic evaluation of a patient. Based on clinical and instrumental parameters, physicians must be able to timely recognize a worsening of the disease and act consequently, modifying patients' treatment, in order to avoid an adverse outcome. This evaluation could be supported by an algorithm that integrate vast amount of information more efficiently and yield more precise predictions. The way has been paved by the recent awareness of the potential enclosed in large amounts of raw data and the consequent effort to make electronic health records (EHRs) available for machine learning research. EHRs are the whole set of digital clinical data produced at single-patient level in health care institutions, that despite being raw and noisy, represent a valuable source of information for several reasons: data can be extracted in massive volumes; information spans the assistance provided by different units of the hospital (clinical units, laboratories, prescriptions, etc.); data can be recorded through time to form a well-ordered sequence of events.

Abbreviations: AUC, Area under the curve; ECG, Electrocardiogram; EHR, Electronic health records; ICU, Intensive care unit; LOCF, Last observation carrier forward; LSTM, Long short-term memory; NN, Neural network; RNN, Recurrent neural network; ROC, Receiving operator curve.

* Corresponding author.
E-mail address: igandin@units.it (I. Gandin).

https://doi.org/10.1016/j.jbi.2021.103876
Received 16 April 2021; Received in revised form 14 July 2021; Accepted 20 July 2021
Available online 27 July 2021

However, research in machine learning for disease risk prediction has proceeded with some problems. Compared to other fields, massive datasets that are required for these algorithms are difficult to obtain, mainly due to the sensitivity of personal medical information. The resulting scarcity of shared benchmarks to evaluate competing models and the difficulty in measuring the methodological progress had consequences on the transferability in practical settings [1].

If data availability is an obstacle that could be possibly overcome in the next future with the adoption of new strategies for data sharing [2], there is another important issue which is far from being solved. Most of the machine learning algorithms, in particular deep learning algorithms, produce models that are hard to be interpreted. Commonly referred as black-box models, such models charge the performance improvement with a cost: models are that complex that underlying mechanism cannot be grasped and only indirect analysis can be applied to get an insight into the role of the different input features.

Clinicians take the responsibility of final decisions and are supposed to justify their actions with causal relationships, in accordance with the principles of the domain. Nevertheless they are not new to black-box algorithms [3]. Currently, in their practice, medical doctors base clinical decision on information without fully understanding how it is generated (e.g. result of specialistic laboratory analysis, output from medical machineries). This practice is obviously well accepted and the reason is the application of regulatory procedures that guarantee reliability, thus promoting trust in such information. For the same reason, research in machine learning interpretability is fundamental. Otherwise, even in front of an improvement in accuracy, the compromise on model transparency and accountability will hardly be accepted in the medical context. In addition to this, it is worth mentioning that safety and liability of AI application are becoming imperative aspects for the new regulating legal frameworks, which are emerging (in particular in the European area [4]) to protect individual fundamental rights related to human dignity and privacy protection.

Despite an intrinsic limit in the black-box problem, due to the impossibility of finding explanations at the same time simple and with perfect fidelity with respect to a model that is supposed to be highly complex [5], improvements on interpretability can provide important information on models' behaviour and help to provide the trustworthiness needed for its usage [6]. In this view, desirable objectives of interpretability techniques are: the detection of potential bias in the training data, unfairness of the algorithms (for certain social groups or individuals), generalization failures and more in general, providing insights to further improvements of the model [7].

Perhaps the most emblematic example of black-box models are deep neural networks (NN), which are made by a multitude of computational blocks organized in layers and inter-connected, each one learning weights to create a custom processing of input information. Among the types of NN particularly suited for data-streams and time-series data, Recurrent Neural Networks (RNN) are gaining popularity in prediction tasks based on EHRs [8], as well as the development of these architectures that enhance interpretability. One of the most promising approaches seem to be the attention models, that help the algorithm to focus on relevant elements in the sequence of data. In addition to improved performance, attention mechanisms have been shown to provide elements for clinical interpretation [9–11]. In particular, a recent study has described an attention model that acts at the level of input variables and facilitate interpretability in the case of EHRs [12].

In the light of the above considerations, we have investigated a deep learning approach for mortality risk prediction in cardiovascular patients that incorporate an attention layer, which can be related to variable importance. To this aim, we analyse the MIMIC III dataset [13], a collection of EHRs of an Intensive Care Unit (ICU) at the Beth Israel Deaconess Medical Centre. Patients data as demographic information, vitals, labs, procedures and medications, were extracted from the informative system of the hospital, de-identified and arranged in a database which is currently available for research purposes. Our task is focused on cardiovascular diseases for two main reasons. The first one is related to possible benefits in the clinical practice. Evaluation of the severity of patients admitted to ICUs is a problem that has been widely studied and the standard approach consists in the application of general scores (like SAPS [14], APACHE [15]) that take in consideration only a small set of input variables. However, currently there is no risk score specifically developed for cardiological critical care, that could provide an accurate prognostic prediction giving the heterogeneous mix of conditions encountered in modern ICU [16]. The second reason is linked to the main purpose of this study: having restricted our analysis on a group of quite homogeneous patients, we can more easily compare the finding of our interpretability investigation (in term of impact of input features) with the clinical knowledge in disease managing. Our research group can count on a long-standing collaboration with expert cardiologists. In line with the setting of standard prognostic scoring systems, that are typically measured in the first 12–24 h, we collect measurements in the first 10 h of admission at the ICU and use them to estimate the risk of mortality within the first 7 days of hospitalization.

2. Material and methods

2.1. Data

The study considers all patients of the MIMIC III database with primary diagnosis falling into a cardiovascular category. The selection is done based on the discharging ICD9 diagnostic codes (see Table A1 in Appendix A for the full list of codes included). In case of multiple stays per patient, given that our outcome of interest is mortality we only consider the last hospitalization. The setting of our study is illustrated in Fig. 1: starting from the moment of the first record of vital signs, measurements are collected in 10 h that are divided in 10 1-hour windows (windows length was arbitrarily set to match the time-resolution of variables after a first exploration of the data). Such variables are used to predict death in the next 146-hours window (7 days). Since our aim is to predict mortality using the first 10 h measurements, we excluded subjects who died or were discharged within the 10 h.

The outcome variable is defined as a binary label: subjects were labelled as 1 if death occurs within the period $[t_{10}, t_f]$, 0 if subjects survive or discharge happens. Thus, we are interested in death occurrence, disregarding the time of the event. Predictor variables can be divided in 3 categories (see Table A2 in Appendix A for the full list): demographic variables, basic information recorded at the admission; monitored vitals, measurements of vital functions with continuous-monitoring systems; clinical variables, measurements of other clinical features (with highly-variable recording time). Categorical variables are transformed into dummy variables. After this, the final number of predictors is 48.

We extract all records related to predictor falling within the 10-hours time-window. Different variable types have different recording frequencies. Monitored vitals, that are recorded at high frequency, are reshaped and aggregated using the mean so that a measurement is obtained for each 1 h-window (forming a time-series of 10 steps). Demographic variables are clearly recorded only once, thus in our setting

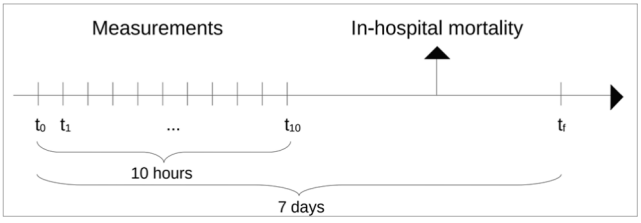

Fig 1. Framework and timeline. Predictors collected in the first 10 h from hospitalization were included in a binary classification model to predict mortality within 7 days.

they are considered constant over time. Remaining clinical variables are characterised by a low-frequency recording, thus they are aggregated over the 10 h and considered constant over time. In all these procedures, the general rule has been to exclude predictors with missing rate $\geq$25%. For the remaining variables, missing values are imputed with a two-step procedure: 1) Last Observation Carried Forward (LOCF) (only for monitored vitals); 2) Imputation based on sample mean. The resulting input dataset consists of 48 variables having 10 repeated measurements for each subject (for demographic and clinical variables that are constant in time, measurements are duplicated). All variables are normalized.

The problem is designed as a binary classification task where labels are fixed, instead predictors are (in part) time-varying. The cohort is randomly divided in training and test set (70% and 30% respectively). Stratified 10-fold cross-validation was performed on the training set and the best model was evaluated on the test set. Similarly to [17], this procedure is repeated for 10 times in order to avoid sampling bias due to the training-test data split. At the end of the iteration, 10 models are evaluated in terms of Area Under the ROC curve (AUC-ROC).

Unfortunately, some input variables required for standard prognostic scoring systems (e.g. Glasgow come score) were not available in MIMIC III for the selected cohort, thus a comparison between standard scores and our model was not possible at this stage.

All data processing has been performed with R 4.0.2, including packages "pROC" [18] and "caret" [19]. Python 3.7.7 has been used for the implementation of the model (described below in Sections 2.2 and 2.3), which is based on TensorFlow backend and Keras API. The core program codes and a tutorial are available on Github (https://github.com/ilariagnd/CardioICURisk).

2.2. Deep learning approach

Deep learning models have pushed the research on clinical predictions from EHR. In particular, starting from the work of Lipton et al. [20], several studies have described successful applications of Recurrent Neural Network (RNN) architectures (see [21–23] for very recent works, see [24] for a comparative review). In particular, encouraged by the recent findings of Harutyunyan et al. [1] for mortality prediction from time-series data (a problem very similar to our research question), we chose to implement a RNN with Long Short-Term-Memory (LSTM) cells.

RNN has been originally developed for tasks related to language processing, a type of data where observations (i.e. words) appear in an ordered sequence that must be taken into account. This setting can be applied in the case of time-series problems like the case of our study. RNN accomplish this task with a recursive weights estimation together with an additional latent variable (called "hidden state") that is iteratively updated and keeps memory of the previous steps.

However, in case of long sequences, with several time-steps, the algorithm could suffer from a problem called vanishing/exploding gradient. As the word suggests, in the back-forward propagation the recursion could easily make the gradient to increase or decrease very fast. One of the solutions to this problem is to enrich the structure of the net and introduce the LSTM cells that can be represented by the following equations [25]:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$H_t = O_t \odot tanh(C_t)$$

Together with the input variables and the hidden state (H_t), LSTM include an additional source of information that represent the "long-term" memory (C_t) and this is achieved with the use of layers called "gates" (I_t input gate, F_t forget gate, O_t output gate) that are able to filter relevant information to be passed in the next time-step.

In our implementation, hyperparameters tuning followed a grid search approach. The result was a LSTM layer with 128 units, sigmoid activation function, dropout 20% [12]. Training was performed with RMSPROP optimizer (learning rate = 0.001, rho = 0.9, epsilon = 1e−08) in 20 epochs [26], batch size was 64.

2.3. Attention mechanism

Concerning model's interpretability, a big step forward has been made by Kaji et al. [12] and Remy [27]. As showed by the authors, attention-based LSTM provide useful output that could facilitate the understanding at the level of input variables. Originally developed for machine translation models to improve their performance [28], the attention mechanism is an additional layer to the LSTM that support long-term dependencies by focusing on specific element of the sequence (words in case of language translation). Attention based neural network have successfully applied for very different problems, like medical computer vision tasks [29], analysis of ECG patterns [30] and blood pressure response [31].

The attention mechanism of Kaji et al. is obtained at the level of input variables with a dense layer with softmax activation function, so that for each feature j an attention vector a_j of length T (number of time-steps) is learned with $|a_j| = 1$. Before being fed to the LSTM, input features are weighted by attention vectors a_j:

$$X_{new} = A \odot X$$

where X represents the $T \times p$ input data (p number of variables) for a single observation and a_j is the j-th column of A. In this setting, softmax works as a normalization on the time dimension, making it possible to interpret a_{jt} as contribution of feature j within a fixed time step t.

We included in our LSTM a similar attention mechanism. However, since we are more interested in understanding the global contribution of each feature, we propose to apply the same procedure on the transposed input. Using the compact notation of Kaji et al., let's denote by a_t the vector of length p representing the input at time t, then attention vectors are obtained as following:

$$a_t = softmax(x_t W_t)$$

with $|a_t| = 1$. In this way attention vectors can be used to understand the global contribution of the j-th feature aggregating values $a_{j1}, \ldots, a_{jT}$ through time. Then $X_{new} = A \odot X$ with $a_t$being the the t-th row of A. Raw softmax activations a_t form an activation map that can be extracted and further analysed. A graphical representation describing the structure of the model can be found in Appendix A (Figure A1).

2.4. Analysis of activations

For each subject we can extract a matrix of values $p \times T$ that sum to one over the rows (feature dimension) and we calculate the average of each column. In this way, for each subject and for each feature we obtain a single activation that can be further median-aggregated (because of the presence of skewness) over the cohort to obtain a marginal attention map for features at population-level. To obtain a robust variable ranking based on attention, we calculate the attention map in all the 10-iteration models and then order features with respect to the average over the iterations.

We further investigate whether activation could be interpreted as the relative importance of input features comparing it with an inherently interpretable model. For the sake of brevity, we focus on one single model (the first one in our 10 iterations). Given the large number of input variables, we opted for a penalized logistic regression model (LR). Since we are also dealing with time-varying independent variables, for

this model we flatten the time dimension using 4 sample statistics: mean, standard deviation, maximum, minimum. Hyperparameters λ (penalty weight) and α (indicator parameter for L1 or L2 norm) were also tuned with a grid search in a 10-fold cross-validation.

Finally, we represent the individual attention activations (in form of heatmap) for two sample subjects admitted to the ICU with similar clinical conditions but with different outcome (survived and deceased).

3. Results

We obtained a dataset of 10,616 individuals where mortality rate was 6.6%. Quality control procedures identified 48 variables as reported in Table S2.

Results of the models are reported in Table 1. On average, the LSTM net achieved to predict mortality with AUC-ROC of 0.790 (SD = 0.015). In the case of LR, the ability to predict mortality in terms of AUC-ROC is 1.3% lower compared to LSTM model. Although this is not the main purpose of this study, it is worth to note that a regression model is able to provide a performance similar to the one of the LSTM.

Ranking input features by attention weights, we find norepinephrine, phenylephrine, creatinine, male, BUN (Blood Urea Nitrogen), age, respiratory rate, oxygen saturation (SpO2), systolic BP (Blood Pressure), paced rhythm in the first 10 positions (Fig. 2).

The comparison between the attention map values and the regression coefficients for a single model is represented in Fig. 3. In the case of LR, for each time-varying variable we estimated the coefficients for 4 statistics (min, max, mean, sd), whereas in the LSTM we obtained one weight for each variable. Given the way penalized regression algorithm deals with correlated variables, it is reasonable to consider only one statistics for each input feature and specifically the one with absolute maximum value of the regression coefficient. The correlation between activations and maximum regression estimates (logit scale) for the 48 features is 0.611 (95% CI [0.395, 0.763]), which indicates a quite strong relationship between the two quantities. Moreover, in the LR model it can be observed that for predictors that were originally time-varying and then aggregated, the stronger impact on the outcome in most of the cases is obtained considering SD, MIN or MAX through the time-windows (instead of MEAN).

So far, attention activations were analysed in aggregated form. An interesting aspect of the attention mechanism is the possibility to extract the matrix of activations related to a single individual. Fig. 4 represents in details the case of two example individuals with very similar clinical conditions attributable to heart failure (list of the discharging ICD-9 diagnostic codes reported in Table A3 in Appendix A): Patient A survived, whereas patient B did not. In this chart it is possible to note that Patient A had a very low saturation only in the first hour and such vital parameter turned out to be the most attended feature. Saturation was the main driving feature also for patient B, but in this case the parameter did not improved over time. Such representation suggests that Patient A was admitted with a condition of respiratory failure that rapidly improved and had a favourable outcome; in the case of Patient B, it was not possible to restore a normal blood oxygenation, consequently his respiratory rate gradually increased in the attempt to compensate and finally he had an adverse outcome.

4. Discussion

In this work we implemented a deep learning architecture based on multivariate time series to predict an adverse event (death) based on electronic health records. Thanks to work of Kaji et al. that has recently introduced the attentioned LSTM neural network for the EHRs, we were able to apply the same strategy and obtain activation maps at predictors-level.

Table 1
Models' results. For each model, the average and standard deviation of AUC-ROC of the 10 iterations are reported.

Model	AUC-ROC Mean	AUC-ROC SD
LSTM	0.790	0.015
LR	0.779	0.013

The results of LSTM are quite far from what obtained by Harutyunyan et al., that for a similar task (mortality risk prediction using MIMIC-III) found a model showing higher performance: AUC-ROC 0.855, 95% CI [0.835,0.873]. In their case, the cohort was around double-sized with respect to the one in analysis here and 17 variables were considered for 48 h instead of 10. Those aspects could be related to the difference observed in performance.

Although this is not the main purpose of this study, it is worthy to note that adopting a penalised regression model we obtain a performance very similar to the one of the LSTM, which confirms previous findings in similar settings [1,12,32,33]. However, it should be noted that despite the high number of predictors included in our study, only a fraction of them is time-varying and this could reduce the chance for LSTM to detect long range dependencies.

As regards our primary objective, model interpretability, this work contributes to encourage the use of attention layers in LSTM to obtain information on the relationship between outcome and predictors. The first 10 most important predictors based on attention weights are in line with the clinical experience of cardiovascular disease treatment. Norepinephrine and phenylephrine are two inotropic drugs used in case of haemodynamic instability, a serious condition characterized by unstable blood pressure, which can cause inadequate blood flow to peripheral organs. Several studies indicate that their use, especially at high dosages, is associated with increased mortality [34,35]. Creatinine and BUN (Blood Urea Nitrogen) are laboratory biomarkers used to investigate renal function. An increase in their values indicates the presence of renal failure, which has been shown to correlate with increased mortality in patients admitted to intensive care [36]. Male gender and advanced age are well-known for being associated with a higher cardiovascular risk [37]. Respiratory rate, oxygen saturation (SpO2), systolic blood pressure are crucial vital parameters constantly monitored in intensive care unit, since they can change suddenly in case of clinical worsening of the patient. These parameters are normally considered in mortality scores and principal morbidities for ICUs [15,38]. Paced rhythm is one of the possible heart rhythms detected by telemetry monitoring (a continuous registration of ECG) performed in Intensive Care Unit. A permanent or temporary pacemaker is needed when the conduction system of the heart is damaged and it is necessary to implant an external device to generate the electrical impulse which triggers heart contraction.

Activation distributions are also investigated through a comparison with an interpretable model, showing that there is a satisfactory agreement in terms of variable importance. Of note, when predictors that were originally time-varying were aggregated using standard LR approach, in most of the cases the stronger impact on the outcome was obtained considering their variability (SD) or the extreme values (MIN or MAX) through the time-windows instead of their mean, indicating that time-to-time variability of the predictor should be considered instead of a marginal aggregated value, thus highlighting the need for time-series approaches. We are not expecting (nor pursuing) a complete agreement between the two variable rankings since it would make no sense to get such concordance for a complex non-linear model [5]. Instead, this step confirms that activations analysis can be of use in the direction of interpretability objectives [39]: detection of bias in the training set to ensure impartiality, assess robustness against model manipulations and verify the meaningfulness of underlying variables. In this view, even though the study was performed on a limited dataset where the deep learning approach brought little increase in the performance, our results confirm the benefits of using attention mechanisms for interpretability as an approach suitable for ultrahigh dimensional settings, where regularisation methods (widely used and "explainable"

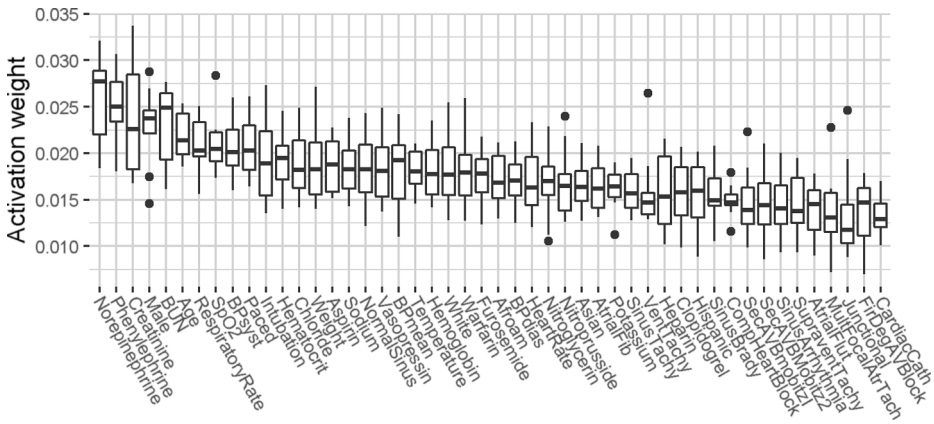

Fig 2. Variable ranking based on attention weights. For each variable, the attention weights of the 10-iteration models are represented in a box-plot. Variables are ordered with respect to the mean of the 10 values.

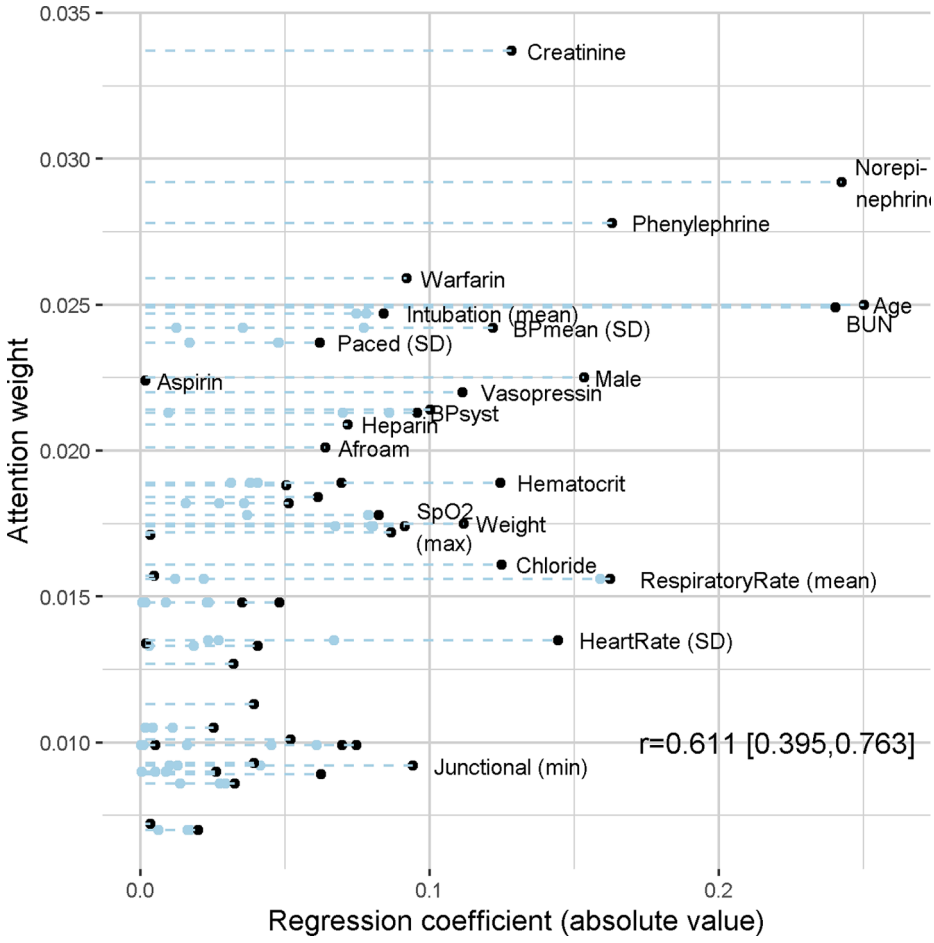

Fig 3. Comparison between LSTM attention weight and LR estimates. On the x-axis, the absolute value of regression estimates is reported and black dots indicates the maximum among the 4 aggregation functions (mean, standard deviation, maximum, minimum). The y-axis corresponds to the predictor activations obtained for the LSTM.

methods) underestimate the level of noise and select spurious correlations [40,41].

The individualized representation of attention activations in form of heatmap show the potential enclosed in these models for personalized predictions and the present work contributes to complement the results of previous studies. We provide a map that highlights the overall impact of variables, which was not possible to obtain in the setting of Kaji et al. [12] where the focus was on time-steps and variable ranking could be made only within the single time-step. The progress made in the last years in the in deep learning research for the medical field shows that deep learning models could adequately embody the concept of personalized risk prediction [42–44]. Attention mechanisms offer the possibility to retrieve driving features at patient-level: individual risk prediction is paired to individual attention map. Interpretation of activations at the single subject level is not straightforward: more insightful clinical interpretations of single cases require the knowledge of the

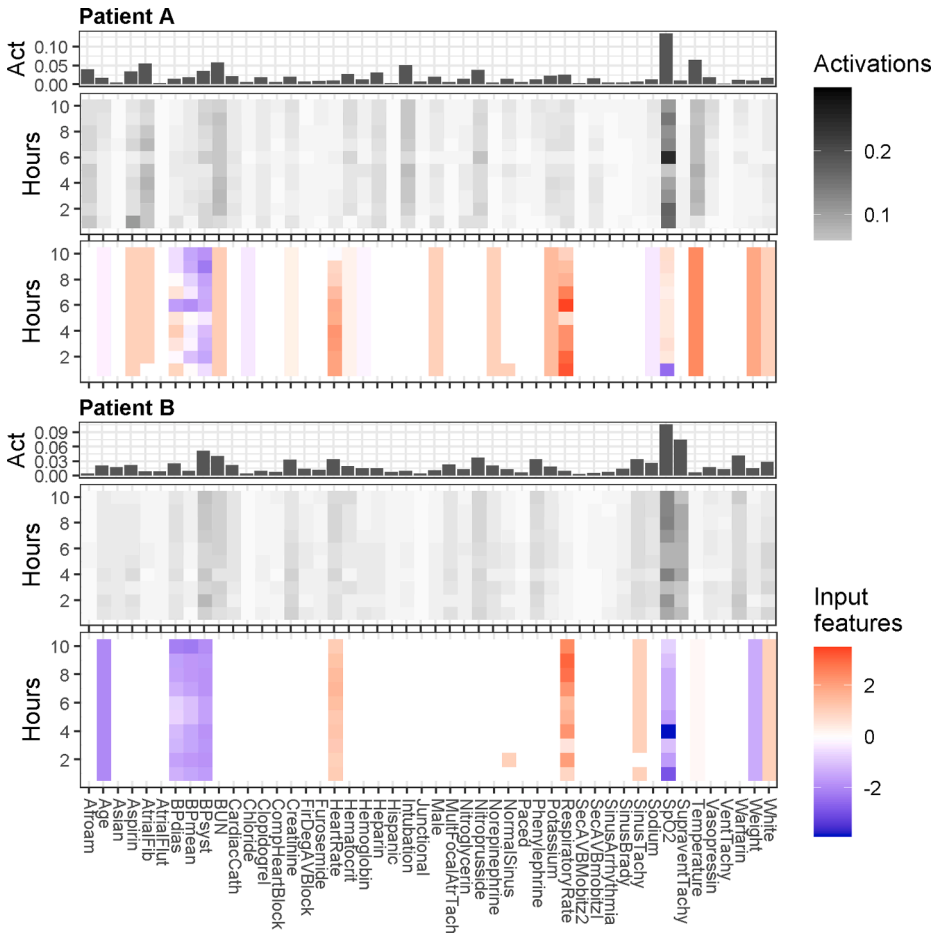

Fig 4. Individual-level attention map. The cases of Patient A and Patient B are presented. For each individual, we report (from bottom to top): heatmap representing the value on input features; heatmap representing the value of attention activations; barplot reporting the time-average contribution of each feature.

complete clinical history of patient (which is not possible for MIMIC III). However, it should be noticed that such patient-level information could not be obtained with more standard statistical approaches: even in the presence of high-order interactions, adequate computational capacity, and data availability for all possible strata, standard statistical regression approach will produce fixed "subgroup-level" weights.

There are several limitations in our study. First of all, in our framework patients that are discharged or had died before the 10th hour are excluded from the study and this is a possible source of selection bias. Secondly, the comparison of models' performance cannot be considered conclusive since our working setting suffers from drawbacks: modest number of observations, low number of time-varying covariates and lack of out-of-sample data for external validation. Third, our investigation on activations extracted from the attention mechanism does not provide practical recommendations to achieve the objectives of model interpretability aforementioned. In particular, the model is not able to distinguish between positive and negative associations among predictors and outcome. However, the research in interpretability is very active, with new tools and strategies emerging. For example, in a very recent work, Lauritsen et al. [21] describes a model that using the concepts of global parameter importance estimation and local explanation summary, provides not only relevant clinical parameters at patient-level for a given point in time, but also a population-based prospective, estimating relevance scores across the entire population. Fourth, the common choice of AUC-ROC as measure of performance for a binary classification problem is not free from concern when dealing with imbalanced data. In this situation a global measure summarises results over regions of the ROC space that would not be used in practice. Also in this direction new methods and strategies are emerging, like the one recently proposed by Carrington et al. that introduces a new interpretation of AUC based on groups of predicted risks, making it possible to interpret the model's performance in each group and to realize where the model could be weak [45].

Further development of our investigation could be the implementation of LSTM with time-sequence also as output targets. In such settings, where the aim is to model the time-to-event for the endpoint, transparent statistical methodologies suitable to carry out the comparison would be survival regression approaches. LSTM could be compared with joint longitudinal and survival models [46] and frailty survival models [47], that handle endogenous time-varying covariates and time-to-event outcomes, as long as the dimensions of the dataset (both in number of observations and in number of variables) are limited, otherwise computational complexity for these approaches could be a serious issue. Another avenue for future research could be the use of more complex input data from different sources and modalities (like time-based data, unstructured text, images, omics data) and extend interpretability analysis to models incorporating such information fusion. As suggested by Holzinger et al. [48], this could be achieved investigating the algorithm class known as Graph Neural Networks, that are based on multimodal embeddings and can be used either with attention mechanisms to identify relevant graphs structures (in terms of edges and nodes) or in combination with model-agnostic interpretability techniques.

5. Conclusions

In this study, we approached the problem of interpretability in time-

series deep learning models. EHRs were considered to predict the risk of mortality within 7 days from ICU admission for a cohort of cardiovascular patients. We implemented an attention model aimed to explore the role of individual variables in predictions. Thanks to the comparison with a transparent model and clinical interpretation of a single-case, we obtained evidence that the method leads to the identification of relevant predictors.

More efforts are required in interpretability research to identify tools and strategies that can be used both by data scientist, in models' developing phase to ensure reliability, and by final users in clinical practice, in order to make timely and context-aware decisions, and our study supports the way of attention mechanisms. We believe this challenge cannot be avoided in the next future if deep learning algorithms will be integrated in clinical decision support systems.

CRediT authorship contribution statement

Ilaria Gandin: Conceptualization, Methodology, Formal analysis, Software, Writing - review & editing. **Arjuna Scagnetto:** Data curation. **Simona Romani:** Validation, Writing - review & editing. **Giulia Barbati:** Conceptualization, Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful for the valuable comments of the anonymous reviewers.

This work was supported by a grant from the University of Trieste: "Heart And machine Learning (HEAL)", Finanziamento per la Ricerca di Ateneo – FRA 2018.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103876.

References

[1] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (2019).
[2] M. Wilkinson, M. Dumontier, I. Aalbersberg, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016) 1–9. https://www.nature.com/articles/sdata201618%22.
[3] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent Rep. 49 (2019) 15–21.
[4] D. Schneeberger, K. Stöger, A. Holzinger, The European Legal Framework for Medical AI, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, pp. 209–226.
[5] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.
[6] M. Sendak, M.C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, et al., The human body is a black box": Supporting clinical decision-making with deep learning. FAT* 2020 - Proc 2020 Conf Fairness, Accountability, Transpar. 2020, pp. 99–109.
[7] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56.
[8] B. Shickel, P.J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. arXiv, 2017.
[9] E. Choi, M.T. Bahadori, J.A. Kulas, A. Schuetz, W.F. Stewart, J. Sun, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process Syst. (2016) 3512–3520.
[10] H. Song, D. Rajan, J.J. Thiagarajan, A. Spanias, Attend and diagnose: clinical time series analysis using attention models, in: 32nd AAAI Conf Artif Intell AAAI 2018, 2018, pp. 4091–4098.
[11] S.A. Kamal, C. Yin, B. Qian, P. Zhang, An interpretable risk prediction model for healthcare with pattern attention, BMC Med. Inform. Decis. Mak. 20 (2020) 307, https://doi.org/10.1186/s12911-020-01331-7.
[12] D. Kaji, J. Zech, J. Kim, S. Cho, N. Dangayach, A. Costa, et al., An attention based deep learning model of clinical events in the intensive care unit, PLoS One 14 (2019), e0211057.
[13] A.E. Johnson, T.J. Pollard, L. Shen, L.H. Lehman, M. Feng, M. Ghassemi, et al., Data Descriptor: MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 1–9.
[14] J.R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, et al., A simplified acute physiology score for ICU patients, Crit. Care Med. 12 (1984) 975–977.
[15] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients, Crit. Care Med. 34 (2006) 1297–1310.
[16] K. Strand, H. Flaatten, Severity scoring in the ICU: a review, Acta Anaesthesiol. Scand. 52 (2008) 467–478.
[17] J. Zhao, Q. Feng, P. Wu, R. Lupu, R. Wilke, Q. Wells, et al., Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction, bioRxiv (2018) 366682.
[18] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, et al., pROC: An open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics 12 (2011).
[19] M. Kuhn, caret Package, J. Stat. Softw. 28 (2008) 1–26. http://www.jstatsoft.org/v28/i05/paper.
[20] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to diagnose with LSTM recurrent neural networks, in: 4th Int Conf Learn Represent ICLR 2016 - Conf Track Proc. 2016.
[21] S.M. Lauritsen, M. Kristensen, M.V. Olsen, M.S. Larsen, K.M. Lauritsen, M. J. Jørgensen, et al., Explainable artificial intelligence model to predict acute critical illness from electronic health records, Nat. Commun. 11 (2020).
[22] N. Tomašev, X. Glorot, J.W. Rae, M. Zielinski, H. Askham, A. Saraiva, et al., A clinically applicable approach to continuous prediction of future acute kidney injury, Nature 572 (2019) 116–119.
[23] S. Saadatnejad, M. Oveisi, M. Hashemi, LSTM-based ECG classification for continuous monitoring on personal wearable devices, IEEE J. Biomed. Heal Inform. 24 (2020) 515–523.
[24] J.R. Ayala Solares, F.E. Diletta Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, et al., Deep learning for electronic health records: a comparative review of multiple deep neural architectures, J. Biomed. Inform. 101 (2020).
[25] A. Zhang, Z.C. Lipton, M. Li, A. Smola, Dive into deep learning, 2020. https://d2l.ai/d2l-en.pdf (accessed 6 Sep 2020).
[26] G. Hinton, T. Tieleman, RMSPROP: Divide the Gradient by a Running Average of its Recent Magnitude. Coursera Neural Networks Mach. Learn. 4 (2012) 26–31. https://www.coursera.org/learn/neural-networks/lecture/YQHki/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude.
[27] P. Remy, keras-attention-mechanism, GitHub repository, 2017.
[28] B. Dzmitry, C. Kyunghyun, B. Yoshua, Neural machine translation by jointly learning to align and translate, in: 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. 2014, pp. 1–15. http://arxiv.org/abs/1409.0473.
[29] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation, IEEE J. Biomed. Heal Inform. (2020).
[30] Y. Zhang, J. Li, Application of heartbeat-attention mechanism for detection of myocardial infarction using 12-lead ECG records, Appl. Sci. 9 (2019).
[31] U.M. Girkar, R. Uchimido, L.H. Lehman, P. Szolovits, L. Celi, W. Weng, Predicting Blood Pressure Response to Fluid Bolus Therapy Using Attention-Based Neural Networks for Clinical Interpretability, arXiv Mach Learn, 2018, pp. 1–6. doi:arXiv:1812.00699v1.
[32] M.A. Jie, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. van Calster, et al., A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. (2019).
[33] G. Lorenzoni, S.S. Sabato, C. Lanera, D. Bottigliengo, C. Minto, H. Ocagli, et al., Comparison of machine learning techniques for prediction of hospitalization in heart failure patients, J. Clin. Med. 8 (2019).
[34] J. Benbenishty, C. Weissman, C.L. Sprung, M. Brodsky-Israeli, Y. Weiss, Characteristics of patients receiving vasopressors, Hear Lung J. Acute Crit. Care 40 (2011) 247–252.
[35] X. Xue-Zhong, W. Hai-Jun, H. Chu-Lin, Y. Quan-Hui, Q. Shi-Ning, Z. Hao, et al., Prognosis of patients with shock receiving vasopressors. World, J. Emerg. Med. 4 (2013).
[36] G. Clermont, C.G. Acker, D.C. Angus, C.A. Sirio, M.R. Pinsky, J.P. Johnson, Renal failure in the ICU: comparison of the impact of acute renal failure and end-stage renal disease on ICU outcomes, Kidney Int. 62 (2002) 986–996.
[37] T. Joint, T. Force, Prevention CVD, Practice C, Society E, European Guidelines on CVD Prevention in Clinical Practice, 2003, December.
[38] R.P. Moreno, P.G.H. Metnitz, E. Almeida, B. Jordan, P. Bauer, R.A. Campos, et al., SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission, Intensive Care Med. 31 (2005) 1345–1355.
[39] A.B. Arrieta, Ser J Del, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, Inf. Fusion 58 (2020) 82–115.
[40] T. Hastie, R. Tibshirani, M. Wainwright, Statistical learning with sparsity: The lasso and generalizations, Stat Learn with Sparsity Lasso Gen, 2015, pp. 1–337.

[41] J. Fan, S. Guo, N. Hao, Variance estimation using refitted cross-validation in ultrahigh dimensional regression, J. R. Stat. Soc. Ser. B Stat. Methodol. 74 (2012) 37–65.
[42] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016).
[43] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Kluger Y. DeepSurv, Personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (2018).
[44] R. Alsaad, Q. Malluhi, I. Janahi, S. Boughorbel, Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma, BMC Med. Inform. Decis. Mak. 19 (2019).
[45] A.M. Carrington, D.G. Manuel, P.W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, et al., Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation, 2021. http://arxiv.org/abs/2103.11357.
[46] D. Rizopoulos, Joint models for longitudinal and time-to-event data: With applications in R, Chapman & Hall/CRC, 2012.
[47] V. Rondeau, Y. Mazroui, J.R. Gonzalez, Frailtypack: An r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation, J. Stat. Softw. 47 (2012).
[48] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI, Inf. Fusion 71 (2021) 28–37.