An Interpretable Deep Learning Model for Time-Series Electronic Health Records: Case Study of Delirium Prediction in Critical Care

Seyedmostafa Sheikhalishahi[1,2], **Anirban Bhattacharyya**[3], **Leo Anthony Celi**[4,5,6], **and Venet Osmani**[1,7]

[1]Fondazione Bruno Kessler Research Institute, Trento, Italy
[2]University of Trento, Italy
[3]Critical Care Services, Mayo Clinic, Jacksonville, FL, USA
[4]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[5]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA
[7]Information School, University of Sheffield, UK

ABSTRACT

Deep Learning (DL) models have received increasing attention in the clinical setting, particularly in intensive care units (ICU). In this context, the interpretability of the outcomes estimated by the DL models is an essential step towards increasing adoption of DL models in clinical practice. To address this challenge, we propose an ante-hoc, interpretable neural network model. Our proposed model, named double self-attention architecture (DSA), uses two attention-based mechanisms, including self-attention and effective attention. It can capture the importance of input variables in general, as well as changes in importance along the time dimension for the outcome of interest. We evaluated our model using two real-world clinical datasets covering 22840 patients in predicting onset of delirium 12 hours and 48 hours in advance. Additionally, we compare the descriptive performance of our model with three post-hoc interpretable algorithms as well as with the opinion of clinicians based on the published literature and clinical experience. We find that our model covers the majority of the top-10 variables ranked by the other three post-hoc interpretable algorithms as well as the clinical opinion, with the advantage of taking into account both, the dependencies among variables as well as dependencies between varying time-steps. Finally, our results show that our model can improve descriptive performance without sacrificing predictive performance.

Introduction

Deep learning (DL) methods and specifically recurrent neural networks (RNNs) are revolutionising many scientific fields such as natural language processing[1], machine translation[2], and as well clinical domain[3]. In this regard, the use of DL models has demonstrated an upward trend in the clinical field for the past several years[4]. These models can capture non-linear relationships in clinical data and significantly outperform the conventional machine learning (ML) models. However, DL models show a limited degree of interpretability and to a large degree are considered black-boxes[5]. Therefore, we need to probe these models better to extract a degree of interpretability from them to make these models more reliable for clinicians.

Conventional machine learning models have been used in the intensive care unit (ICU), which are interpretable[6] but cannot capture non-linear relationship in data. This is because the data in ICU is recorded in a time-series and conventional ML models do not have an intrinsic ability to deal with time-series inputs. More advanced models can deal with time-series data, for example RNNs can model evolution of patients' state, however they are not intrinsically interpretable.

The interpretability of DL models remains a significant challenge in the ML domain. In this context, interpretability and explainability concepts are often used interchangeably within the general Artificial Intelligence (AI) community[7]. Interpretable models are categorised into post-hoc and ante-hoc models. Post-hoc models incorporate the interpretable module only at inference and as such, they aim to keep a trained model unchanged, while explaining their behaviour externally. Examples of post-hoc methods include Shapley Value Sampling (SVS)[8] belonging to the occlusion-based family of interpretability methods, Integrated Gradients (IG)[9], and Guided Back-propagation (GB)[10].

In contrast, ante-hoc models incorporate the interpretable module during training. As a consequence, a single model is employed for both prediction and interpretation. Attention-based models, such as[11] belong to ante-hoc interpretable models. Self-attention is an attention mechanism that relates different positions of a single sequence in order to compute a representation of the sequence. The self-attention mechanism has been employed successfully in a variety of tasks, including machine

translation[12], abstractive summarisation[13], and textual entailment[14].

Recently developed attention models offer the promise of providing interpretability while retaining the flexibility and versatility of DL models. The attention-based models were employed to predict outpatient disease progression[11]. The Attend and Diagnose model of Song et al.[15] used a self-attention mechanism to improve an RNN's predictive accuracy for four clinical tasks but did not explore interpretability. While important time points were easily extracted from this model, identifying important variables at a given point in time required additional calculation which is not considered in the proposed method. Choi et al.[16] proposed RETAIN model, which uses two separate RNN layers integrated with an attention layer over both variables and time using embedded variables. In contrast to our model, Choi et al. do not consider the dependencies between time-steps and dependencies between different variables and use two separate RNN networks, which could be computationally expensive; their model is trained and validated on EHR data to predict heart failure. The attention-based model of Kaji et al.[11], which is applied to three clinical tasks, focused on variable-level interpretability. However, it does not consider the time-level importance, and dependencies among variables and times-steps are not considered.

Time-level importance is useful to understand how importance of variables changes over time, for example during an ICU stay, such that clinical interventions can dynamically target the corresponding clinical aspects as they arise, described by those variables. Additionally, knowing which clinical parameters are important and at what time allows physicians to design personalised interventions, potentially leading to improved outcomes.

Similar to[11], a possible way of interpreting the structured data is to employ an attention-layer straight after the input layer, which computes the coefficient of each variable before being fed into RNN. In the study done by Zhang et al.[17] an LSTM-based model with event embedding and time encoding is leveraged to model clinical time series for early prediction of sepsis in the emergency department. Additionally, an attention mechanism and global max pooling techniques are employed to enable interpretation for the LSTM-based model. Unlike[17] that converted numerical values into categorical values and created an embedding out of them, in our study, we used actual numerical values and converted categorical variables into embeddings. Additionally, we employed double self-attention architecture to provide a clinically validated interpretability.

However, the above-mentioned ante-hoc interpretable models have three limitations as follows:

1. The dependencies among clinical variables and time-steps are not captured.
2. Time-step importance is not considered as the attention is applied on variable-level.
3. The predictive performance is generally worsened.

To address the limitations as mentioned earlier, we propose a Double Self-attention Architecture (DSA), which employs a self-attention[18] mechanism at a variable-level and another self-attention mechanism at the time-step level. Additionally, we use effective-attention mechanism to interpret the model outcomes as it was found to be more performant than self-attention[19]. Effective-attention is computed from a matrix decomposition of self-attention mechanism which is explained comprehensively in the explanation module section.

For brevity and concerning the use of self-attention in both algorithms, we term both double self-attention and double effective-attention architectures as *DSA* in the rest of this article. In summary, the contributions of this work are as follows:

- DSA simultaneously attends over the variable level and the time-step level.
- DSA takes into account the dependencies between different time-steps and as well as correlation among clinical variables while computing the importance of each variable and time-step.
- DSA outperforms the ante-hoc interpretable models, while providing clinically validated interpretability.
- Comparison with clinical knowledge and other post-hoc interpretable models verifies the soundness of variable ranking provided by DSA, avoiding spurious associations that are not clinically relevant.
- DSA maintains comparable predictive performance with baseline models, such as BiLSTM while providing interpretability, as such minimising the trade off between predictive performance and interpretability.

We developed and validated an interpretable DL model to provide a variable ranking based on prediction of the onset of delirium in critically ill-patients to prioritise the patients at risk. This is a clinically important case study because delirium occurrence is common in the ICU. At the same time its aetiology is not well understood, while the preventive strategies, such as ABCDEF bundle, are highly resource intensive[20]. Our model allows for (1) an interpretable DL model, (2) variable ranking by considering varying aspects such as variable inter-dependence and time-step dependencies, and (3) an interpretable screening tool that can prioritise patients at risk, thus reducing the burden on care providers.

Materials and methods

Ethics statement

The current dataset was constructed by processing the eICU Collaborative Research Database (eICU-CRD)[21] and Medical Information Mart for Intensive Care (MIMIC-III)[22] critical care dataset. As the study was based on publicly available datasets, there was no need for further Institutional review board (IRB) approval for this research.

Data description, cohort selection and outcome definition

The eICU-CRD is a freely available multi-centre database comprising 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 in over 200 hospitals located throughout the US[21]. The MIMIC-III database is an open-access single-centre ICU database including 53,423 distinct hospital admissions for 46,476 unique patients admitted from 2001 to 2012[22]. Both datasets comprise patient demographics, vitals, clinical flowsheets, laboratory values, medications, interventions, and outcomes. Any patient admitted to the ICU for 24 hours or more and with at least one CAM (Confusion Assessment Method) was included in our study population. In the patient records, in the case of multiple positive CAM-ICU records, the first CAM-ICU was considered as the onset of delirium. The patients older than 18 and younger than 89 are included in the study, resulting in 22840 patients (16546 patients from eICU-CRD and 6294 patients from MIMIC-III). The patients characteristics for both datasets are shown in supplementary material, Table 5

Variable selection and preprocessing

We compiled 21 clinical variables identified by critical care clinicians as relevant to delirium prediction, commonly used in the literature, and available in both data-sets, including demographics, vital signs, laboratory measurements, and medication data. A detailed list of the included clinical variables in this study is depicted in Supplementary material in Table 6. Variable preprocessing is detailed in[23], and included aggregating values into hourly intervals with the last known value used for that interval. If a variable is not measured during the interval, the value is imputed by forward and then (if required) backward imputation. We converted categorical variables into a vector to capture the semantics of each category, while for continuous variables we used the recorded value in the database without any adaptation.

Outcome assessment

In this study, we evaluated the ability of the proposed model to provide a clinically validated variable ranking in the case of delirium prediction in different settings, such as varying observation window (12h and 24h) and different prediction window (12h and 48h) illustrated in Figure 1.

In this work, the observation windows of 12h or 24h are chosen based on the Intensive Care Delirium Screening Checklist (ICDSC)[24] which is an 8-24 hours window[25] to predict the incidence of delirium in the following 12h or 48h. Therefore, our methods estimate patients' risk of delirium in the next 12 to 48 hours, based on multivariate analysis of a sequence of clinical parameters collected during the observation window of either 12 hours or 24 hours[26].

Model development

Each patient can be viewed as a sequence of medical records (vital signs, laboratory measurements, and demographics) ordered by time, and each record contains a set of clinical variables. A three dimensional data with patient ICU stays (n = 15,726), time steps (n = 24 or 12), and variables (n = 21) serves as input to the model. As shown in Figure 2, the proposed model is divided into three modules, namely input preparation, explainable module, and prediction module.

Input representation

We process and model numerical and categorical variables separately. Categorical variables are represented using either one-hot encoding or entity embedding. One-hot encoding is the baseline approach that converts the variables into binary representation. Since this approach results in a large sparse matrix, we have compared the performance of one-hot encoding with entity embedding in our previous work[27] and found entity embedding to provide superior performance. Therefore, in this work we use entity embedding[28], where each categorical variable in the dataset is mapped to a vector and the corresponding embedding is added to the patient's record. This entity embedding is learned by the neural network during the training phase along with other parameters. The model in the training phase learns the vectors related to each categorical variable. The vectors of the categorical variable are concatenated with the numerical variables to be fed into the model. Therefore, the input representation at time t is as follows:

$$x_t = Concat[(Numerical_t, U(Categorical_t)] \quad (1)$$

$Numerical_t$ stands for the numerical variable, $Categorical_t$ stands for the categorical variable at time t, and U is the embedding matrix.

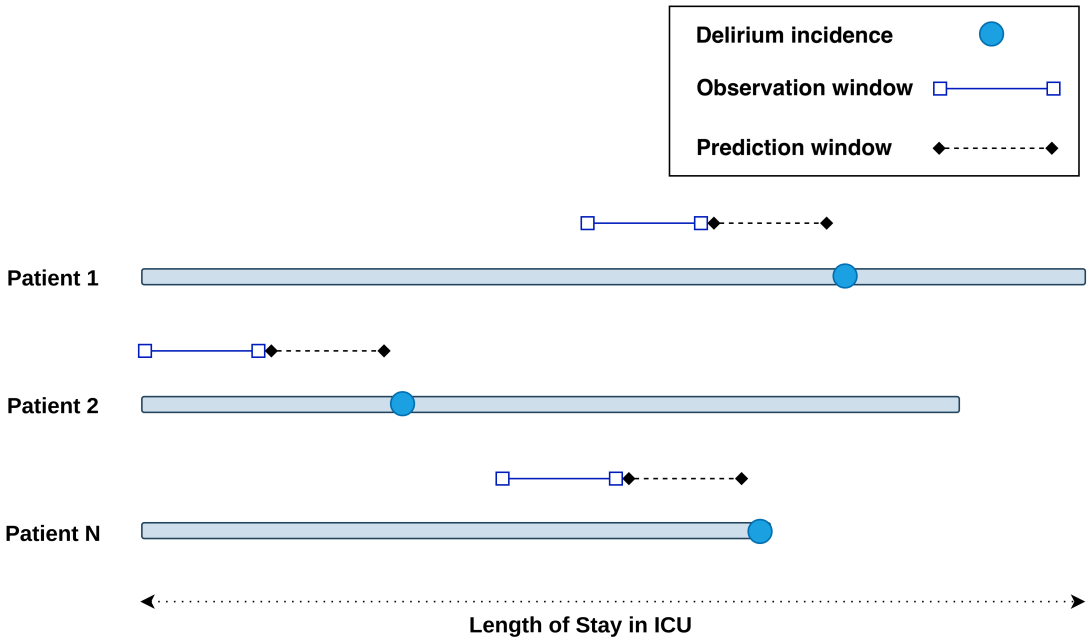

Figure 1. Delirium prediction schema; observation window represents the collected data for each study (12h, 24h), and the prediction window represents the time ahead to predict delirium (12h, 48h). The blue bar represents the length of stay in ICU that can vary between patients

In the explanation module, the input x_t is fed into two different self-attention layers, as shown in Figure 2. The self-attention mechanism on the right side is applied on the variables to compute the variable importance, namely α_v. The self-attention mechanism on the left side is applied on the time-steps to compute the time-step importance named as α_t. The coefficient of the contribution is computed via both α_v and α_t using a dot-product applied as follows:

$$c(x_t) = \underbrace{\alpha_v}_{\text{self-attention on variable}} \odot \underbrace{\alpha_t}_{\text{self-attention on time}} \tag{2}$$

further expanded in[29].

The input data is weighted with the computed attention using a residual connection as shown in the equation 3

$$wi_t = x_t \odot \underbrace{c(x_t)}_{\text{contribution coefficient}} \tag{3}$$

where wi_t is the weighted input at time t, x_t is the input which was computed in equation 1, and c is the computed coefficient of the contribution in equation 2.

In the prediction module, similar to[11] the weighted input is fed to a masking layer to filter the time steps where patients have less than 12 or 24 hours of data available and are fed into a BiLSTM layer to get the data representation for each patient.

Formal definition of BiLSTM is available in[30]. In our work BiLSTM layer with 128 units is connected to a hyperbolic tangent activation function. The output layer of our network consists of one dense neuron with a softmax activation to output the probability of a given event over ICU stays.

Model training and evaluation

We evaluate the models using descriptive and predictive performance metrics. We compared predictive performance of our models with both, BiLSTM as well as the model proposed by Kaji et al.[11], using the same architecture for both, however with manual hyper-parameter optimisation to achieve model convergence. We used Adam optimiser with a learning rate of $7.5 * 10^{-4}$, and decay of $1 * 10^{-6}$ in all models, with batch size of 128, training the models for 50 epochs using cross-entropy as the loss function. We evaluated the results based on 5-fold stratified cross-validation. Typically, metrics computed based on the

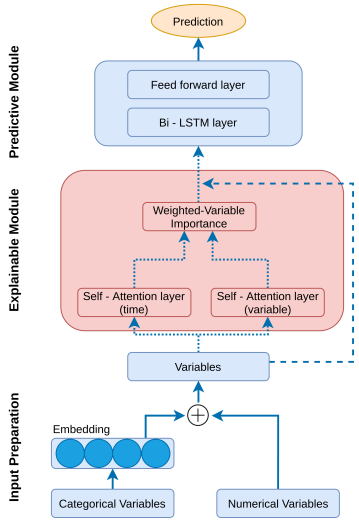

Figure 2. Proposed architecture

k-fold stratified cross-validation can assess overfitting and have lower variance[31]. We report the predictive performance using the Area Under Receiver Operating Characteristic (AUROC), Area Under Precision-Recall Curve (AUPRC), Precision and Recall with Confidence Interval (CI) of 95%.

Explanation module

Understanding how the model predicts a patient's delirium onset is an essential step in validating its use. DL techniques are typically considered black boxes where it is challenging to determine how a predictive model generates a prediction. A model should provide clinically validated explanations related to the clinical variables where these explanations can be utilised by clinicians during daily routines. Recent advances in ML techniques, such as attention mechanisms have enabled an improved way to probe interpretability. Attention-based[11] models give importance to the classification associated with each input variable given to the model, allowing us to identify the most predictive variables that contribute to the severity of the diagnosis. In this section, we employ two attention-based models to understand what has been learned by our models. In detail, we can observe which time-steps and variables the model relies on assigning a degree of significance to the time-steps and variables.

As it is shown in Figure 3.a, the data of each patient is fed as input into attention-based (self-attention or effective-attention) layers. We employ two 3-head attention mechanisms in order to compute time-step importance and variable importance. As it is shown in Figure 3.b, to compute the importance of a single time-step, we need to score each time-step of the input sequence against this single time-step. The score is computed as it is shown in equation 5 and determines focus that needs to be placed on other time-steps as we encode a time-step at a specific position. In this way, we can capture the dependencies between time steps while computing the importance of each time step. As depicted in Figure 3.c, to compute the importance of a single variable, we need to score each variable of the input sequence against this single variable, and the score determines focus that needs to be placed on other variables as we encode a variable. While computing the importance of each variable, the self-attention enables capturing dependencies among variables, including the temporal dimension.

In the following sections, we provide a detailed description of the varying version of attention mechanisms which provide interpretable outputs, namely self-attention and effective-attention, where the latter is computed from matrix decomposition of self-attention, the component orthogonal to the nullspace contributing to the model output. Effective-attention has been shown to be less associated with less important variables, consequently capturing better the most relevant variables in comparison to self-attention[19].

Self-attention

The implications of time steps and clinical variables vary depending on the context. To capture this contextual information, we applied two self-attention layers to which one self-attention layer attends over time-steps and the other self-attention layer attends over variables.

As it is demonstrated in Figure 4, Self-attention is an attention mechanism that relates different positions of a single sequence in order to compute a representation of the sequence. The self-attention mechanism has been employed successfully

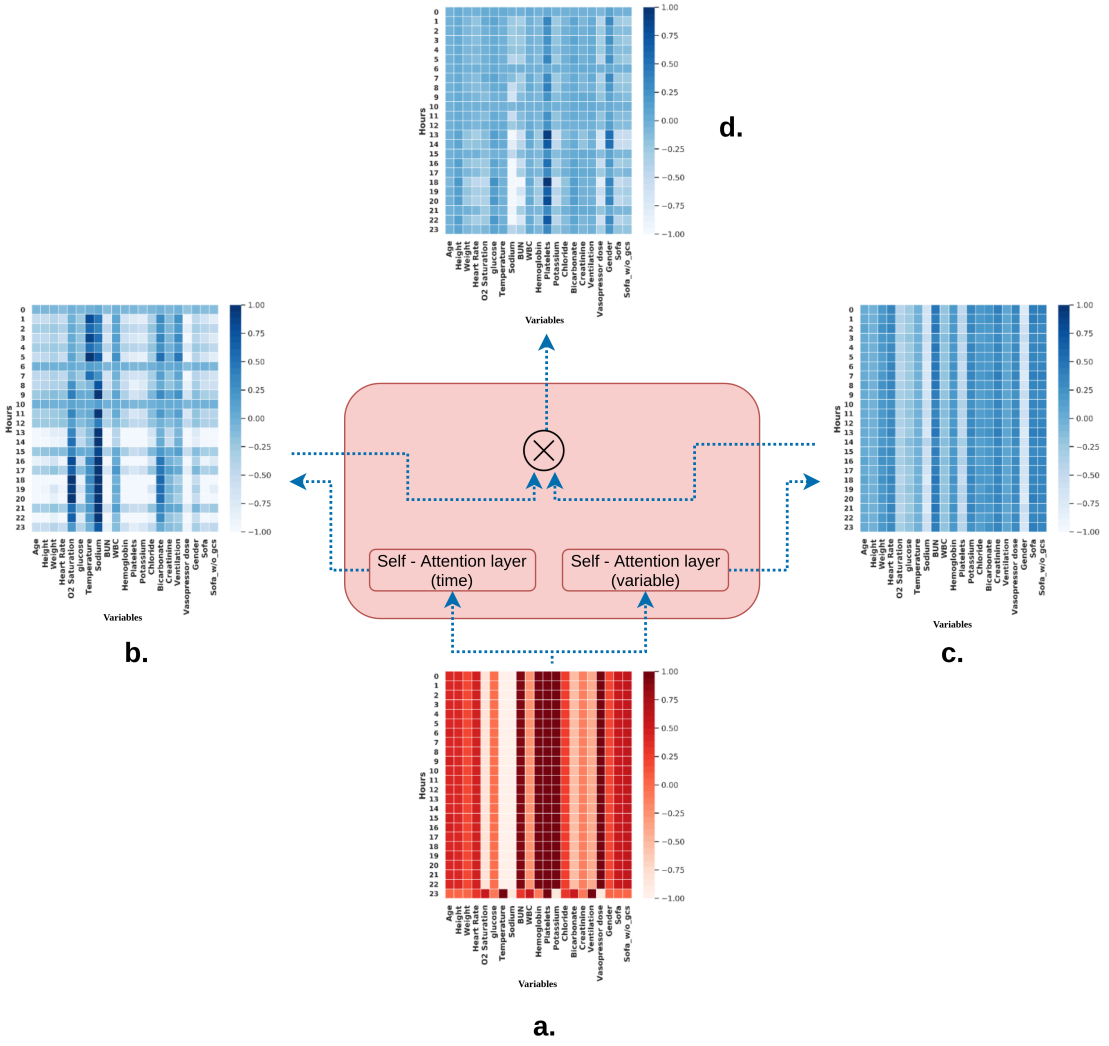

Figure 3. Explanation module: a.Input data; b.Time importance; c.Variable importance; d.Variable importance by considering time importance

in a variety of tasks, including machine translation[12], abstractive summarisation[13], and textual entailment[14]. Formally,

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

Where Q, K, V are computed by multiplying input with the learned matrices W_Q, W_k, W_V during training.
DSA employs multi-head self-attention, which projects queries, keys, and values h times with different, learned linear projections. The scores are computed in parallel and are concatenated to get one matrix score, Formally:

$$Multihead(Q,K,V) = Concat(head_1, head_2)W^O where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

Where parameters matrices such as W_i^Q, W_i^K, W_i^V, and W^O are the projections[18].

Effective-attention

As it is demonstrated in[32], self-attention can be decomposed into two matrices: i) the component in the left nullspace of V which is indicated with ($A^{\|}$) and ii) the component orthogonal to the nullspace ($A^{\perp}$). The matrix $A^{\|}$ is irrelevant for the model

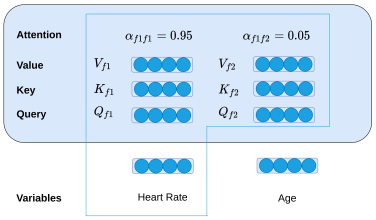

Figure 4. The architecture of self-attention

output because its product with the value matrix is equal to zero. The matrix $A^{\perp}$ contributes to the model output, which is so-called *effective-attention*.

Additionally, as Sun et al.[19] noted effective-attention is associated less with the variables related to the language modelling pre-training such as separator [SEP], and it has the potential to illustrate linguistic variables much better than self-attention. Equivalent to our study, we believe that in comparison to self-attention, effective-attention is associated less with less important clinical variables, allowing it to capture better the most relevant clinical variables in outcome prediction.

$$AV = (A^{\parallel} + A^{\perp})V = \vec{0} + A^{\perp}V = A^{\perp}V \tag{6}$$

As it is illustrated in equation 6, $A^{\parallel}V$ is equal to zero, therefore the effective-attention matrix is equal to $A^{\perp}$.
The effective-attention matrix $A^{\perp}$, is computed as the following[32]:

- We first compute the singular value decomposition (SVD) of the value matrix V which is $V = U\sum WT$
- The rows of U that correspond to singular values equal to zero span LN(V):
  $LN(V) = span\{u_1, ..., u_k\}$,
  Where k is the number of singular values that equal zero.
- We project each row a_i of the attention matrix A to LN(V) to construct a projection of the matrix A to LN(V):
  $P_{LN(V)}(a_i) = \sum_{j=1}^{k} \langle a_i, u_j \rangle u_j, \forall i \in \{1, ..., d_s\}$,
  $P_{LN(V)}(A) = [P_{LN(V)}(a_1), ..., P_{LN(V)}(a_{ds})]$
  where $\langle ., . \rangle$ denotes the dot product.
- effective-attention is equal to
  $A^{\perp} := A - P_{LN(V)}(A)$

It is worth mentioning that similar to the approach in[19] we replace self-attention with effective-attention at the model test phase.

Data and code availability

We made use of several open-source libraries based on Python to conduct our experiments, ML framework Scikit-learn[33] and DL framework Pytorch[34]. The experiments and implementation details are available in:
https://github.com/mostafaalishahi/Delirium_prediction_models.

Results

In this section we report both, the descriptive evaluation of the model as well as predictive performance. The descriptive performance is evaluated against the well known algorithms, including Shapley Value Sampling, Integrated Gradients and Guided Back-propagation, while the predictive performance is based on evaluation metrics such as AUROC, AUPRC, precision and recall with 95% CI.

Descriptive performance

As mentioned earlier the importance of interpretable deep learning models in the clinical domain, in this section we explore further interpretability by providing the most important clinical variables for delirium-onset prediction task. In this regard, we compute variable importance by considering the importance of one variable over other variables across the patient cohort as shown in Equation 2. Although there are many definitions of interpretability, we focused on how the model ranks each input variable with respect to outcome prediction. Given that interpretability of neural networks is still an open research question,

especially for temporal neural networks[35], we also provide results from three other post-hoc models to compare with our proposed model. In this context, we employed as the benchmark the Shapley Value Sampling (SVS)[8], Integrated Gradient (IG)[9], and Guided Backpropagation (GB)[10], to ensure that the variable importance results computed by DSA are consistent across the three benchmark models. The top-10 influential variables ranked for MIMIC-III and eICU-CRD are reported in Table 1 and Table 2 respectively. The variable ranking is reported using three different post-hoc interpretable algorithms, namely IG, SVS, and GB, compared to two proposed ante-hoc attention-based interpretable models, namely self-attention and effective-attention.

The most influential variables that have contributed to delirium prediction according to their relative importance in the eICU-CRD dataset as reported in Table 2 are heart rate, Ventilation, age, white blood cell count, and vasopressor dose according to the five algorithms. Most of these variables are also ranked in the top-10 in the MIMIC-III dataset as depicted in Table 1. Both proposed attention-based interpretable models (DSA) captured most of the important variables ranked by IG, SVS, and GB and in both datasets, validating the soundness of the proposed model, with the additional advantage of also providing inter-variable dependencies as well as temporal importance.

It is interesting to note that, effective-attention which was previously used in[19] shows a slightly higher number of variables in common with the other three post-hoc algorithms that is an extra point for effective-attention to be studied further in the case of clinical time-series data.

	Observation window 12h – Prediction window 48h				
Variable ranking \ **Algorithm**	**IG**	**SVS**	**GB**	**DSA (self-attention)**	**DSA (effective-attention)**
1	ventilation	ventilation	ventilation	weight	***heart rate***
2	WBC	gender	WBC	*hemoglobin*	***platelets***
3	creatinine	WBC	creatinine	*heart rate*	***BUN***
4	vasopressor dose	vasopressor dose	vasopressor dose	*sodium*	*hemoglobin*
5	sodium	sofa	sodium	*BUN*	*SpO_2*
6	BUN	sodium	BUN	*WBC*	potassium
7	glucose	age	glucose	potassium	*WBC*
8	platelets	sofa w/o GCS	platelets	*vasopressor dose*	height
9	hemoglobin	creatinine	hemoglobin	*glucose*	***sofa w/o GCS***
10	heart rate	BUN	heart rate	bicarbonate	***creatinine***
	Observation window 24h – Prediction window 12h				
1	ventilation	ventilation	ventilation	temperature	*WBC*
2	gender	gender	heart rate	*ventilation*	weight
3	sodium	heart rate	sodium	*sodium*	***age***
4	heart rate	sodium	creatinine	*SpO_2*	***ventilation***
5	sofa	sofa	bicarbonate	*platelets*	*bicarbonate*
6	age	age	age	*BUN*	*SpO_2*
7	bicarbonate	sofa w/o GCS	BUN	*age*	height
8	vasopressor dose	vasopressor dose	vasopressor dose	chloride	*gender*
9	sofa w/o GCS	bicarbonate	platelets	weight	potassium
10	creatinine	creatinine	WBC	*vasopressor dose*	***vasopressor dose***

Table 1. Variable ranking presented by different algorithms versus DSA (top-10 variables) on MIMIC-III dataset. The variables ranked using self-attention and effective-attention in common with IG, SVS, and GB are represented in *italic*. Variables that are considered relevant or partially relevant based on the published literature and clinical experience (shown in Table 7) are highlighted in **bold**.

Variable ranking \ Algorithm	IG	SVS	GB	DSA (self-attention)	DSA (effective-attention)
	Observation window 12h – Prediction window 48h				
1	ventilation	ventilation	ventilation	*heart rate*	**sodium**
2	heart rate	heart rate	heart rate	sodium	***heart rate***
3	age	age	WBC	platelets	hemoglobin
4	WBC	WBC	age	height	*WBC*
5	sofa	sofa	vasopressor dose	*age*	**sofa w/o GCS**
6	vasopressor dose	vasopressor dose	bicarbonate	*chloride*	***ventilation***
7	bicarbonate	bicarbonate	chloride	*weight*	***glucose***
8	BUN	BUN	BUN	*WBC*	***age***
9	chloride	chloride	glucose	sofa w/o GCS	*weight*
10	weight	weight	weight	*SpO_2*	height
	Observation window 24h – Prediction window 12h				
1	ventilation	ventilation	vasopressor dose	*potassium*	***age***
2	vasopressor dose	vasopressor dose	ventilation	temperature	*WBC*
3	age	age	WBC	creatinine	***vasopressor dose***
4	heart rate	heart rate	heart rate	*vasopressor dose*	*potassium*
5	WBC	WBC	age	*SpO_2*	*weight*
6	potassium	potassium	potassium	*weight*	**sodium**
7	sofa	sofa	platelets	*heart rate*	hemoglobin
8	bicarbonate	bicarbonate	bicarbonate	*age*	**creatinine**
9	gender	weight	weight	*platelets*	*gender*
10	weight	platelets	BUN	*BUN*	***ventilation***

Table 2. Variable ranking presented by varying algorithms vs. proposed model (top-10 variables) on eICU-CRD dataset. The variable ranked using self-attention and effective-attention in common with IG, SVS, and GB are represented in *italic*. Variables that are considered relevant or partially relevant based on the published literature and clinical experience (shown in Table 7) are highlighted in **bold**.

Predictive performance

We evaluated 12014 (24h observation – 12h prediction) and 9481 (12h observation – 48h prediction) from eICU-CRD and 3712 (24h observation – 12h prediction) and 2128 patients (12h observation – 48h prediction) from MIMIC-III databases. Considering AUPRC, Precision, and Recall, DSA outperforms the proposed model by Kaji et al. in two different scenarios as shown in Table 3 for MIMIC-III and for eICU-CRD datasets as depicted in Table 4 . Another point to mention, although the predictive performance of the DSA is better than Kaji's model, its predictive performance is slightly worse than BiLSTM in terms of precision. This is due to the nature of ante-hoc interpretable models in which there is a trade-off between predictive performance power and descriptive performance[11].

Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
	Observation window 12h – Prediction window 48h			
BiLSTM	71.37 (67.99 - 74.72)	29.81 (27.33 - 31.88)	28.45 (25.74 - 31.17)	65.98 (59.82 - 72.13)
DSA	68.66 (64.99 - 72.33)	28.58 (23.64 - 33.20)	26.70 (22.56 - 30.85)	59.85 (51.93 - 67.77)
Kaji model	67.56 (64.91 - 70.22)	27.90 (25.68 - 30.46)	24.31 (22.01 - 26.60)	58.07 (52.48 - 63.66)
	Observation window 24h – Prediction window 12h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	81.24 (76.44 - 86.11)	44.45 (40.35 - 48.81)	35.03 (30.73 - 39.33)	71.16 (64.54 - 77.77)
DSA	80.50 (77.23 - 83.85)	44.90 (41.28 - 49.40)	35.58 (33.71 - 37.44)	68.98 (62.69 - 75.27)
Kaji model	78.33 (76.11 - 80.62)	41.69 (38.55 - 45.15)	32.39 (28.52 - 36.25)	65.63 (57.78 - 73.48)

Table 3. Predictive performance on MIMIC-III dataset

	Observation window 12h – Prediction window 48h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	84.20 (82.52 - 85.86)	33.24 (29.41 - 36.54)	**28.37 (27.15 - 29.58)**	74.44 (70.91 - 77.98)
DSA	82.51 (80.33 - 84.67)	31.21 (28.01 - 33.70)	24.92 (24.22 - 25.61)	75.99 (72.20 - 79.78)
Kaji model	81.64 (80.05 - 83.27)	30.19 (27.41 - 32.27)	24.60 (23.23 - 25.97)	75.00 (70.56 - 79.44)
	Observation window 24h – Prediction window 12h			
Model	AUROC% (95% CI)	AUPRC% (95% CI)	Precision% (95% CI)	Recall% (95% CI)
BiLSTM	88.02 (86.31 - 89.75)	42.69 (38.71 - 46.1)	**38.19 (36.78 - 39.6)**	80.39 (76.62 - 84.16)
DSA	87.10 (85.15 - 89.03)	42.20 (38.78 - 45.69)	34.17 (32.88 - 35.46)	82.89 (78.54 - 87.25)
Kaji model	85.85 (84.16 - 87.57)	38.03 (34.64 - 41.02)	35.82 (34.48 - 37.15)	76.63 (72.84 - 80.42)

Table 4. Predictive performance on eICU-CRD dataset. Highlighted in bold are results with statistically significant differences

Discussion

Our study shows that the proposed models outperform the state-of-the-art interpretable model proposed by[11], while being interpretable and comparable to a handful number of post-hoc interpretable algorithms such as IG, SVS, and GB. This demonstrates the strength of DSA in both the predictive performance and the associated interpretable matrix.

Improving the prediction of delirium is a critical step towards improving ICU outcomes, and optimising costs[36]. Our study found that incorporating two self-attention layers with a BiLSTM layer can achieve informative performance in predicting delirium onset. The AUROC, AUPRC and Recall for the delirium prediction suggest that employing self-attention does not significantly sacrifice predictive performance with respect to the BiLSTM. Furthermore, the focus of this work is principally on variable ranking and explainability, where between two methods that perform comparably, the one that provides better explanation of its predictions would be preferable, even at the cost of slightly sacrificing the predictive performance.

In this study, we demonstrated how to achieve a level of interpretability for a DL model for clinical events in ICU by incorporating self-attention mechanisms. As noted[37], interpretability of DL models can facilitate understanding of inferential processes of a neural network and improve the model in terms of descriptive.

While many BiLSTM based models to predict clinical outcomes have incorporated attention, we are aware of only a handful of them that used attention to identify the variables driving the prediction[11,15–17,38–40]. Several of the studies as mentioned earlier employed attention; however, none of them compared the proposed model with other interpretable models.

We demonstrated how self-attention could be applied to the input variables to provide a degree of interpretability by capturing variable dependencies and time-step dependencies. It should be noted that clinical understanding of delirium is currently incomplete. Therefore, while the majority of top variables considered relevant by our model are also clinically relevant, the relevance of some of the remaining variables cannot be excluded in future studies.

Our study has the following limitations: i) many variables that clinicians would have wanted to incorporate into this study were not available in eICU-CRD and MIMIC-III datasets or had a very high rate of missing data, in part due to the heterogeneity of the datasets, ii) we note that the proposed self-attention model can underline the importance of each variable but cannot identify how a variable affects the probability of an event, delirium prediction in our case, without performing further analysis and, iii) this study included data from US-based hospitals and ICUs only, therefore generalisability in a different context would require additional analysis.

Conclusion

In conclusion, we believe that self-attention mechanisms could create interpretable decision support systems for clinical practice. This study demonstrated that such an approach could learn informative models for predicting delirium and has shown how the individual variables underlying these predictions can be explored using self-attention and effective-attention mechanisms. Furthermore, the explainability module proposed in this paper can be used effectively to visualise variable importance. This in turn would aid understanding of the input variables and support clinical decision making by focusing on particular variables that the model has deemed important at time points of interest of the disease trajectory.

Supplementary material

Variables	eICU			MIMIC		
Number of patients	CAM-ICU + 3153	CAM-ICU - 13393	*p* value —	CAM-ICU + 1268	CAM-ICU - 5026	*p* value —
Age, mean (SD), years	65.53 (15.14)	62.20 (16.16)	< 0.05	64.81 (15.62)	63.27 (15.82)	< 0.05
Female (%)	1405 (44)	6295 (47)	—	545 (43)	2211 (44)	—
Height, mean (SD), m	168.47 (18.23)	169.25 (15.90)	< 0.05	170.06 (14.22)	168.88 (14.87)	0.054
Weight, mean (SD), kg	83.06 (29.88)	85.00 (25.58)	< 0.05	82.68 (30.25)	81.53 (24.89)	0.15
Heart Rate, mean (SD), bpm	88.22 (18.06)	85.09 (17.73)	< 0.05	88.60 (17.53)	85.12 (17.29)	< 0.05
Oxygen Saturation, mean (SD), %	97.16 (2.72)	96.80 (2.79)	< 0.05	97.17 (2.71)	96.58 (4.50)	< 0.05
Glucose, mean (SD), mg/dL	140.32 (45.97)	146.46 (56.31)	< 0.05	144.51 (58.70)	141.25 (51.43)	< 0.05
Temperature, mean (SD), °C	37.01 (0.69)	36.97 (2.65)	< 0.05	37.06 (0.76)	36.88 (0.76)	< 0.05
Serum Sodium, mean (SD), mEq/L	140.32 (5.80)	138.57 (5.04)	< 0.05	139.39 (5.48)	138.32 (4.89)	< 0.05
BUN, mean (SD), mg/dL	31.93 (22.10)	25.88 (18.64)	< 0.05	33.96 (24.46)	28.10 (20.77)	< 0.05
WBC, mean (SD), per microliter	13.01 (6.47)	11.08 (5.51)	< 0.05	12.13 (7.73)	10.74 (6.29)	< 0.05
Hemoglobin, mean (SD), g/dL	9.73 (1.89)	10.00 (2.08)	< 0.05	9.76 (1.68)	10.27 (1.76)	< 0.05
Platelets, mean (SD), per microliter	201.34 (122.76)	210.23 (108.70)	< 0.05	202.59 (137.23)	199.53 (114.33)	< 0.05
Serum Potassium, mean (SD), mEq/L	3.98 (0.59)	4.00 (0.57)	0.1431	4.03 (0.57)	4.07 (0.56)	< 0.05
Chloride, mean (SD), mEq/L	105.54 (6.86)	103.24 (6.29)	< 0.05	104.57 (6.69)	104.36 (6.37)	< 0.05
Serum Bicarbonate, mean (SD), mEq/L	35.23 (5.02)	25.52 (5.02)	< 0.05	25.16 (5.21)	24.88 (4.95)	< 0.05
Serum creatinine, mean (SD), mg/dL	1.45 (1.16)	1.37 (1.21)	< 0.05	1.63 (1.28)	1.37 (1.05)	< 0.05
Ventilation, mean (SD)	0.87 (0.34)	0.71 (0.45)	< 0.05	0.56 (0.50)	0.33 (0.47)	< 0.05
Total norepinephrine dose (SD), mcg/kg/min	0.02 (0.31)	0.01 (0.28)	< 0.05	0.08 (0.63)	0.06 (0.57)	< 0.05
SOFA, mean (SD)	4.9 (3.3)	3.42 (2.84)	< 0.05	6.46 (3.77)	6.67 (3.34)	< 0.05
SOFA without GCS, mean (SD)	3.27 (2.83)	2.58 (2.33)	< 0.05	5.42 (3.65)	4.99 (3.13)	< 0.05

Table 5. *Characteristics of the selected patient cohorts divided by the CAM-ICU status*

Variable group	Variable name
Demographic data	age, gender, height, weight
Vital signs	oxygen saturation (SpO_2), heart rate (HR), temperature
Other Measurements	sofa, sofa without GCS, Ventilation
Laboratory Measurements	white blood cell count (WBC), sodium (Na), blood urea nitrogen (BUN), glucose, hemoglobin, platelets, potassium, chloride, bicarbonate, creatinine
Administered drugs	Dopamine, epinephrine, norepinephrine, phenylephrine (all calculated as norepinephrine equivalent)

Table 6. Full set of variables included in the prediction models

Variables	Relevance
age BUN glucose norepinephrine sodium sofa without GCS SpO2 ventilation	Considered relevant to delirium prediction
creatinine heart rate platelets vasopressor dose	Considered partially relevant to delirium prediction
bicarbonate gender height hemoglobin potassium WBC weight	Not currently considered relevant to delirium prediction

Table 7. Relevance of variables based on the published literature as well as clinical experience

References

1. Jagannatha, A. N. & Yu, H. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, vol. 2016, 856 (NIH Public Access, 2016).
2. Singh, S. P. *et al.* Machine translation using deep learning: An overview. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, 162–167 (2017).
3. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal biomedical health informatics* **22**, 1589–1604 (2017).
4. Sheikhalishahi, S. *et al.* Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics* **7**, e12239 (2019).
5. Rai, A. Explainable ai: from black box to glass box. *J. Acad. Mark. Sci.* **48**, 137–141 (2020).
6. Ahmad, M. A., Eckert, C. & Teredesai, A. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 559–560 (2018).
7. Silva, W., Fernandes, K., Cardoso, M. J. & Cardoso, J. S. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, 133–140 (Springer, 2018).
8. Castro, J., Gómez, D. & Tejada, J. Polynomial calculation of the shapley value based on sampling. *Comput. & Oper. Res.* **36**, 1726–1730 (2009).
9. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328 (PMLR, 2017).
10. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).
11. Kaji, D. A. *et al.* An attention based deep learning model of clinical events in the intensive care unit. *PloS one* **14**, e0211057 (2019).
12. Song, K., Tan, X., Peng, F. & Lu, J. Hybrid self-attention network for machine translation. *arXiv preprint arXiv:1811.00253* (2018).
13. Paulus, R., Xiong, C. & Socher, R. A deep reinforced model for abstractive summarization (2017). 1705.04304.
14. Lin, Z. *et al.* A structured self-attentive sentence embedding (2017). 1703.03130.
15. Song, H., Rajan, D., Thiagarajan, J. J. & Spanias, A. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence* (2018).
16. Choi, E. *et al.* Retain: An interpretable predictive model for healthcare using reverse time attention mechanism (2016). 1608.05745.
17. Zhang, D. *et al.* An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns* **2**, 100196 (2021).
18. Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).
19. Sun, K. & Marasović, A. Effective attention sheds light on interpretability. *arXiv preprint arXiv:2105.08855* (2021).
20. Martinez, F., Tobar, C. & Hill, N. Preventing delirium: should non-pharmacological, multicomponent interventions be used? a systematic review and meta-analysis of the literature. *Age ageing* **44**, 196–204 (2015).
21. Pollard, T. J. *et al.* The eicu collaborative research database, a freely available multi-center database for critical care research. *Sci. data* **5**, 180178 (2018).
22. Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Sci. data* **3**, 160035 (2016).
23. Bhattacharyya, A. *et al.* Delirium prediction in the ICU: designing a screening tool for preventive interventions. *JAMIA Open* **5**, DOI: 10.1093/jamiaopen/ooac048 (2022). Ooac048, https://academic.oup.com/jamiaopen/article-pdf/5/2/ooac048/44002754/ooac048.pdf.
24. Bergeron, N., Dubois, M.-J., Dumont, M., Dial, S. & Skrobik, Y. Intensive care delirium screening checklist: evaluation of a new screening tool. *Intensive care medicine* **27**, 859–864 (2001).
25. Brummel, N. E. *et al.* Implementing delirium screening in the intensive care unit: secrets to success. *Critical care medicine* **41**, 2196 (2013).

26. Bhattacharyya, A. *et al.* 400: Predicting delirium risk for the following 24 hours in critically ill patients using deep learning. *Critical Care Medicine* **48**, 182–182, DOI: 10.1097/01.ccm.0000619952.70488.fb (2020).

27. Sheikhalishahi, S., Balaraman, V. & Osmani, V. Benchmarking machine learning models on multi-centre eicu critical care dataset (2019). 1910.00964.

28. Guo, C. & Berkhahn, F. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737* (2016).

29. Sheikhalishahi, S. Machine learning applications in intensive care unit. *Univ. Trento Thesis Collect.* DOI: 10.15168/11572_339274 (2022).

30. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* **18**, 602–610 (2005).

31. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The J. Mach. Learn. Res.* **11**, 2079–2107 (2010).

32. Brunner, G. *et al.* On identifiability in transformers. *arXiv preprint arXiv:1908.04211* (2019).

33. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

34. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

35. Ismail, A. A., Gunady, M., Bravo, H. C. & Feizi, S. Benchmarking deep learning interpretability in time series predictions. *arXiv preprint arXiv:2010.13924* (2020).

36. Schubert, M. *et al.* A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients-a cohort study. *BMC health services research* **18**, 1–12 (2018).

37. Vilone, G. & Longo, L. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093* (2020).

38. Baghaei, K. T. & Rahimi, S. Sepsis prediction: an attention-based interpretable approach. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6 (IEEE, 2019).

39. Chen, P. *et al.* Interpretable clinical prediction via attention-based neural network. *BMC Med. Informatics Decis. Mak.* **20**, 1–9 (2020).

40. Kang, Y. *et al.* A clinically practical and interpretable deep model for icu mortality prediction with external validation. In *AMIA Annual Symposium Proceedings*, vol. 2020, 629 (American Medical Informatics Association, 2020).