Supplementary Material for DeepAISE - An Interpretable and Recurrent Neural Survival Model for Early Prediction of Sepsis

Supreeth P. Shashikumar[1], Christopher Josef[2], Ashish Sharma[3], Shamim Nemati[1*]

[1]Department of Biomedical Informatics, University of California San Diego Health, La Jolla, USA.

[2]Department of Surgery, Emory University School of Medicine, Atlanta, USA

[3]Department of Biomedical Informatics, Emory University, Atlanta, USA

[*]To whom correspondence should be addressed; E-mail: snemati@health.ucsd.edu

Materials and Methods

Appendix A

Survival analysis

In survival analysis, the objective is to model the time until an event (such as death) as a function of a set of covariates. We consider the onset of sepsis as the event of interest in this work. The longitudinal data of every patient is split into consecutive windows of 1 hour duration with the survival data for each of the windows comprising of three elements - a set of features *x*, the time to sepsis event τ and sepsis event indicator *e*. If a sepsis event occurs within the prediction horizon, the time interval τ will correspond to the duration of time between the time at which sepsis event occurs and the time of collection of features *x*, with the sepsis event indicator *e* being equal to 1. If sepsis does not occur within the prediction horizon, the time interval τ will correspond to one hour more than the duration of the prediction horizon, with the sepsis event indicator *e* being equal to 0 (this indicates $right - censoring$). Figure S1 shows an example of survival data of a patient for prediction horizon of 4 hours.

We further define two fundamental concepts in survival analysis - survival function, and hazard function. The survival function $S(m)$ is the probability of not getting sepsis in the proceeding m hours from the current time, and is given by $S(m) = P(\tau \geq m)$. The probability that a sepsis event occurrs within the proceeding m hours, $F(m)$, is then given by $F(m) = 1 - S(m)$.

The hazard function $H(m)$, defined in Equation (S1), is the instantaneous risk of sepsis at time m. In other words, $H(m)$ gives the conditional probability that sepsis will occur in the time bin (of 1 hour duration) m, given that it did not occur until time m:

$$H(m) = \lim_{\Delta m \to 0} \frac{P(m \leq \tau < m + \Delta m | \tau \geq m)}{\Delta m} \tag{S1}$$

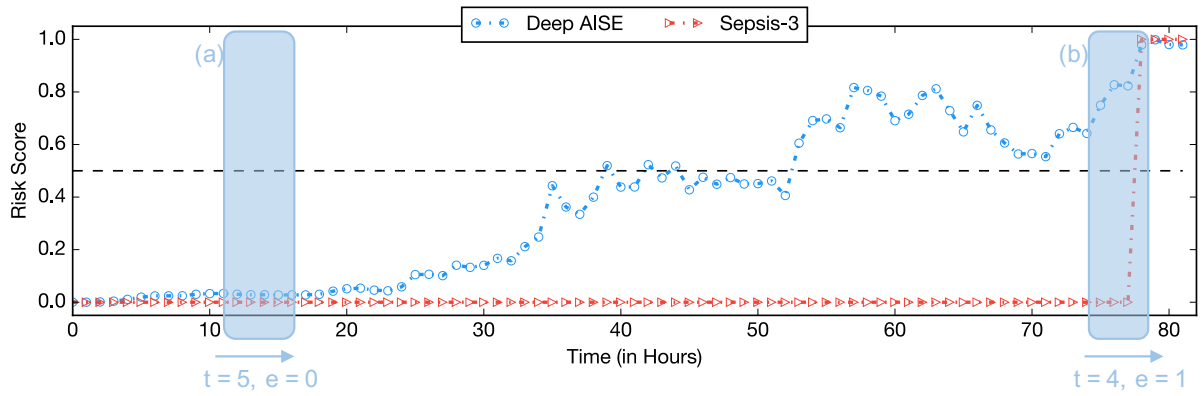


Figure S1: **Example of Survival data for a patient admitted to the ICU.** A sliding window (in this case, 4 hours duration) moves through the time series, hour-by-hour. If no sepsis event occurs within the prediction horizon (case (a)), we set e=0 indicating right censoring and τ=5. If a sepsis event occurs within the prediction horizon, the time to sepsis is recorded (τ=4 in case (b)) and we set e=1 indicating a sepsis event. Finally, if the patient record was terminated within a given prediction horizon (due to death, transfer or discharge from the ICU) that window was also marked as censored.

In the case of cox proportional hazards model, the hazard function is characterized by a combination of baseline hazard function (or the population level risk), $H_0(m)$, and the patient specific sepsis risk which is a function of the patient's features (vitals, labs, etc.), denoted by $g(x)$. The hazard function in a cox proportional hazards model will be of the form $H(m) = H_0(m) \cdot \exp(g(x))$. The survival function can then be expressed in terms of the hazard function as follows :

$$S(m) = \exp\left(-\int_0^m H(x)\,dx\right) \tag{S2}$$

Appendix B

Deep Artificial Intelligence Sepsis Expert (DeepAISE)

For each patient admitted to the Intensive Care Unit (ICU), the goal of the proposed DeepAISE model was to predict (at a regular interval of 1 hour) the probability of onset of sepsis, using all data available for the patient up until the time of prediction. Figure S2 provides an overview of the proposed DeepAISE model.

Notations

The notations that are followed throughout the rest of the paper is described as follows: For a total of N patients admitted to the ICU, we considered a dataset $D = \{D_i\}_{i=1}^{N}$ with D_i = { $\mathbf{X}_i$, T_i, $\mathbf{e}_i$, $\mathcal{T}_i$ }. The length of time series for patient i is denoted by T_i. A total of 65 features are measured/computed every hour for every patient in the ICU. The features for patient i is represented by $\mathbf{X}_i$ = [$\mathbf{x}_i^1$,$\mathbf{x}_i^2, \ldots$,$\mathbf{x}_i^{T_i}$], where $\mathbf{x}_i^t \in \mathbb{R}^{65}$ is the feature vector at time step t. The sepsis event indicator for patient i is denoted by $\mathbf{e}_i$ = $[e_i^1, e_i^2, \ldots, e_i^{T_i}]$ where at time step t, e_i^t = 0 if onset of sepsis does not occur within the prediction horizon otherwise e_i^t = 1. $\mathcal{T}_i$ = $[\tau_i^1, \tau_i^2, \ldots, \tau_i^{T_i}]$ represents the time to sepsis event for patient i.

Model

We consider the prediction of onset of sepsis as a sequential prediction task in our study. To achieve this, the proposed DeepAISE model, shown in Figure S2, employs a combination of a 2 layer stacked Gated Recurrent Unit (GRU) framework and a modified weibull-cox proportional hazards model (WCPH) to predict the onset of sepsis at a regular interval of 1 hour.

Let us consider a sequence of data of length T_i belonging to patient i. At each timestep t the stacked GRU model takes in as input, the feature vector $\mathbf{x}_i^t$ and stores the temporal information

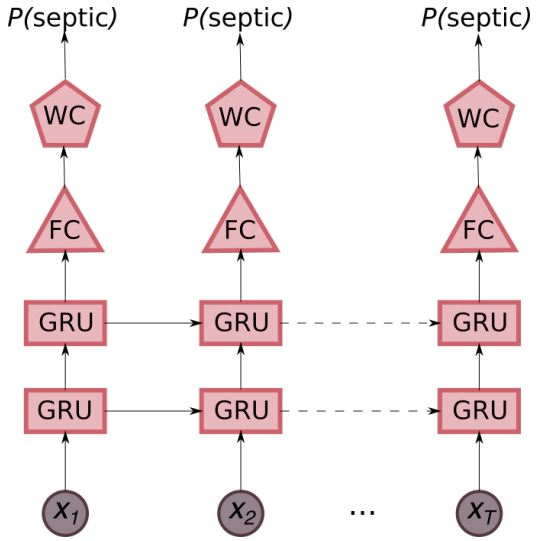


Figure S2: **Schematic diagram of the Deep Artificial Intelligence Sepsis Expert (DeepAISE) model.** The 65 features that are measure/computed every hour are fed sequentially into a 2 layer stacked GRU framework, the output from the stacked GRU layer is then fed into a fully connected layer, and a modified Weibull Cox Proportional Hazards Model (WCPH) is employed to compute the probability of occurrence of sepsis within the proceeding m hours (denoted by $F_t(m)$, with $t = [1, 2, \dots T]$). In our work, we are interested in the prediction of onset of sepsis 4 hours in advance.

within it's hidden layers. The GRU model used in our study is composed of 3 components at every timestep t: the $reset$ gate $\mathbf{r}_t$, the $update$ gate $\mathbf{z}_t$, and the hidden layer $\mathbf{h}_t$. The hidden layer $\mathbf{h}_t$ is computed as follows :

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z\mathbf{x}^t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{x}^t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r) \\
\widetilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h\mathbf{x}^t + \mathbf{r}_t \odot \mathbf{U}_h\mathbf{h}_{t-1} + \mathbf{b}_h) \\
\mathbf{h}_t &= \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \widetilde{\mathbf{h}}_t
\end{aligned} \tag{S3}$$

where $\sigma()$ is the logistic sigmoid function, $\odot$ denotes element-wise multiplication (or Hadamard product) and $\mathbf{W}_{\{z,r,h\}}$, $\mathbf{U}_{\{z,r,h\}}$, $\mathbf{b}_{\{z,r,h\}}$ are the weight matrices and bias terms associated with the calculation of $update$ gate, $reset$ gate, and hidden unit respectively. In this formulation of the GRU, if the $reset$ gate is close to 0 the intermediate hidden layer $\widetilde{\mathbf{h}}_t$ will ignore the previous

hidden state $\mathbf{h}_{t-1}$ and reset with the current input $\mathbf{x}^t$, essentially allowing the hidden state to ignore any information that would be irrelevant later. The $update$ gate controls the extent of information carried over from the previous hidden state $\mathbf{h}_{t-1}$ to the current hidden state $\mathbf{h}_t$. This behavior of the GRU helps the DeepAISE model to remember long-term dependencies that are present in the sequential data, and aids the model in identifying and capturing information that is only necessary for prediction of onset of sepsis.

In our proposed DeepAISE model, we stack 2 layers of GRU on top of each other to increase the representational power of the model. We represent the GRUs in the stacked layer as $\mathrm{G}^{(1)}$ and $\mathrm{G}^{(2)}$, with $\mathrm{G}^{(l)}$ = $\{\mathbf{W}^{(l)}_{\{z,r,h\}}, \mathbf{U}^{(l)}_{\{z,r,h\}}, \mathbf{b}^{(l)}_{\{z,r,h\}}\}$ $\forall\ l = 1, 2$. Correspondingly $\mathbf{r}^{(l)}_t$, $\mathbf{z}^{(l)}_t$, and $\mathbf{h}^{(l)}_t$ would be the output of $reset$ gate, $update$ gate and the hidden state of the GRU in layer l at timestep t.

The output $\mathbf{h}^{(2)}_t$ from the stacked GRU layer is then fed into a fully connected layer before being fed into the modified Weibull-cox proportional hazards model, for predicting onset of sepsis in the proceeding m hours (where m = 2, 4, 6, 8, 10 or 12 hours). The weibull-cox proportional hazards is a more robust parametric counterpart to the more familiar cox proportional hazards model *(44)* . The Weibull-Cox model defines the baseline hazard function as $H_0(m|\lambda, \nu) = (\nu/\lambda)(m/\lambda)^{\nu-1}$, where $\lambda > 0$ is a scale parameter and $\nu > 0$ is a shape parameter. For a patient i, the hazard function $H(m)$ at time step t is then defined as -

$$H_{it}(m|x_{it}, \theta, \beta, \lambda, \nu) = H_0(m|\lambda, \nu)\exp(\beta^T f(x^t_i)) \tag{S4}$$

where $f()$ is the output from the fully connected layer in the deep learning pipeline, β is a L dimensional weight vector ($\beta \in \mathbb{R}^L$), and θ = $\{\mathrm{G}^{(1)}, \mathrm{G}^{(2)}, \mathbf{W}_{fc}, \mathbf{b}_{fc}\}$. $\mathbf{W}_{fc}$ and $\mathbf{b}_{fc}$ denote the weights and bias term of the fully connected layer respectively. The survival function (i.e. probability of not getting sepsis in the proceeding m hours from current time step t) is given by

-

$$S_t(m|x_{it}, \theta, \beta, \lambda, \nu) = \exp(-\Lambda_0(m) \exp(\beta^T f(x_i^t)) \tag{S5}$$

where $\Lambda_0(m) = (m/\lambda)^\nu$ is the cumulative base hazard rate. The probability that onset of sepsis occurs with the proceeding m hours is then given by $F_t(m) = 1 - S_t(m)$.

Learning model parameters

Given the dataset D, we would like to compute the posterior probability over the parameters of the model which is defined as $p(\theta, \beta, \lambda, \nu|D) \propto p(D|\theta, \beta, \lambda, \nu)p(\theta)p(\beta)p(\lambda)p(\nu)$. We assume that $p(\theta)$, $p(\beta)$, $p(\lambda)$ and $p(\nu)$ are constant, and therefore maximization of the posterior probability is nothing but maximization of the likelihood of the data. The parameters of the proposed model is then learned through the maximum likelihood approach, wherein the log likelihood of the data is maximized (or the negative log likelihood of the data is minimized). The data likelihood is given in Equation (S6).

$$\begin{aligned} P(D|\theta, \beta, \lambda, \nu) &= \prod_{i=1}^{N} \prod_{t=1}^{T_i} [H_0(\tau_t^t|\lambda, \nu) \exp(\beta^T f(x_i^t))]^{e_i^t} S(\tau_i^t|x_i^t, \lambda, \nu, \beta, \theta) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T_i} [H_0(\tau_i^t|\lambda, \nu) \exp(\beta^T f(x_i^t))]^{e_i^t} \exp(-\Lambda_0(\tau_i^t) \exp(\beta^T f(x_i^t)) \end{aligned} \tag{S6}$$

where for patient i at time step t, $\mathbf{x}_i^t \in \mathbb{R}^{65}$ is the feature vector, e_i^t is the sepsis event indicator, and τ_i^t represents the time to sepsis event.

Further, the negative log-likelihood of data is then denoted by -

$$\mathcal{L}(\theta, \beta, \lambda, \nu) = -\frac{1}{N} \log P(D|\theta, \beta, \lambda, \nu) \tag{S7}$$

We then follow a mini-batch stochastic gradient descent approach to learn the optimal parameters of the model, by minimizing $\mathcal{L}(\theta, \beta, \lambda, \nu)$. Intuitively, maximizing the data likelihood (or minimizing the negative log-likelihood) will correspond to a) maximizing the probability that

sepsis does not occur before time τ_i^t and b) maximizing the probability of actual sepsis events, when the events are not censored (i.e. $e_i^t = 1$).

Appendix C

Input features

The complete list of the input features to the model is as follows -

1. High-resolution dynamical features (calculated using 6 hours sliding windows, with 5 hours overlap; 6 features)

    - standard deviation of RR intervals and Mean Arterial Blood Pressure (RR_{STD} and MAP_{STD}), average multiscale entropy of RR and MAP (HRV_1 and BPV_1) and average multiscale conditional entropy of RR and MAP (HRV_2 and BPV_2).

2. Clinical features (10 features)

    - Mean Arterial Blood Pressure (MAP), Heart Rate (HR), Oxygen Saturation (O_{2Sat}), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiratory Rate ($RESP$), Temperature ($Temp$), Glasgow Coma Scale (GCS), Partial Pressure of Arterial Oxygen (PaO_2), Fraction of Inspired O2 (FiO_2).

3. Laboratory (General; 25 features)

    - White Blood Count (WBC), Hemoglobin, Hematocrit, Creatinine, Bilirubin and Bilirubin Direct, Platelets, International Normalized Ratio (INR), Partial Prothrombin Time (PTT), Aspartate Aminotransferase (AST), Alkaline Phosphatase, Lactate, Glucose, Potassium, Calcium, Blood rea nitrogen (BUN), Phosphorus, Magnesium, Chloride, B-type Natriuretic Peptide (BNP), Troponin, Fibrinogen, CRP, Sedimentation Rate, Ammonia.

4. Laboratory (Arterial Blood Gas or ABG; 5 features):

- pH, pCO_2, HCO_3, Base Excess, SaO_2.

5. Demographics/History/Context (19 features)

    - Care Unit (Surgical, Cardiac Care, or Neurointensive care), Surgery in the past 12 hours, Wound Class (clean, contaminated, dirty, or infected), Surgical Specialty (Cardiovascular, Neuro, Ortho-Spine, Oncology, Urology, etc.), Number of antibiotics in the past 12, 24, and 48 hours, Age, Charleston Comorbidity Index (CCI), Mechanical Ventilation, maximum change in SOFA score over the past 6 hours

All dynamic features were organized into 1-hour non-overlapping time series bins to accommodate for different sampling frequencies of available data. The 1-hour time bin interval was selected as a balance between having short windows with too many missing data points (low-frequency clinical data) and having time windows too long to make any meaningful prediction. Non-overlapping bins simplified the modeling schema by minimizing autocorrelation. EMR features with sampling frequencies higher than once every hour were uniformly resampled into 1-hour time bins, by taking the median values if multiple measurements were available. Features were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). The renal component of the SOFA score was slightly modified to account for poor data quality of urine output, and only used serum creatinine. Otherwise, the SOFA score was calculated as outlined in [1] . Mean imputation was used to replace all remaining missing values (mainly at the start of each record)

The bedside monitor data (HR and MAP with 0.5 Hz resolution) was matched and time synchronized to each patients EMR data. The following features from the HR and MAP time series were derived from the bedside monitor's proprietary software using the ECG and blood pressure waveforms: standard Deviation of HR (HR_{STD}), Standard Deviation of MAP (MAP_{STD}), Multiscale Entropy [2] [3] of R-R intervals (60/HR) and MAP (HRV1, HRV2 and BPV1, BPV2,

respectively). The time series-related features were updated every hour, using a 6-hour sliding window with five hours overlap. For each window, 17 different scales (scales 1, 4, 7, ... 49) were considered for all variability measurements of heart rate (HR) and blood pressure (MAP), and the average value of multiscale entropy and conditional entropy over all scales were included as features in the model.

Table S1: Tabulation of features that are present or absent in a cohort. (Y: Yes/Present, N: No/Absent)

Feature Name	Emory cohort	UCSD cohort	MIMIC cohort
Clinical features			
MAP	Y	Y	Y
HR	Y	Y	Y
O_2Sat	Y	Y	Y
SBP	Y	Y	Y
DBP	Y	Y	Y
Resp	Y	Y	Y
Temp	Y	Y	Y
GCS	Y	Y	Y
PaO_2	Y	Y	N
FiO_2	Y	Y	N
Laboratory features			
WBC	Y	Y	Y
Hemoglobin	Y	Y	Y
Hematocrit	Y	Y	Y
Creatinine	Y	Y	Y
Bilirubin Total	Y	Y	Y
Bilirubin Direct	Y	Y	N
Platelets	Y	Y	Y
INR	Y	Y	Y
PTT	Y	Y	Y
AST	Y	Y	Y
Alkaline Phosphate	Y	Y	Y
Lactate	Y	Y	Y
Glucose	Y	Y	N
Potassium	Y	Y	Y
Calcium	Y	Y	Y
BUN	Y	Y	Y
Phospshorous	Y	Y	Y
Magnesium	Y	Y	Y
Chloride	Y	Y	Y
BNP	Y	N	N
Troponin	Y	Y	Y
Fibrinogen	Y	Y	N
CRP	Y	N	N
Sedimentation rate	Y	N	N
Ammonia	Y	N	N
pH	Y	Y	N
$PaCO_2$	Y	Y	N
HCO_3	Y	Y	N
Base Excess	Y	Y	N
SaO_2	Y	Y	N
High resolution dynamical features			
RR_{STD}	Y	N	N
MAP_{STD}	Y	N	N
HRV_1	Y	N	N
HRV_2	Y	N	N
BPV_1	Y	N	N
BPV_2	Y	N	N
Demographics/History/Contextual features			
Care Unit	Y	Y	Y
Surgery	Y	Y	N
Wound class	Y	Y	N
Surgical Specialty	Y	Y	N
#ABX 12 hours	Y	Y	Y
#ABX 24 hours	Y	Y	Y
#ABX 48 hours	Y	Y	Y
Age	Y	Y	Y
CCI	Y	Y	Y
Mechanical Ventilation	Y	Y	Y
SOFA	Y	Y	Y

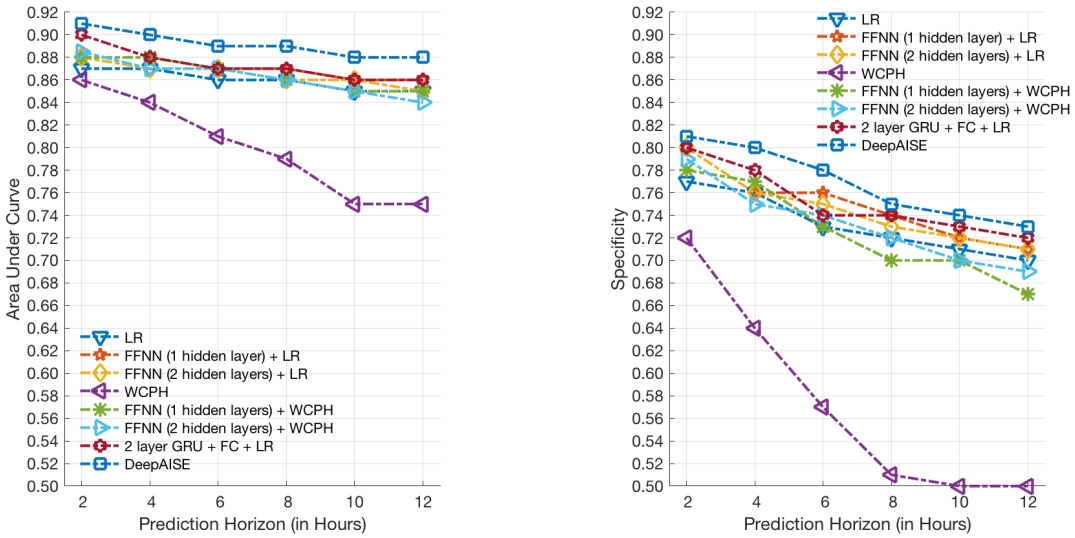
Appendix D

Emory cohort

Table S2: Summary of patient characteristics of the Emory dataset

Model	All Patients	Non-Septic	Septic
Patients, (#)	25820	24375	1445
			6%
Male, no. (%)	53.3	53.2	55.2
Age, median (IQR) y	61	61	61.5
	[49 - 71]	[49 - 71]	[50.5 - 72]
Race, no. (%)			
Caucasian	48.6	48.9	45.0
Black	43.3	43.1	45.4
Asian	1.3	1.3	1.3
Hispanic	0.024	0.02	0.08
Surgery (%)			
Cardiovascular	13.1	-	-
Neuro	6.1	-	-
Ortho-spine	1.8	-	-
Oncology/General Surgery	3.6	-	-
Urology	0.4	-	-
ICU LOS, median (IQR) h	48	46	141
	[28 - 90]	[27 - 77]	[77 - 258]
Inpatient Mortality, %	4.1	3.5	15.2
Inpatient Hospice, %	3.8	3.3	12.5
SOFA, median (IQR)	1.9	1.7	5.0
	[0.6 - 4.0]	[0.5 - 3.6]	[3.1 - 7.4]
CCI, median (IQR)	2	2	3
	[1 - 4]	[1 - 4]	[2 - 5]
ICU Admission to $t_{sepsis-3}$, median (IQR) h	-	-	24 [9 - 63]

(a) (b)



** LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure S3: **Comparison of Emory testing set performance of all baseline models and DeepAISE to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours.** The Area Under the Curve (AUC) is shown in the left panel. The Specificity (SPC) is shown in the right panel.

Table S3: Summary of Emory testing set prediction performance of DeepAISE model in predicting $t_{sepsis-3}$ 4 hours in advance. The DeepAISE model consists of a 2 layer GRU, a fully connected layer and WCPH model. The Area Under the Curve (AUC), Specificity (SPC) and Accuracy (ACC) are reported for both training set and testing set

	Testing set (Training Set)		
Model	**AUC**	**SPC**	**ACC**
LR	0.87 (0.89)	0.76 (0.79)	0.76 (0.79)
FFNN (1 hidden layer) + LR	0.88 (0.92)	0.76 (0.85)	0.77 (0.85)
FFNN (2 hidden layers) + LR	0.87 (0.92)	0.76 (0.85)	0.78 (0.85)
WCPH	0.84 (0.86)	0.64 (0.71)	0.64 (0.71)
FFNN (1 hidden layer) + WCPH	0.88 (0.92)	0.77 (0.85)	0.77 (0.85)
FFNN (2 hidden layers) + WCPH	0.87 (0.93)	0.75 (0.88)	0.75 (0.88)
2 layer GRU + FC + LR	0.88 (0.90)	0.78 (0.88)	0.78 (0.88)
DeepAISE	**0.90 (0.94)**	**0.80 (0.89)**	**0.78 (0.82)**

* *LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

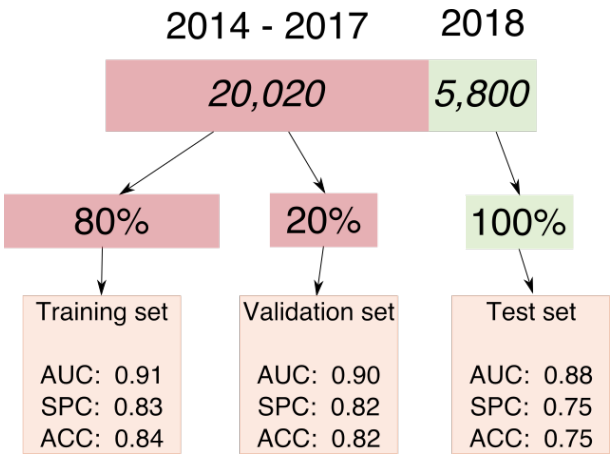


Figure S4: **Performance of DeepAISE that was first trained on Emory year-based training set (patients in Emory cohort from the year 2014 through 2017) and then applied to a heldout test set collected from 2017 to 2018 (Emory year-based holdout set).**

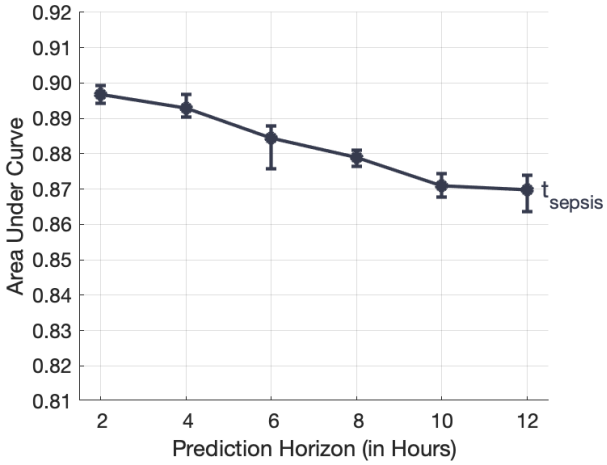


Figure S5: 10-fold Cross Validation performance of DeepAISE on Emory cohort. We performed bagging with replacement on the Emory cohort to produce 10 different datasets with each set comprising of 80% training set and 20% holdout set, and we measured the performance of the model (Area under the Curve) on the hold outs sets across the 10 folds (Emory CV holdout set). Median values are shown for different prediction horizons with error bars representing the interquartile range.

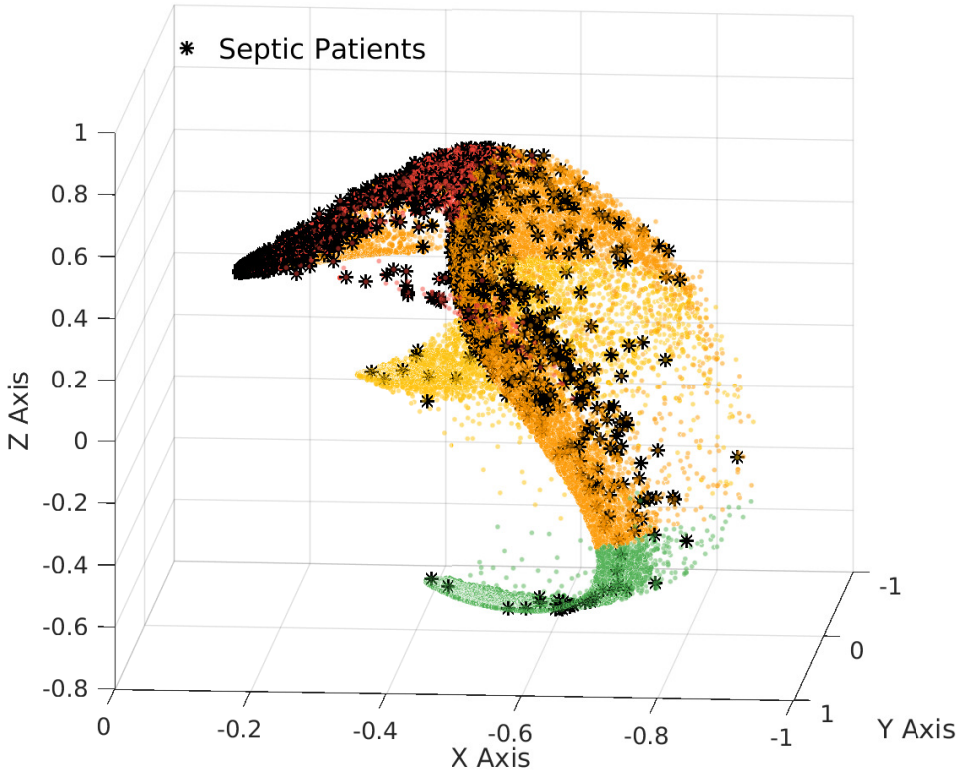


Figure S6: **Visualization of DeepAISE time series covariates performed by spectral clustering, with septic patients represented by asterisk.** The colors for the patients in the plots were chosen based on the predicted sepsis risk score (green represents the lowest predicted sepsis risk score, and red represents the highest predicted sepsis risk score).

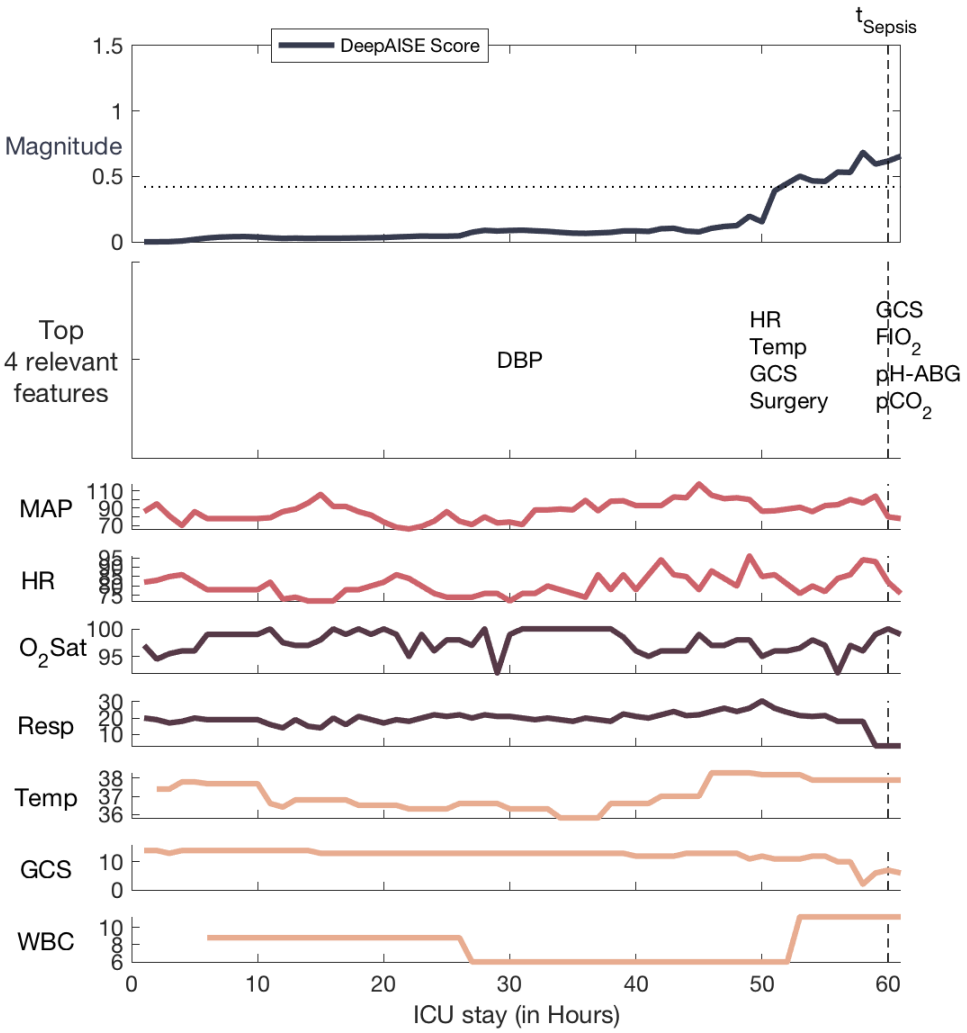


Figure S7: **DeepAISE score shown for Patient #1 (P1).** Commonly recorded hourly vital signs of the patient, including heart rate (HR), mean arterial blood pressure (MAP), respiratory rate ($RESP$), temperature ($TEMP$), oxygen saturation (O_2Sat) are shown. The most significant features contributing to the DeepAISE score are listed immediately below the DeepAISE Scores (for clarity of presentation, only selected time points are shown). The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Refer to Appendix C of Supplementary Material for more details on the abbreviated features.

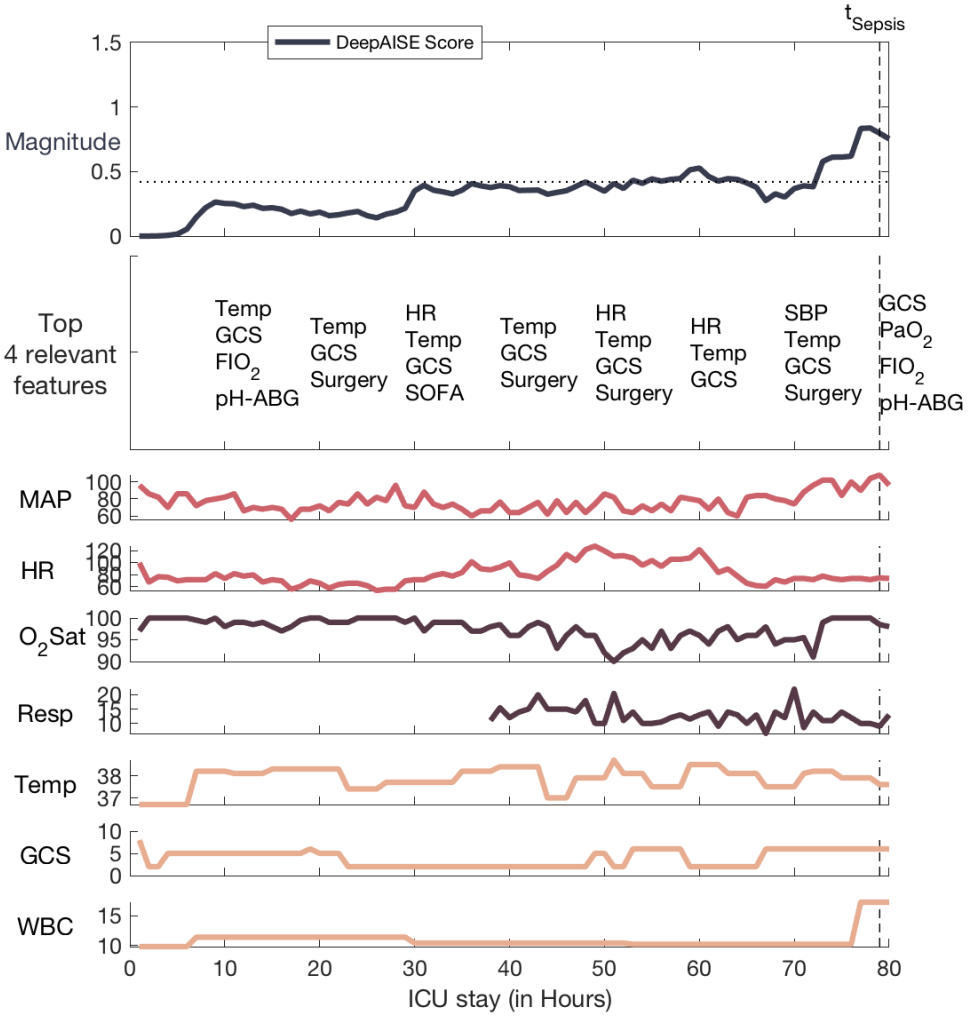


Figure S8: **DeepAISE score shown for Patient #2 (P2).** Commonly recorded hourly vital signs of the patient, including heart rate (HR), mean arterial blood pressure (MAP), respiratory rate ($RESP$), temperature ($TEMP$), oxygen saturation (O_2Sat) are shown. The most significant features contributing to the DeepAISE score are listed immediately below the DeepAISE Scores (for clarity of presentation, only selected time points are shown). The horizontal dashed line indicates the prediction threshold corresponding to a sensitivity of 0.85. Refer to Appendix C of Supplementary Material for more details on the abbreviated features.

Appendix E

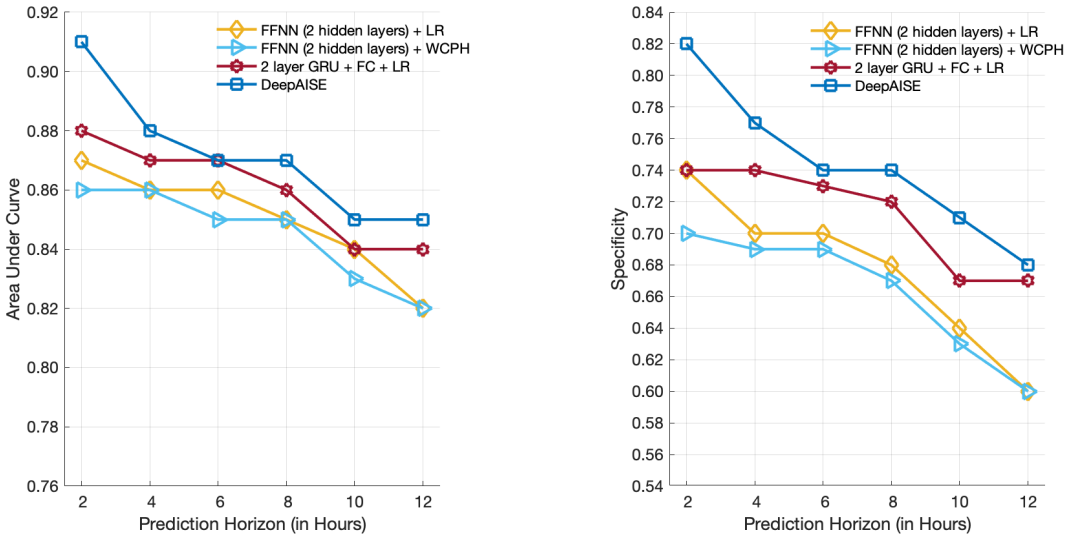
UCSD Cohort

The UCSD cohort contained a total of 18,752 patients, 1073 of whom developed sepsis in the ICU. Out of the 18,752 patients, 80% of them were used for developing the model (training set), and the remaining 20% form the testing set.

Table S4: Summary of patient characteristics of UCSD cohort

Model	All Patients	Non-Septic	Septic
Patients, no.	18752	17679	1073
			5.7%
Male, no. (%)	61.1	60.8	66.5
Age, median (IQR) y	59.8	59.7	60.8
	[46.4 - 70.8]	[46.3 - 70.8]	[46.9 - 70.3]
Race, no. (%)			
Caucasian	52.2	52.5	47.4
Black	7.9	7.9	6.2
Asian	5.4	5.4	6.4
ICU LOS, median (IQR) h	44.8	43.3	143.8
	[24.3 - 79.6]	[23.8 - 72.9]	[78.5 - 241.2]
Mortality, %	4.7	3.7	21.1
SOFA, median (IQR)	2	1	4
	[0 - 4]	[0 - 3]	[2 - 6]
CCI, median (IQR)	3	2	3
	[1 - 6]	[1 - 3]	[2 - 6]
ICU Admission to $t_{sepsis-3}$, median (IQR) h	-	-	38 [16 - 74]

(a) (b)



* *LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure S9: Comparison of performance of baseline models and DeepAISE on the UCSD cohort to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours. a) The Area Under the Curve (AUC) is shown in the left panel. b) The Specificity (SPC) is shown in the right panel.

Appendix F

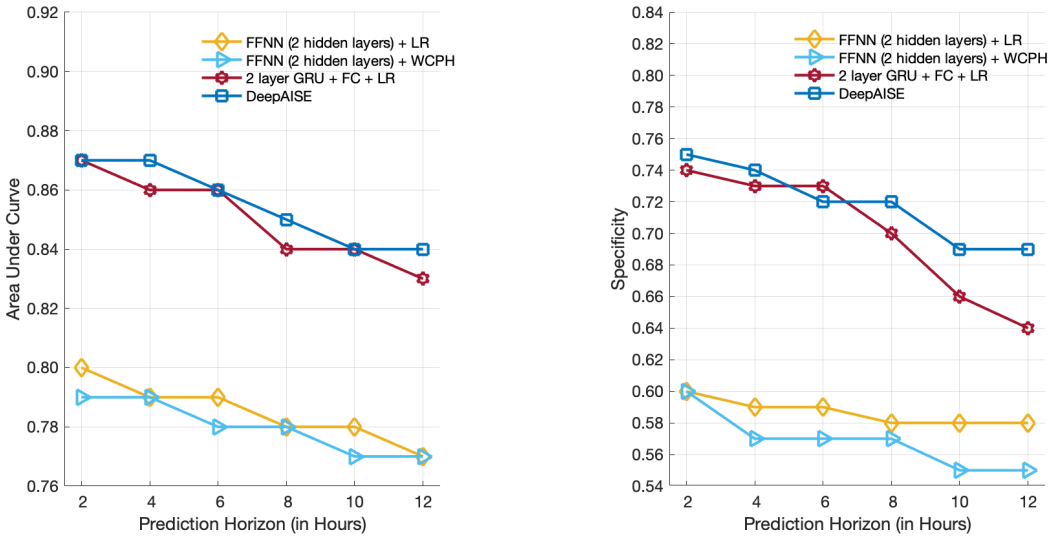
MIMIC-III Cohort

The MIMIC-III cohort contained a total of 40,474 patients, 2276 of whom developed sepsis in the ICU. Out of the 40,474 patients, 80% of them were used for developing the model (training set), and the remaining 20% form the testing set.

Table S5: Summary of patient characteristics of MIMIC-III cohort

Model	All Patients	Non-Septic	Septic
Patients, no.	40474	38198	2276
			5.6%
Male, no. (%)	56.5	56.3	58.8
Age, median (IQR) y	66	65	66
	[52 - 77]	[52 - 77]	[50 - 71]
Race, no. (%)			
Caucasian	71.6	71.5	72.4
Black	9.2	9.2	8.6
Asian	2.2	2.2	2.4
Hispanic	3.4	3.4	3.4
ICU LOS, median (IQR) h	47	45	158
	[27 - 88]	[26 - 74]	[83 - 266]
Mortality, %	8.9	8.1	22.0
SOFA, median (IQR)	1.6	1.5	3.3
	[0.65 - 3.1]	[0.6 - 2.9]	[2.0 - 5.1]
CCI, median (IQR)	2	2	3
	[1 - 3]	[1 - 3]	[1 - 4]
ICU Admission to $t_{sepsis-3}$, median (IQR) h	-	-	31.2 [13.3 - 70.2]

(a) (b)



* *LR = Logistic Regression layer, FFNN = Feedforward Neural Network, WCPH = Weibull Cox Proportional Hazard layer, FC = Fully Connected layer, AISE = Artificial Intelligence Sepsis Expert*

Figure S10: Comparison of performance of baseline models and DeepAISE on the MIMIC-III cohort to predict $t_{sepsis-3}$ for prediction horizons of 2, 4, 6, 8, 10, and 12 hours. a) The Area Under the Curve (AUC) is shown in the left panel. b) The Specificity (SPC) is shown in the right panel.

Appendix G

Table S6: Description of the various datasets used in the analysis of DeepAISE

Dataset	Description
Emory training	70% of patients from the entire Emory cohort
Emory testing	20% of patients from the entire Emory cohort
Emory hyperparameter optimization	10% of patients from the entire Emory cohort
Emory year-based training	Patients in Emory cohort from the year 2014 through 2017
Emory year-based holdout	Patients in Emory cohort from the year 2017 through 2018
MIMIC training	80% of patients from the entire MIMIC-III cohort
MIMIC testing	20% of patients from the entire MIMIC-III cohort
UCSD training	80% of patients from the entire UCSD cohort
UCSD testing	20% of patients from the entire UCSD cohort

Appendix H

Analysis of contribution of input features to the DeepAISE risk score

Unlike many other sepsis prediction algorithms, DeepAISE is uniquely interpretable wherein apart from computing the sepsis risk score, the model identifies the most relevant features contributing to the sepsis risk score as well. The importance of each feature's contribution to the risk score is measured through a metric called *relevance score*. The *relevance score* (R) is computed as $R = \frac{dY}{dX} * X$, where Y is the sepsis risk score, and X is the input feature.

Table S7: Sensitivity of DeepAISE performance to features with positive relevance score.

Metric	Local replacement	Global replacement	Random replacement †	No replacement
Area Under the Curve	0.88	0.89	0.899 [0.886, 0.901]	0.90
Sensitivity	0.51*	0.82*	-	0.85
Specificity	0.95*	0.83*	-	0.81

* *Sensitivity and Specificity measured at threshold corresponding to 0.85 sensitivity level of DeepAISE run on the entire Emory test set.*

† *10 features selected at random replaced with population mean (repeated 100 times). Result reported as median[IQR]*

Table S8: Sensitivity of DeepAISE performance to features with negative relevance score.

Metric	Local replacement	Global replacement	Random replacement †	No replacement
Area Under the Curve	0.83	0.88	0.899 [0.886, 0.901]	0.90
Sensitivity	0.96*	0.85*	-	0.85
Specificity	0.42*	0.78*	-	0.81

* *Sensitivity and Specificity measured at threshold corresponding to 0.85 sensitivity level of DeepAISE run on the entire Emory test set.*

† *10 features selected at random replaced with population mean (repeated 100 times). Result reported as median[IQR]*

We quantified the importance of the most relevant features (chosen based on the *relevance score*), by performing two experiments. In the first experiment, which we call *local feature replacement analysis*, for each hour that a patient was in the ICU we computed the relevance score of all the input features, and replaced 10 of them with the highest positive (or negative) relevance score by their population mean (this is roughly equivalent to treating them as missing

data). We then ran the DeepAISE model on these two new datasets and report the performance in Table S7 and Table S8. In the second experiment, which we call *global feature replacement analysis*, we identified 10 features that appeared the most common as a top 10 relevant feature starting 10 hours prior to and until $t_{sepsis-3}$ (two separate analysis were run for the positive relevance scores and the negative relevance scores). We then replaced these 10 global relevant features with the population mean, for the entire cohort. Finally, we report DeepAISE's performance on these new datasets in Table S7 and Table S8.

Understanding the effect of masking locally important features

In a nonlinear sequential model such as DeepAISE the relationship between the imput features and the model output is by no means obvious. As such, untangling this relationship requires careful analysis. In our analysis, we were interested in understanding the importance of features that were positively and negatively contributing to the sepsis risk score. **Features with positive relevance score:** These were the features for which a) $\frac{dY}{dX}$ was positive and X was positive (first quadrant in Figure S11a) or b) $\frac{dY}{dX}$ was negative and X was negative (second quadrant in Figure S11b). In both the above cases, when X is replaced with 0 (Note: All the features are normalized to a standard normal distribution, hence population mean is 0), the sepsis risk score Y decreases. Thus, when features that have a positive relevance score are replaced with the population mean, we would expect the sepsis risk score Y to drop. This should result in decreased sensitivity (reduction in true positive rate) and increased specificity (reduction in false alarm rate). **Features with negative relevance score:** These were the features for which a) $\frac{dY}{dX}$ was positive and X was negative (second quadrant in Figure S11a) or b) $\frac{dY}{dX}$ was negative and X was positive (first quadrant in Figure S11b). In both the above cases, when X is replaced with 0 (Note: All the features are normalized to a standard normal distribution, hence population mean is 0), the sepsis risk score Y increases. Thus, when features that have a

negative relevance score are replaced with the population mean, we would expect the sepsis risk score Y to increase. This should result in increases sensitivity (increase in true positive rate) and decreased specificity (increase in false alarm rate).

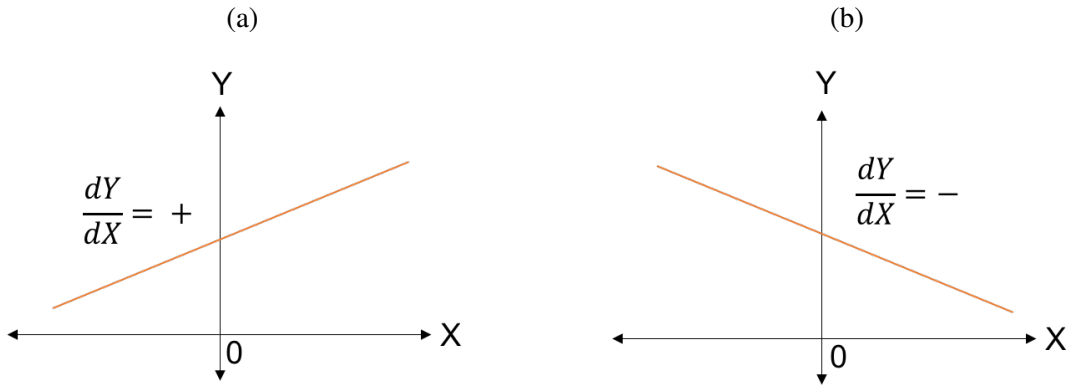


Figure S11: a) Plot of sepsis risk score (Y) against a single variable X, when the slope is positive i.e. $\frac{dY}{dX} = +$, and b) Plot of Y against a single variable X, when the slope is negative i.e. $\frac{dY}{dX} = -$

Features with positive relevance score

Local feature replacement analysis

Referring to Table S7, we see that by performing *local feature replacement analysis*, DeepAISE achieved an AUC of 0.88 on the modified Emory testing set. We also observed that when the threshold corresponding to 0.85 sensitivity level (0.81 specificity) of DeepAISE run on the entire Emory test set was used, the sensitivity dropped to 0.51 and the specificity increased to 0.95. These trends agree with our hypothesis that removing features that positively influence the sepsis risk score, will lead to reduced true positive rate.

Global feature replacement analysis

In the *global feature replacement analysis*, DeepAISE achieved an AUC of 0.89 on the modified Emory testing set. When the threshold corresponding to 0.85 sensitivity level of DeepAISE

run on the entire Emory test set was used, the sensitivity dropped to 0.82 and the specificity increased to 0.83.

Features with negative relevance score

Local feature replacement analysis

Referring to Table S8, we see that by performing *local feature replacement analysis*, DeepAISE achieved an AUC of 0.83 on the modified Emory testing set. We also observed that when the threshold corresponding to 0.85 sensitivity level (0.81 specificity) of DeepAISE run on the entire Emory test set was used, the sensitivity increased to 0.96 and the specificity dropped to 0.42. These trends agree with our hypothesis that removing features that negatively influence the sepsis risk score, will lead to increased true positive rate.

Global feature replacement analysis

In the *global feature replacement analysis*, DeepAISE achieved an AUC of 0.88 on the modified Emory testing set. When the threshold corresponding to 0.85 sensitivity level of DeepAISE run on the entire Emory test set was used, the sensitivity increased to 0.85 and the specificity increased to 0.78.

Appendix I

For each patient record, the percentage of missingness of each of the clinical and laboratory features was computed (for e.g. For patient 1, Missingness(HR) = $\frac{Number\ of\ time\ points\ where\ HR\ is\ missing\ in\ Patient\ 1}{Sequence\ length\ of\ Patient\ 1} x\ 100\ \%$). These percentages of missingness were then averaged across all the patients, and the resulting numbers are shown in Table S9. Each entry in Table S9 represents on average, the number of time points where new measurements are not available when an entire patient record is considered (for e.g. On average, over the length of ICU stay of a patient in the Emory cohort, HR is not measured for 10.95% of the time).

Table S9: Comparison of missing data (in %) per patient across the three cohorts considered in this study.

Feature Name	**Emory cohort (%)**	**UCSD cohort (%)**	**MIMIC cohort (%)**
Clinical features			
MAP	13.06	36.86	9.95
HR	10.95	11.91	7.40
O_2Sat	12.52	13.15	10.40
SBP	12.66	36.65	20.80
DBP	12.66	36.66	20.87
Resp	19.51	13.09	9.29
Temp	65.77	63.56	66.28
GCS	78.90	63.83	67.28
PaO_2	96.89	97.07	-
FiO_2	96.85	80.82	-
Laboratory features			
WBC	93.28	92.50	93.33
Hemoglobin	92.55	92.49	92.30
Hematocrit	92.58	92.54	89.44
Creatinine	93.16	92.50	92.58
Bilirubin Total	97.55	97.87	98.52
Bilirubin Direct	99.12	99.28	-
Platelets	93.53	92.51	92.81
INR	96.85	-	94.89
PTT	98.08	96.11	94.72
AST	97.10	97.87	98.50
Alkaline Phosphate	97.10	97.83	98.55
Lactate	97.53	98.64	97.13
Glucose	74.67	92.53	-
Potassium	92.76	92.06	88.66
Calcium	93.15	92.58	94.37
BUN	93.16	92.48	92.62
Phospshorous	96.09	94.63	94.32
Magnesium	94.18	94.19	93.16
Chloride	92.51	92.59	92.47
BNP	99.19	-	-
Troponin	97.27	98.65	98.74
Fibrinogen	99.03	99.41	-
CRP	99.46	-	-
Sedimentation rate	99.47	-	-
Ammonia	99.48	-	-
pH	96.93	97.05	-
$PaCO_2$	96.88	97.07	-
HCO_3	96.81	97.07	-
Base Excess	99.21	97.07	-
SaO_2	98.27	97.19	-

Appendix J

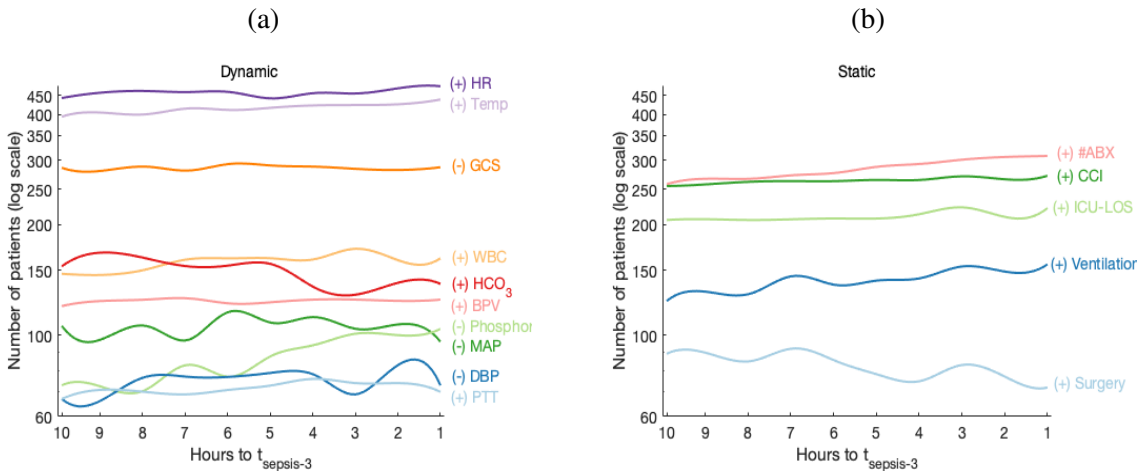


Figure S12: Most common features contributing to an elevated risk score.

Appendix K

Performance of DeepAISE under different levels of missingness

In this section, we evaluated the performance of DeepAISE algorithm (when fine tuned to a new cohort) under different levels of missingness of input features (specifically laboratory measurements). For every patient, we computed the percentage of missing laboratory measurements (the percentage of missing measurements was computed over a rolling 24 hour window, and were then averaged across all the windows). This number represented the percentage of laboratory measurements missing on average for a patient over a 24 hour window. The patients were then split into 3 different groups based on the percentiles of their missingness. Group 1 consisted of patients whose percentage of missingness fell below 33 percentile of the overall cohort. Group 2 consisted of patients whose percentage of missingness was above 33 percentile and below 66 percentile of the overall cohort. Group 3 consisted of patients whose percentage of missingess was above 66 percentile and below 100 percentile of the overall cohort. The performance of DeepAISE (on the entire UCSD cohort) for each of the above groups has been tabulated in Table S10. The Area under the Receiver operating characteristic curves and Area under precision recall curves for Groups 1, 2 and 3 are shown in Fig. S14 and Fig. S15.

Table S10: Performance of DeepAISE on the entire UCSD cohort for differing levels of missingness of input features. For reference, the percentage of missingness in Group 1 <Group 2 <Group3. (AUC: Area under the receiver operating characteristic curve. AUCpr: Area under the precision recall curve)

	Total patients	**Septic patients**	**AUC**	**AUCpr**
Group 1	6188	672 (10.85%)	0.871	0.278
Group 2	6188	561 (9.06%)	0.906	0.242
Group 3	6376	140 (2.19%)	0.916	0.179

Note: The Positive Predictive Value (or Precision) is defined as the ratio of number of true

positives to the sum of the true positives and false positives. In the scenario where the class labels are highly imbalanced (in our case very low prevalence of positive labels compared to negative labels), the positive predictive value (PPV) can get penalized by the false positives to a very large extent. It is also often the case that the predicted risk scores from a sequential prediction algorithm like DeepAISE can cross the decision threshold earlier than the 4-hours prediction horizon. For example, the DeepAISE risk score crossed the decision threshold about 12 hours prior to $t_{sepsis-3}$ for the patient shown in Figure S13. In this case, according to the definition of positive predictive value all the positive predictions up until 4 hours prior to $t_{sepsis-3}$ would be counted as false positives. This is not clinically optimal, as earlier warnings are still relevant. In order to not penalize the algorithm for making positive predictions before the expected 4 hours prediction horizon, during the computation of PPV we considered any positive predictions that occurred upto 24 hours prior to $t_{sepsis-3}$ as true positives (the blue shaded region in Figure S13). The AUCpr tabulated in Table S10 consists of PPV computed as described above.

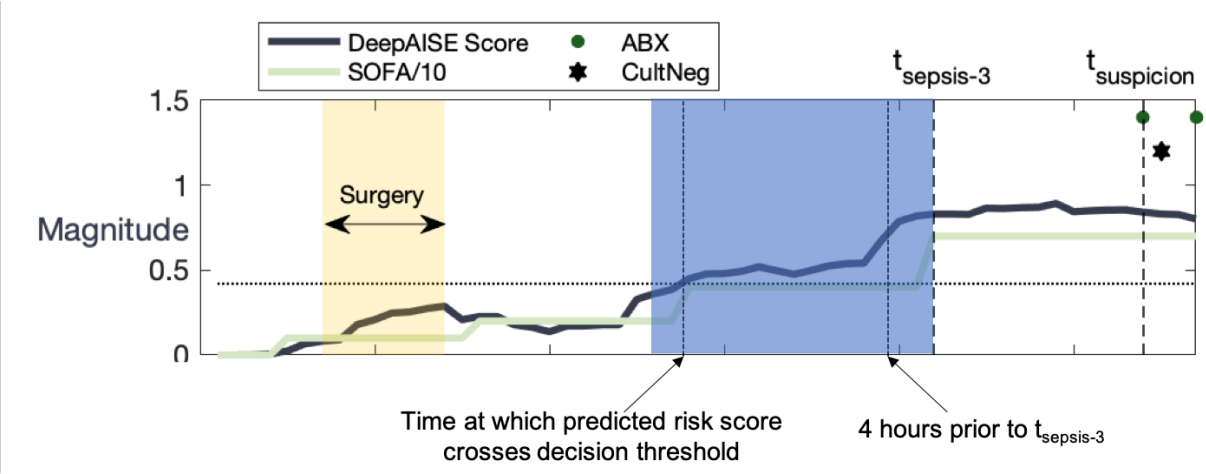


Figure S13: The DeepAISE risk score crosses the decision threshold about 12 hours prior to $t_{sepsis-3}$. In this case, according to the definition of positive predictive value all the positive predictions up until 4 hours prior to $t_{sepsis-3}$ would be counted as false positives. This is not clinically optimal, as earlier warnings are still relevant. In order to not penalize the algorithm for making positive predictions before the expected 4 hours prediction horizon, during the computation of PPV we considered any positive predictions that occurred upto 24 hours prior to $t_{sepsis-3}$ as true positives (the blue shaded region)

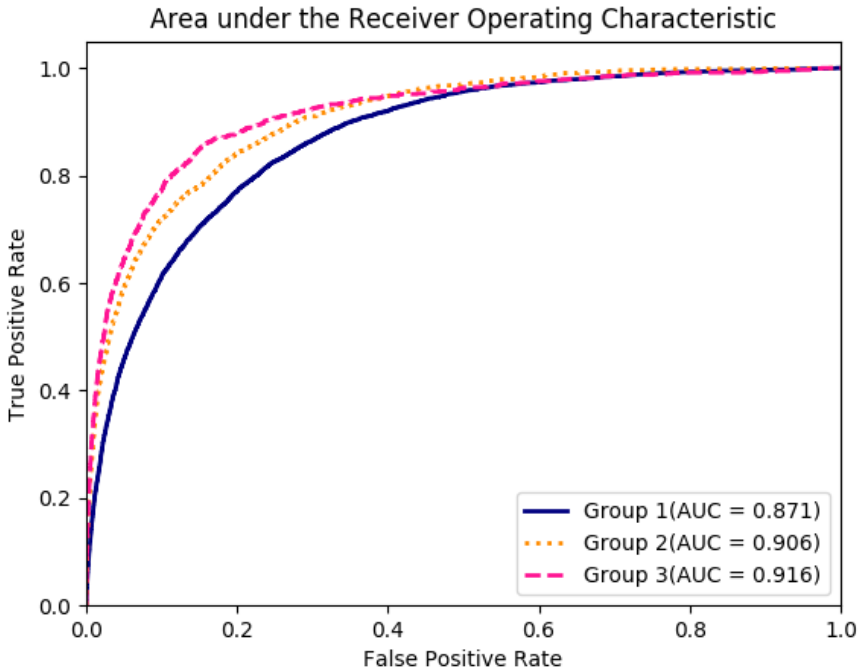


Figure S14: Area under the receiver operating characteristic curves for Groups 1, 2 and 3.

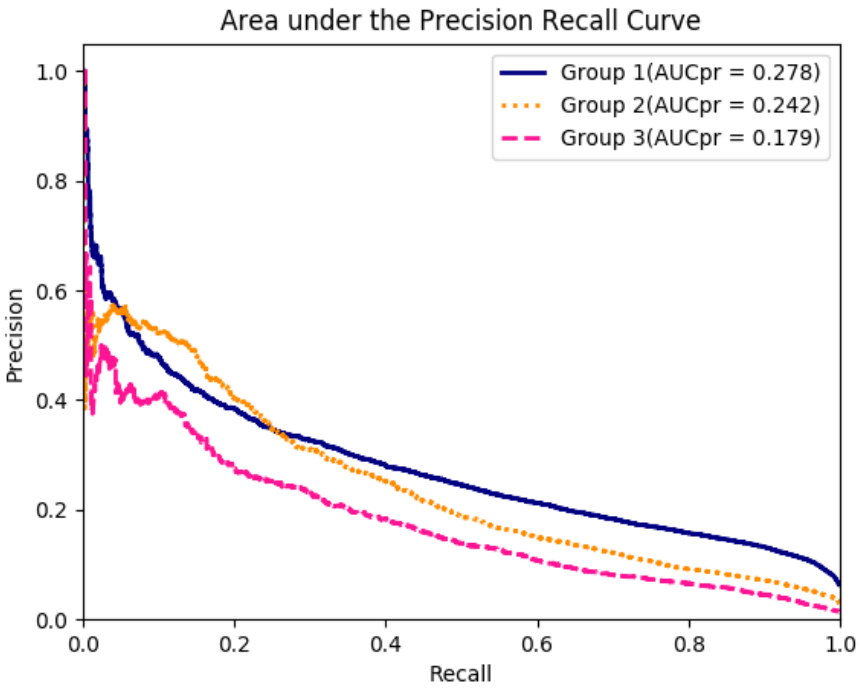


Figure S15: Area under the precision recall curves for Groups 1, 2 and 3.

Appendix L

DeepAISE User Interface

To facilitate real-time decision support with DeepAISE, a software platform was developed. The platform is responsible for interfacing with an EHR to obtain real-time patient data, computing hourly DeepAISE scores, and presenting the results in an interpretable, clinically meaningful fashion. The platform consists of four distinct microservices as shown in Fig. 6 of main paper. These microservices are containerized using Docker. The four microservices are:

- **DeepAISE Model Service:** The DeepAISE model is wrapped in a python wrapper and exposes a REST API that allows an external entity to run the model, in a stateless inference mode, and return a DeepAISE score given a set of observations about a patient. The stateless nature refers to the fact that, in a clustered deployment of the model, predictions about a patient are not tied to a specific worker node. This makes the system highly scalable and it can leverage popular orchestration systems such as Kubernetes and Docker Swarm to scale on-demand, in response to fluctuations in system load. Finally, for security needs, the system maintains an audit log of every request it receives, and whether the request was successfully executed.

- **Results Database Service:** This is the core data management layer of the DeepAISE platform and consists of MongoDB and a Java based web service that is responsible for providing a REST API as well as the necessary security and auditing infrastructure. The data is organized as binary JSON documents, where each document represents the raw information (observations, measurements, etc.) and the DeepAISE predictions (scores, contributing factors etc.) about a specific patient, at a specific point in time. The organization of the data in this fashion allows us to make time series requests and query

for population trends in the past n hours as well as take deeper dives into an individual patients data.

- **Data Orchestrator Service:** This service directly interfaces with the healthcare IT system. It is responsible for fetching patient data (observations, measurements etc.), and transform them into the JSON document structure described above. It posts this data to the DeepAISE Model Service and then incorporates the returned results into the JSON document before posting them to the Results Database Service. It also maintains systems information, tracking uptime, as well as population characteristics that could be used to detect anomalies in data

- **User Interface Service:** This is the service that serves the user facing interfaces. It presents a dashboard that provides a ranked presentation of patients at risk for sepsis (see Fig.6, Under Observation column). Creating a card for each patient that concisely displayed their sepsis risk on the front and provided the top contributing factors on the cards opposite side made it easier for the clinical team to investigate impending cases of sepsis. A second drag-and-drop column allowed for clinicians to identify patients that have been reviewed and for whom no immediate action was deemed necessary (see Fig.6, Snoozed Alarms column). Situational awareness was further improved by the movement of septic patients to a third column (see Fig.6, Treatment Initiated column) indicating that a sepsis related treatment had been initiated.

References

[1] Vincent, J.-L. *et al.* The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure (1996).

[2] Costa, M., Goldberger, A. L. & Peng, C.-K. Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters* **89**, 068102 (2002).

[3] Nemati, S. *et al.* Respiration and heart rate complexity: effects of age and gender assessed by band-limited transfer entropy. *Respiratory Physiology & Neurobiology* **189**, 27–33 (2013).