Supplementary material

Supplementary material and methods

Preprocessing

Alignment to external databases (MIMIC-III and eICU) for validation:

MIMIC-III and eICU databases can be accessed via physio.net by a credentialed user. All data is de-identified according to HIPAA standards.

Feature Selection

A quality check was conducted for labels and frequency of features in the institutional and external datasets. We were able to extract 85 common features. All values were sorted by the most frequent occurrence and applied to our own dataset. The goal was to obtain as many overlapping features as possible. This was achieved by not only sorting for group labels but also for strings inside text strings. Exact extraction and alignment code is also available in the GitHub repository.

Alignment

Since MIMIC-III and eICU have at certain points a higher resolution of data, we took the mean of every hour of values to achieve the same resolution as the institutional dataset. Features like temperature or height were brought in a common SI unit. The distribution of all features was checked in advance to see if there were still certain outliers or unaligned data. Continuous medication differed a lot between European and US standards. Therefore, groups of medication substances were merged and achieved an overlying distribution of the medication given after normalization of each feature per dataset.

Labelling

Common labels were created through string detection and replacement.

Preprocessing of all databases (ICP-ICU, MIMIC-III, and eICU)

Outlier Detection

After selecting common features, we checked for outliers (99% percentile) and clipped the data with the maximum value of that patient's feature.

Outlier over Time

The ICU time course was aligned to hourly data. If more than one measurement was saved in that hour, we calculated the mean.

Normalization

We investigated three normalization approaches: Z-standardization, Yeo-Johnson transformation coupled with a Z-standardization, and min-max normalization. We created all datasets with these three normalization methods and created a common dataset with MICE (R Package MICE imputation) and checked through Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) for discrimination (Supp Fig. 1). This approach showed the least possible discrimination when normalized with Yeo-Johnson coupled with Z-standardization; therefore, this approach was selected for training purposes.

Imputation of Missing data

For the imputation of missing data, several methods were tried such as iterative imputation, mean imputation, median imputation, filling with zeros, and minus one. Iterative imputation performed a bit better than the others, but at the same time it introduced a strong bias and correlations into the input features. Therefore, we decided on using the median imputation. A continuous medication was assigned with 0 when not given.

Model architecture

Multi-layer perceptron (MLP) layers to feed into the LSTM cell and another MLP layer to read the information out of the LSTM cell to predict the likelihood of a short and long phase to occur in three hours. This multi-layered approach allows the extraction of higher-level features from raw data. These properties fulfil our aims of model robustness to deal with arbitrary sequence lengths and various degrees of data missingness. The concrete model architecture consists of an MLP that takes the input and feeds it into a hidden layer to extract features. These features are fed into 1-4 stacked LSTM layers. We apply layer normalization on the output of the final

LSTM, followed by a dropout layer. The output is fed into a two-layered MLP to predict whether a short or a long critical phase will appear within the next N hours.

Training

During training, we first applied Gaussian noise with a mean of 0 and a standard deviation up to 0.2 (equalling 20% of the standard deviation of the Z-standardized input) to simulate measurement errors and to make the network more robust towards perturbations applied later during the saliency calculation. The loss between the targets and the predictions is then calculated as follows:

1. Apply a mask where targets are NaN (where no ICP measure is available).
2. Apply binary cross-entropy loss to phase predictions per phase and per patient.
3. Weigh the loss by the inverse distribution of the targets (weigh timesteps with critical phases more than timesteps with non-critical phases).
4. Divide the loss per patient by the number of valid target points per patient such that every patient has the same influence on the final loss.
5. Average the loss over all patients.

The loss of the network is optimized by using the ADAM [30] optimizer with a learning rate which is reduced by a factor of 0.98 after every training epoch. We use a batch size of 16, train for a total of 32 epochs, and clip the gradient to 0.5. We stop the training early if the validation loss does not improve for 5 epochs.

Tuning

To find good hyperparameters for each model, we tuned several hyperparameters. We use the Tree Parzen Estimator of the Optuna library and prune unpromising trials with the MedianPruner (trials that perform worse than the median performance of all previous trials at a given time step are stopped). We tune for 128 trials per model.

When tuning the network, we optimized over the following options and ranges:

- Learning rate: 1e-8 - 1e-3
- Hidden layer size: [32, 64, 128, 256, 512]

- Dropout: 0.2 - 0.5
- Std of Gaussian input noise: 0.0 - 0.25
- Number of stacked LSTM layers: 1 - 4
- Gradient accumulation: 1 - 16 (This is effectively a multiplier of the batch size.)

The hyperparameters found after tuning for the models (1-10h and 24h) are shown in Suppl. Figure 2.

Feature Importance (Saliency)

IG works by taking the gradient of the model output towards the input for several discrete inputs and summing them. These inputs are interpolations on the path between the current patient and a baseline. This baseline is supposed to be an input for which the model generates zero saliency. We construct this neutral baseline by taking the median over each feature and repeating these median values for as many time steps (in the presented study one hour) as the current patient has. This IG procedure is augmented and smoothed by adding SG on top of it. This is done by repeating IG N times (in the presented study n=50), each time with newly generated Gaussian noise (std=0.01) added to the input. The N resulting input saliencies are then averaged.

For the feature ranking, all patients' saliencies are divided by the total absolute sum. This way, every patient, independent of the length of ICU stay, has the same influence on the final saliency ranking. All saliencies were grouped per hour, feature, before the sum was calculated. Finally, the grouped sum was divided by the count of the total amount of actual inputs the model hat at that time of a certain feature. This way less frequent features like laboratory values can play a greater role then one hot encoded feature like the diagnosis. The ranking shows the mean of that sum for every hour (first 360 hours). The heatmap was calculated the same way, only for visualization reasons we made a 99 percentile and 1 percentile clipping of the data. This is used to get a visual mean (white) around 0. We divided all saliencies again by the maximum value (for the distribution being between -1 and 1). Only the first 360 hours are shown in the heatmap in which every tile represents one hour. If the model took the input features on a more individual basis and was important but the role switched according to the surrounding input, feature importance values could be negative and positive inside the trajectory of one patient. The ranking ifs arranged according to the mean importance of long critical phases.

Hardware

Dual Socket Intel Xeon Workstation

2x8 cores a 4,4 GHz

364 GB RAM

2 NVIDIA Titan V

2x 2TB SSD

Ubuntu 18.04 LTS Operating System installed

Software

Integrated Care Manager (ICM Dräger®)

ICMiq – Tool to analyse the reporting database of ICM (Dräger®)

Soarian (Philips®)

RStudio

Adobe Illustrator

Word

Packages

R

- Tidyverse (for preprocessing and analysis) [33]
- yardstick (for analysis of predictions)
- gt (for creating tables)

Python

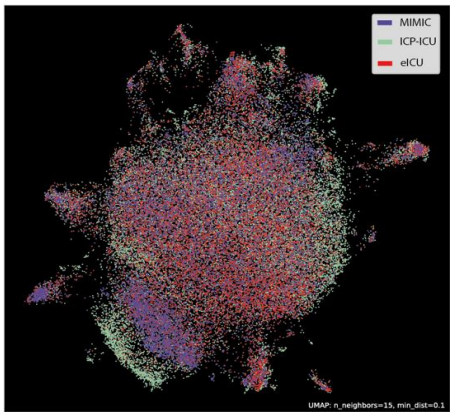
- Pytorch (for model construction)
- Pytorch-lightning (for model training)
- Sklearn
- Numpy [34]
- Captum (for saliencies) [35]
- Mlflow (for experiment logging)
- Optuna (for hyperparameter tuning)

Programming Languages

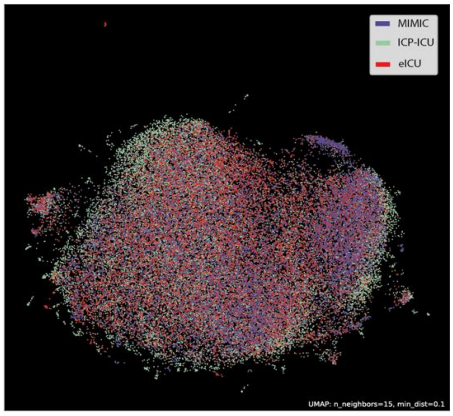
Python and R

Supplementary Figures

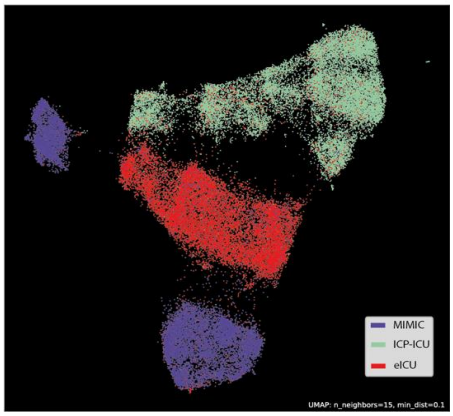
z-Norm



Yeo-Johnson

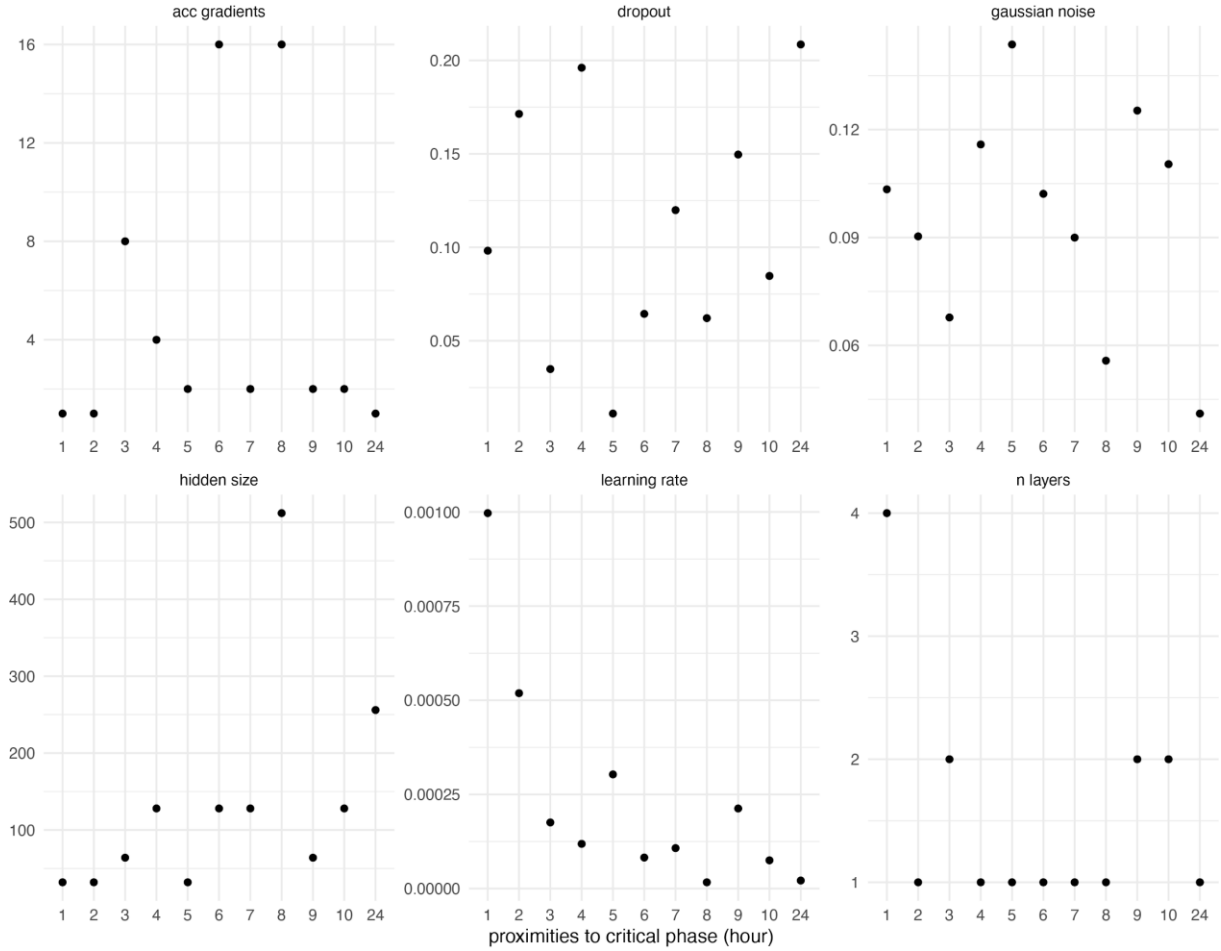


Min-Max



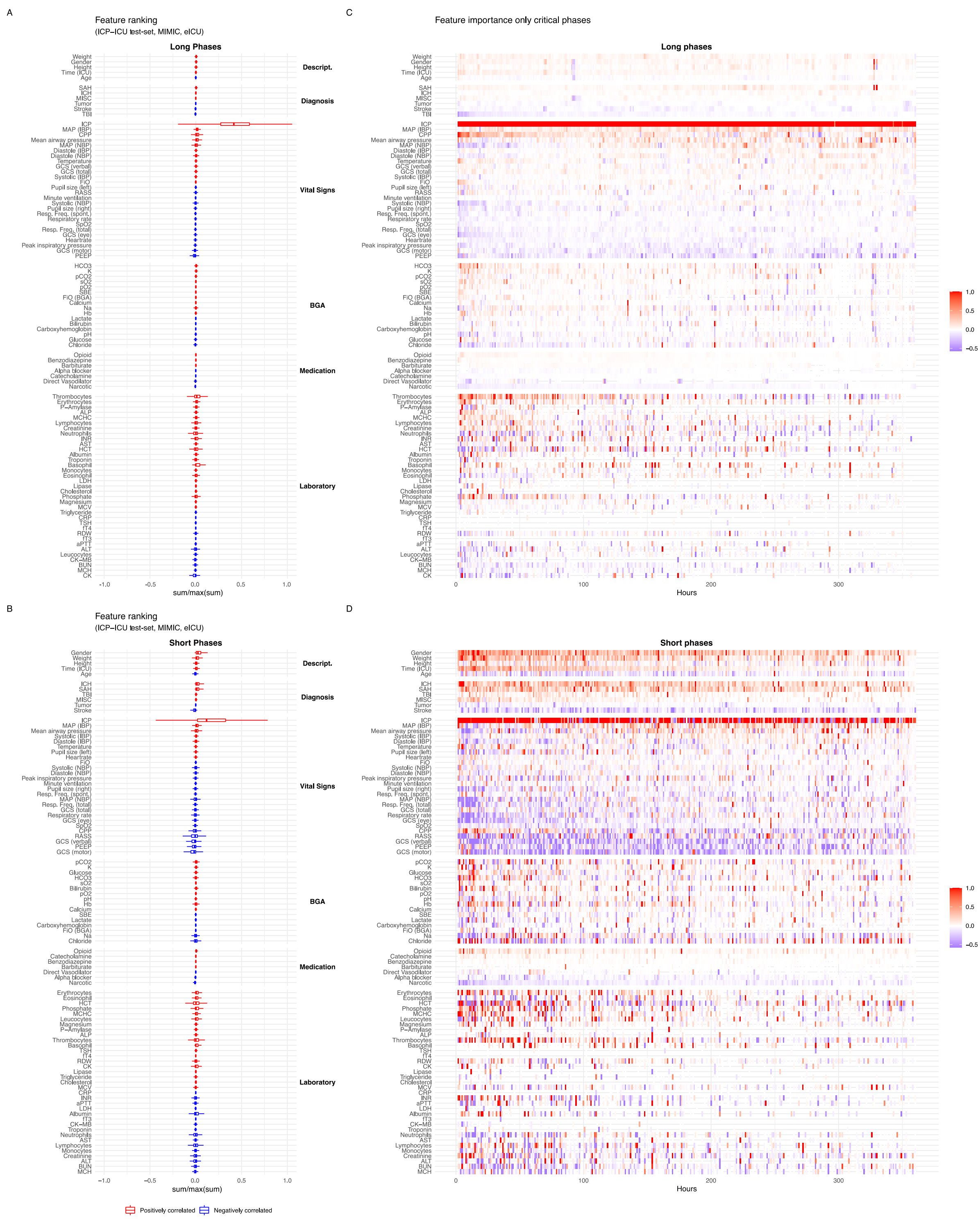
Supplementary Figure 1 UMAP of different normalization approaches. To visualize potential differences between datasets, we used dimensionality reduction (UMAP) to visualize a potential clusterable difference between data in our different normalizations. Yeo-Johnson Transformation followed by a Z-standardization of the data returned the least discrimination between datasets compared to Min-Max Normalization where the datasets could be detected by UMAP. We used Yeo-Johnson Transformation to train and test our model.

Supplementary Figure 2 Hyperparameter of models



Supplementary Figure 2 Hyperparameter of models. Performance of models predicting different proximities to critical phases. Five independent models on different splits on data (Cross validation) were done to train our models. In the tuning process of our models, we took the mean validation loss across all models. Concordant hyperparameters are depicted for each underlying model used in figure 2.

Supplementary Figure 4 Feature importance - Saliency



Supplementary Figure 4 Feature importance - Saliency. Reflecting the feature importance for the prediction of critical phases two hours in advance – a method (Integrated Gradients) was used to calculate gradient-based saliencies for every hour of input feature individually per patient. Boxplots of the sum of all saliencies per timesteps (hours) divided by their occurrence of input feature are depicted and listed in decreasing manner for long critical phases (A) and for short critical phases (B); The lower and upper hinges of the boxplots correspond to the first

and third quartiles (the 25th and 75th percentiles) the middle line to the median. The upper and the lower whisker extends from the hinge to the largest and smallest value no further than 1.5 * IQR from the hinge (A and B). The value can be negatively or positively correlated with the predicted target. If something is positively correlated, it is shown in red; if it is negatively correlated, it is shown in blue (A and B). This could be used intuitively in the clinic since red reflects a more negative and bluer a more positive influence on the predicted critical phase. A histogram of every hour of the input sequence is shown for the first 360 hours (each tile represents one hour) for long critical phases (C) and short critical phases (D). Patient descriptive and diagnosis are one hot encoded input features. All other features are down sampled to an hourly input and can be dynamic according to their nature. Medications (only continuous) are shown in groups of substances.

Supplementary Tables

Supplementary Table 1 ICP values across different groups

Group	ICP-ICU	MIMIC	eICU
Dataset			
Overall	13.0 (± 9.9) n= 562,666	10.3 (± 5.8) n= 357,617	8.2 (± 8.6) n= 485,801
Diagnosis			
TBI	12.5 (± 10.0) n= 208,666	13.8 (± 6.8) n= 86,781	11.3 (± 9.4) n= 90,147
ICH	13.6 (± 7.9) n= 125,902	9.3 (± 5.2) n= 57,038	8.2 (± 7.9) n= 117,808
Stroke	12.6 (± 8.8) n= 80,912	9.8 (± 5.8) n= 25,190	7.4 (± 6.9) n= 46,468
MISC	13.5 (± 11.2) n= 58,318	12.1 (± 7.3) n= 61,848	6.7 (± 10.1) n= 142,954
SAH	13.8 (± 12.9) n= 73,445	9.4 (± 4.9) n= 100,751	7.8 (± 7.7) n= 64,948
Tumour	13.3 (± 9 .7) n= 15,423	9.2 (± 4 .8) n= 26,009	7.9 (± 6.7) n= 23,476
Outcome			
Deceased	19.5 (±18.5) n= 86,863	10.8 (± 6.4) n= 48,358	10.3 (± 14.6) n= 46,518
Survived	11.6 (± 5.9) n= 475,803	10.2 (± 5.7) n= 309,259	7.9 (± 7.4) n= 439,283

Supplementary Table 2 - ICP and GCS values

Gruppe	ICP	GCS
Diagnosis		
TBI	12.4 (± 9.5)	4.0 (IQR: 6)
ICH	10.5 (± 7.9)	7.0 (IQR: 8)
Stroke	10.3 (± 8.2)	6.0 (IQR: 8)
MISC	8.9 (± 10.6)	10.0 (IQR: 10)
SAH	9.8 (± 8.4)	9.0 (IQR: 11)
Tumour	9.7 (± 7.5)	11.0 (IQR: 11)
Outcome		
Deceased	14.7 (± 15.8)	3.0 (IQR: 3)
Survived	9.9 (± 6.7)	8.0 (IQR: 9)
Dataset		
eICU	8.2 (± 8.6)	11.0 (IQR: 7)
MIMIC	10.3 (± 5.8)	10.0 (IQR: 7)
ICP-ICU	13.0 (± 9.9)	3.0 (IQR: 4)

Supplementary Table 3 - Comparison of models with different proximities to predict long critical phases
Area Under the Curve (AUC) of five independent models (cross-validation) and standard deviation (±)

Hour	ICP-ICU	MIMIC	eICU
Receiver Operating Characteristics - Long Phases			
1	0.982 (± 0.0008)	0.965 (± 0.0010)	0.941 (± 0.0025)
2	0.953 (± 0.0009)	0.948 (± 0.0025)	0.903 (± 0.0033)
3	0.936 (± 0.0025)	0.923 (± 0.0019)	0.872 (± 0.0031)
4	0.939 (± 0.0009)	0.916 (± 0.0020)	0.865 (± 0.0026)
5	0.933 (± 0.0040)	0.913 (± 0.0021)	0.865 (± 0.0017)
6	0.926 (± 0.0009)	0.884 (± 0.0046)	0.832 (± 0.0023)
7	0.918 (± 0.0016)	0.893 (± 0.0039)	0.837 (± 0.0021)
8	0.903 (± 0.0049)	0.866 (± 0.0071)	0.818 (± 0.0047)
9	0.914 (± 0.0036)	0.875 (± 0.0074)	0.823 (± 0.0051)
10	0.915 (± 0.0019)	0.863 (± 0.0045)	0.821 (± 0.0023)
24	0.826 (± 0.0071)	0.836 (± 0.0063)	0.779 (± 0.0046)
Precision Recall - Long Phases			
1	0.839 (± 0.0059)	0.553 (± 0.0831)	0.682 (± 0.0091)
2	0.707 (± 0.0067)	0.464 (± 0.0173)	0.563 (± 0.0148)
3	0.625 (± 0.0084)	0.393 (± 0.0053)	0.500 (± 0.0067)
4	0.653 (± 0.0045)	0.329 (± 0.0076)	0.471 (± 0.0043)
5	0.596 (± 0.0082)	0.320 (± 0.0083)	0.454 (± 0.0059)
6	0.592 (± 0.0026)	0.273 (± 0.0067)	0.415 (± 0.0052)
7	0.570 (± 0.0069)	0.244 (± 0.0203)	0.397 (± 0.0154)
8	0.488 (± 0.0085)	0.234 (± 0.0078)	0.376 (± 0.0059)
9	0.521 (± 0.0186)	0.223 (± 0.0398)	0.350 (± 0.0354)
10	0.532 (± 0.0052)	0.238 (± 0.0085)	0.378 (± 0.0040)
24	0.292 (± 0.0069)	0.177 (± 0.0049)	0.298 (± 0.0089)
Receiver Operating Characteristics - Short Phases			
1	0.800 (± 0.0028)	0.677 (± 0.0328)	0.707 (± 0.0069)
2	0.728 (± 0.0015)	0.801 (± 0.0048)	0.674 (± 0.0028)
3	0.712 (± 0.0028)	0.778 (± 0.0072)	0.666 (± 0.0033)
4	0.702 (± 0.0022)	0.789 (± 0.0073)	0.666 (± 0.0013)
5	0.695 (± 0.0029)	0.788 (± 0.0043)	0.666 (± 0.0042)
6	0.714 (± 0.0014)	0.785 (± 0.0044)	0.641 (± 0.0021)
7	0.709 (± 0.0030)	0.784 (± 0.0038)	0.654 (± 0.0031)
8	0.713 (± 0.0086)	0.780 (± 0.0062)	0.626 (± 0.0155)
9	0.698 (± 0.0034)	0.761 (± 0.0083)	0.651 (± 0.0034)
10	0.710 (± 0.0035)	0.781 (± 0.0060)	0.647 (± 0.0037)
24	0.680 (± 0.0050)	0.740 (± 0.0065)	0.633 (± 0.0055)
Precision Recall - Short Phases			
1	0.183 (± 0.0028)	0.100 (± 0.0040)	0.106 (± 0.0030)
2	0.107 (± 0.0007)	0.074 (± 0.0037)	0.096 (± 0.0003)
3	0.098 (± 0.0007)	0.065 (± 0.0039)	0.094 (± 0.0011)
4	0.083 (± 0.0014)	0.070 (± 0.0047)	0.092 (± 0.0005)
5	0.080 (± 0.0017)	0.070 (± 0.0029)	0.093 (± 0.0016)
6	0.086 (± 0.0019)	0.092 (± 0.0018)	0.081 (± 0.0007)
7	0.099 (± 0.0033)	0.071 (± 0.0030)	0.088 (± 0.0013)
8	0.097 (± 0.0037)	0.089 (± 0.0033)	0.077 (± 0.0030)
9	0.086 (± 0.0036)	0.063 (± 0.0047)	0.088 (± 0.0013)
10	0.092 (± 0.0032)	0.081 (± 0.0064)	0.087 (± 0.0013)
24	0.084 (± 0.0009)	0.073 (± 0.0006)	0.081 (± 0.0016)

Supplementary Table 4 – Model metrics with optimal threshold for the prediction of long critical phases
Values depicted as percent and confidence interval (CI)

Datasets	Distance target (h)	Accuracy	Sensitivity	Specificity	PPV	NPV
eICU	1	0.871 (CI: 0.866-0.876)	0.876 (CI: 0.870-0.881)	0.866 (CI: 0.852-0.880)	0.288 (CI: 0.266-0.309)	0.991 (CI: 0.991-0.992)
ICP-ICU	1	0.931 (CI: 0.929-0.933)	0.946 (CI: 0.939-0.953)	0.916 (CI: 0.907-0.925)	0.380 (CI: 0.357-0.403)	0.997 (CI: 0.996-0.997)
MIMIC	1	0.907 (CI: 0.901-0.912)	0.924 (CI: 0.910-0.938)	0.889 (CI: 0.874-0.905)	0.199 (CI: 0.178-0.220)	0.997 (CI: 0.997-0.998)
eICU	2	0.828 (CI: 0.824-0.833)	0.807 (CI: 0.792-0.822)	0.850 (CI: 0.836-0.863)	0.249 (CI: 0.235-0.263)	0.986 (CI: 0.985-0.987)
ICP-ICU	2	0.890 (CI: 0.888-0.892)	0.890 (CI: 0.875-0.905)	0.890 (CI: 0.873-0.908)	0.307 (CI: 0.277-0.336)	0.993 (CI: 0.993-0.994)
MIMIC	2	0.881 (CI: 0.876-0.886)	0.892 (CI: 0.887-0.898)	0.869 (CI: 0.861-0.878)	0.168 (CI: 0.159-0.177)	0.996 (CI: 0.996-0.997)
eICU	3	0.800 (CI: 0.796-0.804)	0.757 (CI: 0.747-0.767)	0.843 (CI: 0.837-0.849)	0.229 (CI: 0.223-0.234)	0.983 (CI: 0.982-0.983)
ICP-ICU	3	0.865 (CI: 0.860-0.870)	0.860 (CI: 0.832-0.888)	0.870 (CI: 0.844-0.896)	0.263 (CI: 0.230-0.296)	0.992 (CI: 0.990-0.993)
MIMIC	3	0.849 (CI: 0.846-0.851)	0.843 (CI: 0.830-0.857)	0.854 (CI: 0.840-0.868)	0.145 (CI: 0.136-0.155)	0.995 (CI: 0.994-0.995)
eICU	4	0.793 (CI: 0.788-0.798)	0.765 (CI: 0.755-0.774)	0.821 (CI: 0.805-0.836)	0.208 (CI: 0.195-0.220)	0.983 (CI: 0.982-0.983)
ICP-ICU	4	0.868 (CI: 0.866-0.870)	0.872 (CI: 0.859-0.885)	0.865 (CI: 0.853-0.877)	0.281 (CI: 0.265-0.297)	0.991 (CI: 0.990-0.992)
MIMIC	4	0.840 (CI: 0.836-0.844)	0.844 (CI: 0.829-0.858)	0.836 (CI: 0.821-0.852)	0.132 (CI: 0.123-0.141)	0.995 (CI: 0.994-0.995)
eICU	5	0.788 (CI: 0.786-0.791)	0.748 (CI: 0.734-0.762)	0.828 (CI: 0.819-0.837)	0.211 (CI: 0.205-0.217)	0.982 (CI: 0.981-0.983)
ICP-ICU	5	0.862 (CI: 0.855-0.868)	0.844 (CI: 0.826-0.862)	0.879 (CI: 0.861-0.898)	0.254 (CI: 0.228-0.279)	0.992 (CI: 0.991-0.992)
MIMIC	5	0.838 (CI: 0.834-0.842)	0.839 (CI: 0.825-0.854)	0.837 (CI: 0.824-0.851)	0.132 (CI: 0.124-0.139)	0.994 (CI: 0.994-0.995)
eICU	6	0.759 (CI: 0.756-0.762)	0.718 (CI: 0.709-0.727)	0.800 (CI: 0.794-0.806)	0.180 (CI: 0.177-0.184)	0.979 (CI: 0.978-0.979)
ICP-ICU	6	0.851 (CI: 0.848-0.854)	0.868 (CI: 0.859-0.876)	0.834 (CI: 0.824-0.845)	0.233 (CI: 0.223-0.243)	0.991 (CI: 0.990-0.991)
MIMIC	6	0.806 (CI: 0.800-0.811)	0.796 (CI: 0.787-0.804)	0.815 (CI: 0.806-0.825)	0.112 (CI: 0.107-0.117)	0.993 (CI: 0.992-0.993)
eICU	7	0.766 (CI: 0.762-0.771)	0.729 (CI: 0.714-0.744)	0.804 (CI: 0.787-0.822)	0.186 (CI: 0.175-0.197)	0.980 (CI: 0.979-0.981)
ICP-ICU	7	0.839 (CI: 0.836-0.842)	0.853 (CI: 0.834-0.873)	0.824 (CI: 0.809-0.840)	0.233 (CI: 0.221-0.245)	0.989 (CI: 0.988-0.990)
MIMIC	7	0.814 (CI: 0.811-0.818)	0.819 (CI: 0.800-0.838)	0.810 (CI: 0.786-0.833)	0.112 (CI: 0.102-0.122)	0.994 (CI: 0.993-0.994)
eICU	8	0.748 (CI: 0.742-0.754)	0.702 (CI: 0.686-0.717)	0.794 (CI: 0.769-0.818)	0.173 (CI: 0.159-0.187)	0.978 (CI: 0.977-0.978)
ICP-ICU	8	0.827 (CI: 0.821-0.833)	0.810 (CI: 0.781-0.840)	0.844 (CI: 0.823-0.865)	0.206 (CI: 0.189-0.222)	0.989 (CI: 0.988-0.990)
MIMIC	8	0.789 (CI: 0.781-0.798)	0.761 (CI: 0.740-0.783)	0.818 (CI: 0.811-0.824)	0.108 (CI: 0.106-0.110)	0.992 (CI: 0.991-0.992)
eICU	9	0.752 (CI: 0.748-0.757)	0.699 (CI: 0.678-0.719)	0.806 (CI: 0.786-0.825)	0.180 (CI: 0.169-0.192)	0.978 (CI: 0.977-0.979)
ICP-ICU	9	0.841 (CI: 0.837-0.846)	0.814 (CI: 0.796-0.832)	0.869 (CI: 0.849-0.889)	0.250 (CI: 0.225-0.275)	0.989 (CI: 0.988-0.990)
MIMIC	9	0.793 (CI: 0.786-0.800)	0.810 (CI: 0.770-0.849)	0.777 (CI: 0.735-0.818)	0.095 (CI: 0.084-0.107)	0.993 (CI: 0.992-0.994)
eICU	10	0.753 (CI: 0.751-0.755)	0.691 (CI: 0.667-0.716)	0.815 (CI: 0.788-0.841)	0.186 (CI: 0.171-0.201)	0.977 (CI: 0.976-0.979)
ICP-ICU	10	0.835 (CI: 0.832-0.838)	0.813 (CI: 0.796-0.829)	0.858 (CI: 0.838-0.878)	0.226 (CI: 0.205-0.247)	0.989 (CI: 0.988-0.990)
MIMIC	10	0.790 (CI: 0.785-0.795)	0.763 (CI: 0.730-0.796)	0.817 (CI: 0.793-0.842)	0.107 (CI: 0.098-0.116)	0.992 (CI: 0.991-0.993)
eICU	11	0.729 (CI: 0.712-0.746)	0.685 (CI: 0.666-0.704)	0.773 (CI: 0.744-0.803)	0.156 (CI: 0.139-0.173)	0.976 (CI: 0.974-0.977)
ICP-ICU	11	0.799 (CI: 0.783-0.816)	0.789 (CI: 0.745-0.832)	0.810 (CI: 0.794-0.826)	0.168 (CI: 0.160-0.176)	0.987 (CI: 0.985-0.990)
MIMIC	11	0.769 (CI: 0.750-0.789)	0.741 (CI: 0.712-0.769)	0.798 (CI: 0.784-0.811)	0.094 (CI: 0.086-0.103)	0.991 (CI: 0.990-0.992)

Supplementary Table 4 – Model metrics with optimal threshold for the prediction of long critical phases
Values depicted as percent and confidence interval (CI)

Datasets	Distance target (h)	Accuracy	Sensitivity	Specificity	PPV	NPV
eICU	12	0.686 (CI: 0.628-0.744)	0.605 (CI: 0.518-0.691)	0.767 (CI: 0.737-0.797)	0.138 (CI: 0.108-0.167)	0.970 (CI: 0.962-0.977)
ICP-ICU	12	0.771 (CI: 0.717-0.824)	0.715 (CI: 0.644-0.785)	0.827 (CI: 0.780-0.873)	0.193 (CI: 0.144-0.241)	0.981 (CI: 0.975-0.987)
MIMIC	12	0.707 (CI: 0.616-0.799)	0.707 (CI: 0.653-0.760)	0.708 (CI: 0.578-0.838)	0.069 (CI: 0.046-0.092)	0.988 (CI: 0.983-0.993)
eICU	24	0.712 (CI: 0.706-0.719)	0.650 (CI: 0.623-0.676)	0.775 (CI: 0.747-0.803)	0.148 (CI: 0.137-0.160)	0.974 (CI: 0.972-0.975)
ICP-ICU	24	0.751 (CI: 0.744-0.758)	0.711 (CI: 0.675-0.747)	0.791 (CI: 0.766-0.815)	0.122 (CI: 0.114-0.131)	0.985 (CI: 0.984-0.987)
MIMIC	24	0.762 (CI: 0.756-0.767)	0.725 (CI: 0.681-0.769)	0.799 (CI: 0.758-0.839)	0.088 (CI: 0.076-0.100)	0.991 (CI: 0.990-0.992)

Supplementary Table 5 Predicting critical phases two hours in advance
Area Under the Curve (AUC) and ± Standard Deviation

	Receiver Operating Characteristics	Precision Recall
Datasets		
eICU	0.903 (± 0.0033)	0.563 (± 0.0148)
ICP-ICU	0.953 (± 0.0009)	0.707 (± 0.0067)
MIMIC	0.948 (± 0.0025)	0.464 (± 0.0173)
Outcome		
Deceased	0.938 (± 0.0016)	0.730 (± 0.0054)
Survived	0.907 (± 0.0033)	0.456 (± 0.0121)
Diagnosis		
TBI	0.918 (± 0.0018)	0.633 (± 0.0050)
ICH	0.914 (± 0.0027)	0.522 (± 0.0132)
SAH	0.914 (± 0.0011)	0.400 (± 0.0069)
Stroke	0.931 (± 0.0024)	0.580 (± 0.0130)
Tumour	0.898 (± 0.0073)	0.392 (± 0.0159)
MISC	0.921 (± 0.0051)	0.586 (± 0.0159)
Sex		
Male	0.923 (± 0.0028)	0.596 (± 0.0101)
Female	0.911 (± 0.0037)	0.473 (± 0.0136)
Missing data		
less missing	0.928 (± 0.0017)	0.591 (± 0.0085)
more missing	0.899 (± 0.0045)	0.472 (± 0.0140)
Weeks		
1. Week	0.922 (± 0.0024)	0.593 (± 0.0080)
2. Week	0.912 (± 0.0039)	0.490 (± 0.0172)
3. Week	0.909 (± 0.0046)	0.371 (± 0.0228)