On the Importance of Step-wise Embeddings for Heterogeneous Clinical Time-Series

Rita Kuznetsova[1*] MKUZNETSOVA@ETHZ.CH
Alizée Pace[1,2,3*] ALPACE@ETHZ.CH
Manuel Burger[1*] BURGERM@ETHZ.CH
Hugo Yèche[1] HYECHE@ETHZ.CH
Gunnar Rätsch[1,2,4,5,6] RAETSCH@INF.ETHZ.CH
[1] *Department of Computer Science, ETH Zürich, Switzerland*
[2] *ETH AI Center, ETH Zürich, Switzerland*
[3] *Max Planck Institute for Intelligent Systems, Tübingen, Germany*
[4] *Medical Informatics Unit, Zürich University Hospital, Zürich, Switzerland*
[5] *Swiss Institute of Bioinformatics, Zurich, Switzerland*
[6] *Department of Biology, ETH Zürich, Zürich, Switzerland*

Abstract

Recent advances in deep learning architectures for sequence modeling have not fully transferred to tasks handling time-series from electronic health records. In particular, in problems related to the Intensive Care Unit (ICU), the state-of-the-art remains to tackle sequence classification in a tabular manner with tree-based methods. Recent findings in deep learning for tabular data are now surpassing these classical methods by better handling the severe heterogeneity of data input features. Given the similar level of feature heterogeneity exhibited by ICU time-series and motivated by these findings, we explore these novel methods' impact on clinical sequence modeling tasks. By jointly using such advances in deep learning for tabular data, our primary objective is to underscore the importance of step-wise embeddings in time-series modeling, which remain unexplored in machine learning methods for clinical data. On a variety of clinically relevant tasks from two large-scale ICU datasets, MIMIC-III and HiRID, our work provides an exhaustive analysis of state-of-the-art methods for tabular time-series as time-step embedding models, showing overall performance improvement. In particular, we evidence the importance of feature grouping in clinical time-series, with significant performance gains when considering features within predefined semantic groups in the step-wise embedding module.

Keywords: Deep Learning, Healthcare, Time-Series, Step-wise Embeddings, Feature Groups.

1. Introduction

Recent years have seen the development of deep learning architectures for Electronic Health Records (EHRs), which explore machine learning solutions for various clinical prediction tasks such as organ failure prediction (Hyland et al., 2020; Tomašev et al., 2019), treatment effect estimation (Bica et al., 2020) or prognostic modeling (Choi et al., 2016b). Most work in this area primarily focuses on either modifying the backbone sequence model (Horn et al., 2020; Xu et al., 2018) or investigating modifications to the training objective (Yèche et al., 2021, 2022; Cheng et al., 2023). Still, the performance gap between proposed deep learning methods and tree-based approaches remains significant (Yèche et al., 2021; Hyland et al., 2020).

Recent work for early prediction of acute kidney injury using sparse multivariate time-series (Tomašev et al., 2021) shows that enhancing the time-step embedding neural network architectures, i.e simple replacement of linear layer to neural network for the input feature space preprocessing, yields significant performance gain. Concurrently, the state-of-the-art on tabular data, which relied on boosted tree methods (Ke et al., 2017; Chen and Guestrin, 2016; Freund et al., 1999), has been surpassed by recent development in the field of deep learning (Gorishniy

[*] These authors contributed equally

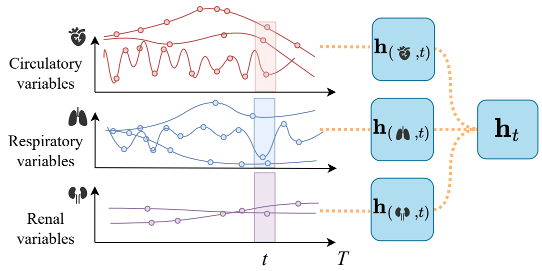


Figure 1: **Schematic time-step embedding architecture**: features interact within predefined semantic groups before being aggregated into time-step embeddings.

et al., 2021, 2022). Despite these observations, recent research in EHRs methods predominantly showcases the development of more powerful backbone sequence models, rather than investigating the influence of step-wise embedding modules. If some approaches have used feature embeddings, with their primary focus being on evaluating the effect of self-supervised pre-training (Tipirneni and Reddy, 2022), a comprehensive evaluation of how feature embedding efforts influence downstream performance is yet to be extensively studied.

Motivated by these observations, our main objective is to showcase the **significance of embedding architectures in clinical time-series analysis**. To achieve this, we conduct an extensive evaluation and comparison of various embedding architectures specifically designed for tabular data, with a focus not on optimizing the backbone sequence model, but rather on optimizing the step-wise embedding module. We find that we obtain timestep representations that serve as an expressive input to downstream sequence models – which boosts the overall performance of deep learning methods on clinical time-series data. Our work is thus orthogonal and complementary to the design of backbone architectures (Horn et al., 2020) or of loss functions for supervised (Yèche et al., 2022) and unsupervised learning (Yèche et al., 2021).

Second, our study demonstrates the importance of **feature groupings** (Imrie et al., 2022; Masoomi et al., 2020) in clinical time-series. In the medical field, it is common to not consider measurement interactions individually but through predefined semantic groups of features (Kelly and Semsarian, 2009; Meira et al., 2001). EHR data consist of multivariate time-series exhibiting such heterogeneity, with variables collected from different data sources and relating to different organ systems. These structures, determined by prior clinical knowledge, delineate feature groups tied to medical concepts or modalities, such as measurement types or organs, which we incorporate into embedding modules. Results demonstrate considering features *in the context of their semantic modality* to improve performance. We illustrate the optimal embedding pipeline uncovered by our work in Figure 1: features interact within groups before being aggregated into time-step embeddings and input to a sequential deep learning module for end-to-end training. This scheme additionally enables the interpretability of model results at a semantic group level. Thus, we also explore how disentangling medical concepts could enhance the interpretability of the model's decision-making.

Contributions The main contributions of this paper are the following: (1) First, we provide an extensive benchmark of embedding architectures for clinical prediction tasks. To the best of our knowledge, no prior work has considered applying the developments from the tabular deep learning literature to the heterogeneous time-series nature of clinical data. (2) Our exhaustive analysis allows us to draw important conclusions that semantically grouping features, especially related to organ systems, greatly enhance prediction performance. (3) Finally, combining these insights, our systematic study sets a new state-of-the-art performance on different established clinical tasks and datasets.

2. Related work

Time-series feature embedding Despite developments in model architectures for supervised clinical time-series tasks (Horn et al., 2020; Zhang et al., 2021), deep learning methods still show performance limitations on the highly heterogeneous, sparse time-series nature of intensive care unit data (Yèche et al., 2021; Hyland et al., 2020). Recent work has, however, demonstrated promising improvements with the introduction of feature embedding layers before the sequence model, together with auxiliary objectives (Tomašev et al., 2021, 2019). This research mirrors recent progress in the field of deep learning for tabular data (Gorishniy et al., 2021, 2022), which significantly outperforms state-of-the-art methods by combining transformer-based approaches with embeddings of tabular data rows. We note that a separate line of research explores self-supervised pretraining methodologies for both clinical time-series representation learning (Tipirneni and Reddy, 2022; Labach et al., 2023)and tabular deep learning (Yin et al., 2020; Huang et al., 2020; Kossen et al., 2021; Somepalli et al., 2022). While we focus on end-to-end supervised training in the present benchmark, we note that this constitutes a promising avenue for further work in clinical time-series feature embedding.

Feature groups within embeddings Recent work on tabular data embeddings highlight the importance of handling features of categorical or numerical types through distinct architectures (Huang et al., 2020; Arik and Pfister, 2021). This motivates our benchmarking study on incorporating additional feature structures, such as measurement or organ type, within the embedding layers. Most research on EHR data modeling focuses on extracting temporal trends (Luo et al., 2016; Ding and Luo, 2021) for patient phenotyping (Aguiar et al., 2022; Qin et al., 2023) from entire time-series. To the best of our knowledge, this work is the first attempt to consider and demonstrate the impact of global feature groupings at a time-step level on prediction performance.

We refer the reader to Appendix D for further discussion of related work.

3. Method

We summarize the overall deep learning pipeline benchmarked in this work in Figure 2, followed by an in-depth explanation of each component in this section.

3.1. Notation

We define a patient stay in the intensive care unit as a multivariate time-series $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$, where T is the length of a given stay. Each time-step is $\mathbf{x}_t = [x_{(1,t)}, \dots, x_{(d,t)}] \in \mathbb{R}^d$. Depending on the specific task, for each patient stay $\mathbf{X}$ we have either an associated label vector $\mathbf{y} \in \mathbb{R}^T$ (per each time-step) or a single label $\mathbf{y} \in \mathbb{R}$ that corresponds to the entire patient stay – see Section 4 for an overview of studied tasks and datasets.

We consider step-wise embedding architectures under two scenarios: first, as a function applied to the entire feature space $\mathbf{x}_t$, referred to as *direct* (D); second, we propose to apply them separately to distinct feature groups, referred to as *feature grouping* (G). In the latter case, there are several ways to group the d observed variables for each time-step $\mathbf{x}_t$ based on their assignment to a particular medical concept from a set of K concepts $\{\mathcal{M}_1, \dots, \mathcal{M}_K\}$, such that all variables are assigned to a single concept: $\bigcup_{k=1}^K \mathcal{M}_k = \{1, \dots, d\}$ and $\forall k \neq k^{'} : \mathcal{M}_{k'} \cap \mathcal{M}_k = \emptyset$. In the context of ICU-related tasks, we define the splitting of features into concept groups by leveraging the prior knowledge: *organ*, *measurement type* (laboratory values, observations, treatments, etc.) and *variable type* as shown by Tomašev et al. (2019). The exact groups are provided in Appendix B.1. We group the features on a time-step level t, and we denote the subset of features belonging to the concept $\mathcal{M}_k$ as $\mathbf{x}_{(\mathcal{M}_k,t)}$. For each k, we learn a representation of $\mathbf{h}_{(\mathcal{M}_k,t)}$ that we refer to as *concept embedding*.

Definition 1 *Let f_{θ_k} denote the embedding model for concept $\mathcal{M}_k$, parameterized by θ_k, taking as input the subset of features $\mathbf{x}_{(\mathcal{M}_k,t)}$ and with output $\mathbf{h}_{(\mathcal{M}_k,t)} = f_{\theta_k}(\mathbf{x}_{(\mathcal{M}_k,t)})$. We term the latent representation $\mathbf{h}_{(\mathcal{M}_k,t)}$ as the* concept embedding.

Definition 2 *The* time-step embedding $\mathbf{h}_t$ *is a representation of all input features $\mathbf{x}_t$ at a given time-step. This embedding can be obtained through two approaches, as illustrated in Figure 2:*

(D) In the first (direct) scenario, $\mathbf{h}_t = f_\theta(\mathbf{x}_t)$, where f_θ is an embedding model parameterized by θ, processing the entire set of features for each time-step t.

(G) In the second (feature grouping) scenario, $\mathbf{h}_t = g_\psi\left[\left\{\mathbf{h}_{(\mathcal{M}_k,t)}\right\}_{k=1}^K\right]$, where g_ψ is an aggregation function applied to the K concept embeddings of each feature group.

The resulting time-step embedding $\mathbf{h}_t$ is subsequently passed as input to the sequential backbone. In the following, we discuss design choices for feature encoder architectures (f_θ and f_{θ_k}) and for the aggregation function g_ψ.

3.2. Direct time-step embedding

As first candidates, following (Gorishniy et al., 2021; Grinsztajn et al., 2022), we consider MLP and ResNet architectures as feature encoders. These are well-studied deep learning models, whose impact on step-wise feature preprocessing remains unexplored in the context of clinical sequence modeling.

We also consider a more advanced architecture borrowed from deep learning for tabular data, the Feature Tokenizer Transformer (FTT (Gorishniy et al.,

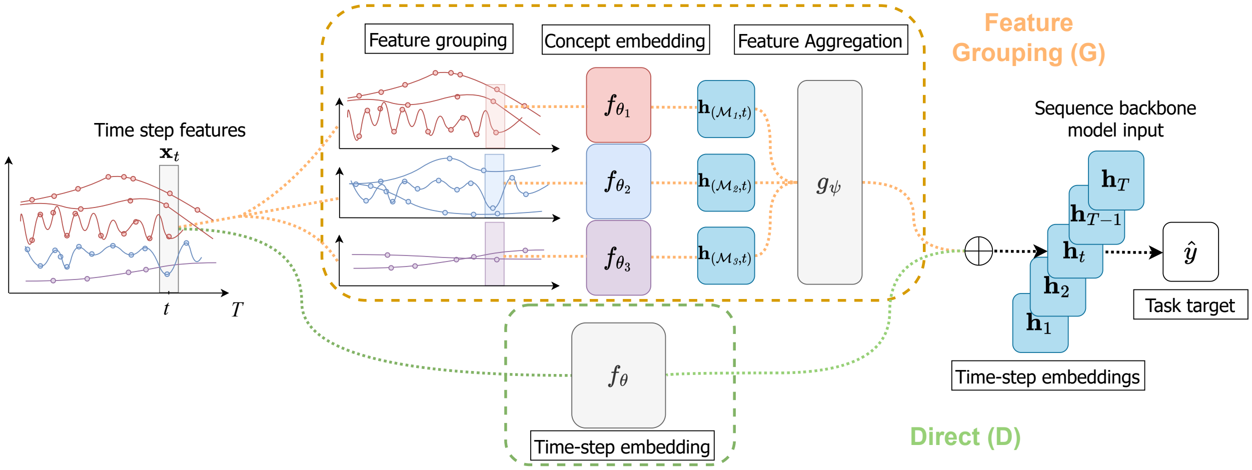


Figure 2: **Pipeline overview.** The entire set of features $\mathbf{x}_t$ for time-step t is: (D) preprocessed to directly form a time-step embedding $\mathbf{h}_t$ (green line); (G) grouped to form concept embeddings $\mathbf{h}_{(\mathcal{M}_k,t)}$, which are aggregated to create a final time-step embedding $\mathbf{h}_t$ (yellow line). The resulting time-step embeddings are then passed to the backbone model. The whole pipeline is trained in an end-to-end fashion to predict the task target $\hat{y}$.

2021)). This complex encoder consists of two distinct modules. First, the *Feature Tokenizer* (FT) embeds individual features $x_{(j,t)} \in \mathbb{R}$ in timestep vector $\mathbf{x}_t$ to high-dimensional continuous variables $\mathbf{e}_{(j,t)} \in \mathbb{R}^m$. This module is linear, parametrized by $\mathbf{W} \in \mathbb{R}^{d\times m}$, such that $\mathbf{e}_{(j,t)} = \mathbf{x}_t^T \mathbf{W}_j$. The final output of the FT module is a matrix $\mathbf{e}_t = \texttt{stack}[\mathbf{e}_{1,t}, \ldots, \mathbf{e}_{d,t}] \in \mathbb{R}^{d\times m}$. Next, the *Transformer* (T) module learns a unique time-step embedding $\mathbf{h}_t$ from matrix $\mathbf{e}_t$, by applying a transformer (Vaswani et al., 2017) along the d dimension. More specifically, to obtain a global representation, $\mathbf{h}_t$ is the output from a "classification token" [CLS] (Devlin et al., 2018) which is concatenated to the input $\mathbf{e}_t$.

We do not consider unsupervised methods such as factor analysis, standard auto-encoders, and variational auto-encoders for the embedding module design, given reports of them not demonstrating significant performance benefits for ICU data feature embeddings (Tomašev et al., 2019). Compared to MLP and ResNet, which consider features equally, FTT, through this two-stage modeling, should handle feature heterogeneity more efficiently, a crucial consideration in the context of ICU data.

3.3. Feature aggregation

As introduced in Section 3.1, in scenario (G), our aim is to explore the impact of embedding distinct groups of features independently. There, we simply use the same architecture for K concept embedding models, each with its own set of parameters θ_k as in Definition 1.

In terms of aggregation function g_ψ, designed to combine concepts $\mathbf{h}_{(\mathcal{M}_k,t)}$ into an overall timestep embedding $\mathbf{h}_t$, we consider the choices: mean (or sum) pooling, concatenation [1], and attention-based pooling. The latter option additionally offers interpretability of concept-level interactions through attention weight analysis, as discussed in Section 5.

3.4. Training

The entire set of features $\mathbf{x}_t$ for time-step t is preprocessed as shown in Figure 2. The resulting time-step embeddings for each for each t are subsequently fed into the sequential backbone model, which is trained in a supervised manner for the final task's

[1] The concatenation function is not, by itself, an aggregation function. It also presents scalability issues and lacks permutation invariance. Nevertheless, we include it in our study for experimental purposes.

target prediction $\hat{y}$. Consistent with previous approaches(Tomašev et al., 2019, 2021; Gorishniy et al., 2021), no specific loss for the embeddings was factored in. The primary objective of this study is to demonstrate that a simple step-wise module integrated in standard end-to-end supervised training pipeline can produce significant performance improvements.

4. Experimental setup

To ensure reproducibility we share our code.[1]

Clinical prediction tasks We demonstrate the effectiveness of our embedding methods for electronic health records by studying their effect on prediction performance for different clinical tasks related to intensive care. Our method and related baselines are benchmarked on the online binary prediction task of (1) circulatory and (2) respiratory failure within the next 12 hours, (3) remaining length of stay and on prediction of (4) patient mortality at 24 hours after admission, as well as (5) patient phenotyping after 24 hours. Tasks (1-5), as defined in HiRID-ICU-Benchmark (Yèche et al., 2021), are based on the publicly available HiRID dataset (Faltys et al., 2021; Hyland et al., 2020). We also consider the task of continuously predicting mortality *within* 24 hours, throughout the patient stay – also known as (6) decompensation, (7) patient mortality at 48 hours after admission and (8) remaining length of stay. We study the latter three task on the well-known MIMIC-III dataset (Johnson et al., 2016). Further details on the definition of each task can be found in benchmark papers which introduced them (Harutyunyan et al., 2019; Yèche et al., 2021). Further details on task definition and data pre-processing are provided in Appendix A.

Success metrics Our primary success metric for the usefulness of our method is *performance on downstream clinical tasks*. As these often consist of significantly imbalanced classification problems (Yèche et al., 2021), performance is measured through the area under the precision-recall curve (AUPRC), the area under the receiver operating characteristic curve (AUROC), and balanced accuracy (Bal. Acc.). For regression problems we report mean absolute error (MAE) in hours. This follows established practice on clinical early prediction tasks (Yèche et al., 2021; Harutyunyan et al., 2019).

Benchmarked methods We evaluate different embedding architectures including linear mapping and Feature Encoders, as referenced in Section 3.2. We also compare these to deep learning models that do not use an embedding layer, where a sequential model gets the raw feature vector at each time-step. Additionally, we consider a Gradient Boosted Tree method using LightGBM (Ke et al., 2017), based on manually-extracted features (Yèche et al., 2021). Downstream, we use deep learning backbones and optimized hyperparameters for our specific prediction tasks, as per prior research (Yèche et al., 2021; Harutyunyan et al., 2019). We use a Gated Recurrent Unit (GRU) (Cho et al., 2014) network for circulatory failure prediction and a Transformer (Vaswani et al., 2017) for all other tasks. That architectural choice for each task is based on previously published papers (Yèche et al., 2022; Yèche et al., 2021) Further implementation details are provided in Appendix B.

5. Results

In this section, we provide results for the proposed benchmarking study, systematically evaluating the performance of different embedding modules for EHR modeling. We validate the following hypotheses: (1) Relying on deep learning for tabular data methods in time-step embeddings can significantly improve the performance of deep learning models for clinical time-series. (2) Specifically, via a comprehensive examination of time-step encoder components, we demonstrate that relying on the FTT approach coupled with feature grouping and appropriate aggregation tends to yield the best overall performance. (3) Attention-based embedding architectures help us gain interpretability on the feature and medical concept level into deep learning models for tabular time-series, which remains largely unexplored in the relevant literature. (4) Models based on strong clinical priors such as feature assignments to organ systems, show superior performance.

Advancing deep learning approaches with step-wise embedding First, we present the experimental results that demonstrate the performance improvement achieved through well-designed embedding methods in deep-learning models for clinical time-series predictions. Despite deep learning models(Cho et al., 2014; Vaswani et al., 2017) often falling

[1] `https://github.com/ratschlab/clinical-embeddings`

Table 1: **Performance benchmark for different embedding architectures,** measured through the Area under the Precision-Recall Curve (AUPRC) or Mean Absolute Error (MAE) in hours. Mean and standard deviation are reported over five training runs. Best and overlapping results are highlighted in bold. *Reference Benchmark* results are best as reported by Yèche et al. (2021) and Harutyunyan et al. (2019) (We train LightGBM, Transformer and Temporal Convolutional Network (TCN) on MIMIC-III for comparison). *Step-wise encoders* are based on prior work (linear Yèche et al. (2021), MLP Tomašev et al. (2019), ResNet Tomašev et al. (2019), and FTT Gorishniy et al. (2021)) and our proposed concept groups. The *backbone* baseline considers the raw input feature vector at each time-step without any embedding layer.

Dataset	HiRID					MIMIC-III		
Clinical pred. task	Circ. Fail.	Resp. Fail.	Mort.	LoS	Pheno.	Decomp.	Mort.	LoS
Metric	*AuPRC* ↑	*AuPRC* ↑	*AuPRC* ↑	*MAE* ↓	*Bal. Acc.* ↑	*AuPRC* ↑	*AuPRC* ↑	*MAE* ↓
Reference Benchmarks								
LSTM	32.6 ± 0.8	56.9 ± 0.3	60.0 ± 0.9	60.7 ± 1.6	39.5 ± 1.2	34.4 ± 0.1	48.5 ± 0.3	123.1 ± 0.2
Transformer	35.2 ± 0.6	59.4 ± 0.3	61.0 ± 0.8	59.5 ± 2.8	42.7 ± 1.4	34.3 ± 0.7	**53.3** ± 0.4	98.3 ± 0.7
TCN	35.8 ± 0.6	58.9 ± 0.3	60.2 ± 1.1	59.8 ± 2.8	41.6 ± 2.3	36.6 ± 0.3	51.8 ± 0.6	97.8 ± 0.9
LightGBM	38.8 ± 0.2	**60.4** ± 0.2	**62.6** ± 0.1	57.0 ± 0.3	45.8 ± 2.0	37.1 ± 0.3	48.2 ± 0.4	99.7 ± 0.1
Step-wise Encoders								
Backbone	36.6 ± 0.5	59.5 ± 0.4	60.1 ± 0.3	59.3 ± 0.6	42.7 ± 0.3	31.8 ± 0.4	52.5 ± 0.1	99.1 ± 0.4
+ linear embedding	39.1 ± 0.4	60.5 ± 0.2	61.0 ± 0.8	58.0 ± 0.4	43.4 ± 1.8	34.5 ± 0.4	51.2 ± 0.8	97.9 ± 0.1
+ MLP embedding	38.8 ± 0.3	60.6 ± 0.3	60.7 ± 0.6	56.9 ± 1.1	41.0 ± 3.1	34.7 ± 0.5	49.6 ± 1.9	97.3 ± 0.3
+ ResNet embedding	37.0 ± 0.5	59.1 ± 0.1	59.2 ± 0.7	57.3 ± 0.7	43.3 ± 2.5	33.6 ± 0.5	51.5 ± 0.8	99.6 ± 0.5
+ FTT embedding	38.8 ± 0.6	59.8 ± 0.1	60.5 ± 0.6	55.7 ± 0.1	39.8 ± 2.6	**38.7** ± 0.3	51.2 ± 0.8	96.9 ± 0.8
+ FTT: type groups	**40.2** ± 0.4	**60.3** ± 0.3	61.6 ± 1.0	54.4 ± 0.3	43.6 ± 0.8	38.0 ± 0.4	52.1 ± 0.1	97.0 ± 1.0
+ FTT: organ groups	**40.6** ± 0.4	**60.7** ± 0.5	**62.3** ± 1.9	**54.0** ± 0.1	**46.5** ± 0.6	37.4 ± 0.1	52.6 ± 0.6	**96.4** ± 0.4

behind classical methods like gradient-boosted trees, as shown in Table 1 and in related work (Yèche et al., 2021), we found that using tabular data deep learning techniques such as FT-Transformer (Gorishniy et al., 2021) helps bridge this performance gap. Building on these insights, our proposed approach of incorporating feature grouping into the embedding process yields further significant performance gains, enabling us to overcome or match the performance of tree-based methods. We refer the interested reader to Appendix C for additional results on other metrics and comparison with other methods, which further support our conclusions. Overall, our analysis establishes **a new state-of-the-art benchmark for clinical time-series tasks**, marking a substantial advancement in the field. Indeed, leveraging well-designed embedding methods and incorporating feature grouping improves performance by a similar scale to optimising the backbone architecture of sequence models in Yèche et al. (2021).

Performance discrepancy between HiRID and MIMIC-III datasets. Incorporating feature groups within the embedding layers shows notable differences in performance gains between the HiRID and MIMIC-III datasets. This discrepancy could be attributed to two primary factors: (1) data resolution and the (2) number of features. With HiRID data resolution being twelve times greater, this leads to sequences of 2016 steps (equivalent to one week) for online tasks. The HiRID dataset processing from Yèche et al. (2021) has a much greater number of features (231), compared to 18 features extracted by Harutyunyan et al. (2019) in MIMIC-III Benchmark. Consequently, FTT models utilizing feature grouping exhibit superior performance on the HiRID dataset, enhancing feature interaction within semantically related groups and rendering the models more resistant to noise, thereby boosting performance. Our results suggest that the use of an embedding module enables deep learning models to extracted relevant signals more effectively. On the contrary, the limited number of features available in MIMIC-III does not allow for significant performance gains with grouping,

Table 2: **Benchmarking analysis of embedding design choices** for circulatory failure prediction. Ablations on the default architecture: FTT (Gorishniy et al., 2021) with organ splitting and attention-based aggregation.

(*a*) Embedding architecture.

Architecture	AUPRC
None	36.6 ± 0.5
MLP	37.6 ± 0.8
ResNet	37.0 ± 0.5
FTT	**40.6** ± 0.4

(*b*) Group aggregation.

Aggregation	AUPRC
Concatenate	39.4 ± 0.2
Average	38.7 ± 0.4
Attention	**40.6** ± 0.4

(*c*) Feature grouping strategies as defined in Section 3.1.

Grouping	AUPRC
None	38.8 ± 0.6
Variable type	39.6 ± 0.1
Meas. type	39.9 ± 0.1
Organ	**40.6** ± 0.4

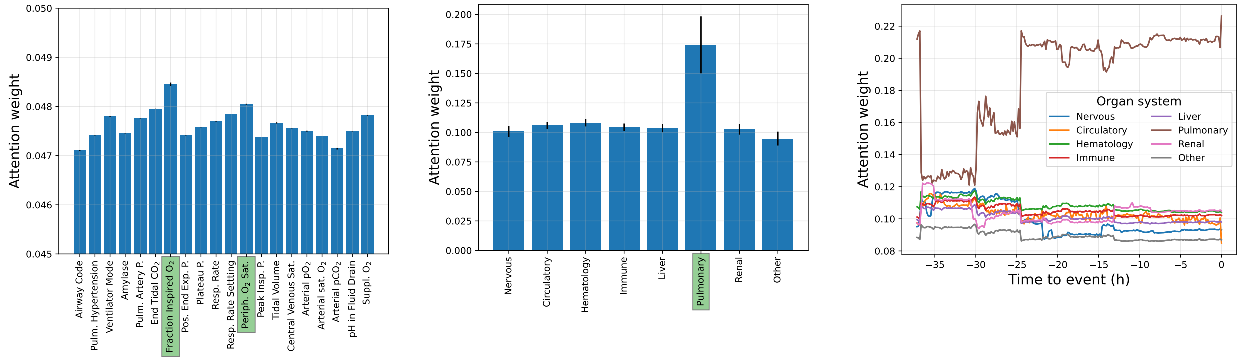
suggesting that this strategy may not be as beneficial in low-dimensional datasets.

Step-wise encoder with feature grouping ablation study. To better understand the impact of each component introduced in Section 3, we investigate various design choices for step-wise embedding architectures, and analyze their impact on performance. Table 2 summarizes our findings of possible concept-level architectures, feature groupings definitions, and aggregation methods. We focus on results for the circulatory failure prediction task for conciseness (referring the reader to Appendix C for exhaustive results on other tasks). In Table 2(*a*), we find that FTT yields the largest performance gains amongst group encoder architectures. Similarly, in Table 2(*b*), we find attention-based aggregation (Vaswani et al., 2017) to consistently improve over other aggregation methods, confirming the need to capture complex concept dependencies present at a time-step level. This supports results from the tabular deep learning literature (Gorishniy et al., 2021). In addition to improving performance, we note that attention mechanisms also provide significant advantages in terms of model interpretability, as further discussed in Section 5. Finally, with respect to different group definitions, we observe pre-defined grouping using domain knowledge to yield the best performance.

Interpretability As a final experiment, we explore the interpretability gained from attention-based models (Choi et al., 2016a; Vig and Belinkov, 2019; Vig, 2019) by analyzing attention at different levels of the model architecture. This provides insights into the relevance of features *within* a single concept embedding as well as the differences in importance *between* concept embeddings to the overall downstream prediction model. Temporal aggregations of attention scores can highlight patient trends in a given time window (Gandin et al., 2021; Lim et al., 2021).

In the context of respiratory failure prediction, we average attention weights over all test patient trajectories and over all timesteps. Within the group of features pertaining to the pulmonary system, we find in Figure 3(*a*) that attention is on average highest on two input features that are highly predictive of this type of organ failure based on its definition (Faltys et al., 2021; Yèche et al., 2021): fraction of inspired oxygen (FiO_2) and peripheral oxygen saturation (SpO_2). We also find in Figure 3(*b*) that the pulmonary organ system has very high importance in predicting respiratory failure, confirming that variables related to lung function, oxygen saturation and ventilation settings are correctly identified as key indicators of event imminence within the embedding model. Note that this analysis is independent of the actual label for a patient at a given time, and thus measures the average contribution of different features and groups in predicting respiratory failure. The various levels of importance scoring can be of assistance to clinicians in different decision-making processes. For instance, on a concept level (e.g. organ systems), it can help to categorize patients in a dynamic way and make it easier to plan resources (e.g. patients with respiratory problems may require ventilators). Additionally, more detailed information on a feature level can be used to make treatment decisions.

Attention scores may not be a perfect explanation (Serrano and Smith, 2019), yet they can still effectively point out important signals. In a clinical decision-support context, these explanations do not need to be taken as absolute truth, but rather as a



(a) Within concept embedding. (b) Between concept embeddings. (c) As a function of time.

Figure 3: Interpretability Analysis of attention-based embeddings for respiratory failure prediction, highlighting the importance of relevant pulmonary variables – particularly close to event occurrence.

way to direct the clinician's attention to the areas that require the most care.

Another form of useful insight gained from attention-based embeddings on clinical time-series consists of patterns of attention as a function of time (Lim et al., 2021), as illustrated in Figure 3(c). For this analysis, we plot attention weights as a function of time for individual patients within the test set. Upon entry to the intensive care unit, the attention mechanism focuses initially on the most relevant organ system, as little patient information is available to predict imminent organ failure. As more information is acquired, attention becomes more balanced across organs, and focuses again on the pulmonary system as a respiratory failure event becomes more likely. This temporal attention pattern highlights the relevance of recent measurements and changes in variables, allowing for a deeper understanding of the predictive patterns and potential early warning signs. We refer the reader to Appendix C for an exhaustive overview of this interpretability analysis, and note that this promising result could benefit from further investigations beyond the scope of the present benchmark, to correlate attention patterns with medical insights and patient evolution trends.

Overall, this study suggests that attention-based embeddings (at different levels in the architecture: features, groups, time) enhance the interpretability of deep learning models for tabular times-series, by shedding light on the most relevant features, medical concept groups and time windows for specific predictions. By understanding which variables are weighted more heavily in the model's decision-making process, clinicians and domain experts can gain trust and validate the machine learning models developed using such embedding methods (Ahmad et al., 2018).

6. Limitations & Broader Impact

Limitations While the FT-Transformer and the use of feature groups provide a powerful setup for the step-wise embedding module, it is crucial to address certain limitations associated with each.

The FT-Transformer is resource-intensive, demanding substantially more hardware and time for training compared to simpler models (MLP, ResNet, especially gradient boosting methods). Scaling it to ICU datasets like HiRID with a large number of features is challenging. Hence, the extensive use of the FT-Transformer for such datasets might increase CO2 emissions from ML pipelines. The research community has already devised a diverse range of solutions aimed at enhancing the speed, memory, and computational efficiency of Transformer-based architectures (Tay et al., 2022). However, when deploying actual models based on these benchmarked architectures, the performance impact of such efficiency-focused modifications remains to be explored.

On the other hand, the concept of feature groups introduces its own set of challenges. Using predefined feature groups, like organ or measurement type, may streamline the model's task, but it could limit its flexibility and requires clinical understanding for effective definition. The role of healthcare professionals is crucial for defining initial feature groups. Limitations also include the challenge of assigning each variable to a single concept, which may not fully capture the multifaceted nature of clinical data.

Broader Impact Integrating a step-wise embedding module with feature groups for ICU models, could impact both the medical and research communities. Firstly, feature grouping approach could help for **precision medicine**. By analyzing a patient's data within the context of their specific feature group, clinicians can better tailor treatment plans to address individual needs. From a machine learning viewpoint, feature splitting can amplify the **performance of predictive models**. Meaningfully grouping data permits these models to discern more complex and nuanced relationships between features, resulting in more accurate predictions. Further, the grouping of data can assist in **continuous patient monitoring**. Healthcare professionals can promptly identify any substantial changes in a patient's condition. Moreover, it can aid in assessing the risk of developing specific conditions, allowing for timely preventative measures. Finally, the **interpretability** derived from attention-based models offers enhanced **trust, validation, and transparency**. By identifying the most relevant features and feature groups and comprehending temporal dynamics through attention patterns, these models become more explainable and trustworthy.

7. Conclusion

Our work benchmarks embedding architectures for deep learning as a new paradigm for clinical time-series tasks, which finally surpasses traditional tree-based methods in terms of performance. Relying on deep learning for tabular data methods, we systematically study different design choices for embedding architectures, demonstrating their essential roles in achieving state-of-the-art results. We find that for distinct groups of features predictive performance significantly improves. We also find that attention-based embeddings offer the best performance as well as greater interpretability, by identifying relevant features and feature groups – such transparency is critical to building trust for real-world clinical applications (Ahmad et al., 2018).

Overall, our study advances the field of machine learning for clinical time-series by leveraging methods and design choices from the tabular deep learning literature. We believe our findings will encourage further work in embedding design for clinical time-series, with the potential to better support clinical decision-making and improve patient outcomes.

Acknowledgments

This project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology) and by ETH core funding (to G.R). This publication was made possible by an ETH AI Center doctoral fellowship to AP.

Institutional Review Board (IRB) This research does not require IRB approval in the country in which it was performed.

References

Henrique Aguiar, Mauro Santos, Peter Watkinson, and Tingting Zhu. Learning of cluster-based feature importance for electronic health record time-series. In *International Conference on Machine Learning*, pages 161–179. PMLR, 2022.

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Joseph Y. Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals, 2021. URL https://arxiv.org/pdf/2007.04871.pdf.

Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*, 2023.

Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3512–3520, Red Hook, NY, USA, 2016a. Curran Associates Inc. ISBN 9781510838819.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Menghan Ding and Yuan Luo. Unsupervised phenotyping of sepsis using nonnegative matrix factorization of temporal trends from a multivariate panel of physiological measurements. *BMC Medical Informatics and Decision Making*, 21:1–15, 2021.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. Hirid, a high time-resolution icu dataset (version 1.1. 1). *Physio. Net*, 10, 2021.

Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780): 1612, 1999.

Ilaria Gandin, Arjuna Scagnetto, Simona Romani, and Giulia Barbati. Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit. *Journal of Biomedical Informatics*, 121:103876, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2021.103876. URL https://www.sciencedirect.com/science/article/pii/S1532046421002057.

The gin-config Team. gin-config python packaged. https://github.com/google/gin-config, 2019.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 2021.

Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 35:24991–25004, 2022.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten M. Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4353–4363. PMLR, 2020. URL http://proceedings.mlr.press/v119/horn20a.html.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Severin Husmann, Hugo Yèche, Gunnar Ratsch, and Rita Kuznetsova. On the importance of clinical notes in multi-modal learning for ehr data. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Fergus Imrie, Alexander Norcliffe, Pietro Lió, and Mihaela van der Schaar. Composite feature selection using deep ensembles. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36142–36160. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/eab69250e98b1f9fc54e473cc7a69439-Paper-Conference.pdf.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

Matthew Kelly and Christopher Semsarian. Multiple mutations in genetic cardiovascular disease: a marker of disease severity? *Circulation: Cardiovascular Genetics*, 2(2):182–190, 2009.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. {CLOCS}: Contrastive learning of cardiac signals across space, time, and patients, 2021. URL https://openreview.net/forum?id=4Nt1F3qf9Gn.

W A Knaus, E A Draper, D P Wagner, and J E Zimmerman. APACHE II: a severity of disease classification system. *Crit. Care Med.*, 13(10):818–829, October 1985.

Jannik Kossen, Neil Band, Clare Lyle, Aidan N Gomez, Thomas Rainforth, and Yarin Gal. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *Advances in Neural Information Processing Systems*, 34:28742–28756, 2021.

Alex Labach, Aslesha Pokhrel, Seung Eun Yi, Saba Zuberi, Maksims Volkovs, and Rahul G Krishnan. Effective self-supervised transformers for sparse time series data. 2023. URL https://openreview.net/pdf?id=HUCgU5EQluN.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.03.012. URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. Predicting icu mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13374–13386. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9b10a919ddeb07e103dc05ff523afe38-Paper.pdf.

Lisiane B Meira, Antonio MC Reis, David L Cheo, Dorit Nahari, Dennis K Burns, and Errol C Friedberg. Cancer predisposition in mutant mice defective in multiple genetic pathways: uncovering important genetic interactions. *Mutation Research/-Fundamental and Molecular Mechanisms of Mutagenesis*, 477(1-2):51–58, 2001.

NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. Cuda, release: 10.2.89, 2020. URL https://developer.nvidia.com/cuda-toolkit.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression. In *International Conference on Artificial Intelligence and Statistics*, pages 3466–3492. PMLR, 2023.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282.

Gowthami Somepalli, Avi Schwarzschild, Micah Goldblum, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6):2765–2787, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://aclanthology.org/P19-3007.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808.

Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2565–2573, 2018.

Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pages 11964–11974. PMLR, 2021.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. *arXiv preprint arXiv:2111.08536*, 2021.

Hugo Yèche, Alizée Pace, Gunnar Rätsch, and Rita Kuznetsova. Temporal label smoothing for early prediction of adverse events. *arXiv preprint arXiv:2208.13764*, 2022.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, 2020.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2021.

Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit. Care Med.*, 34(5):1297–1310, May 2006.

Appendix A. Clinical datasets and prediction tasks

A.1. Task definition

In this section, we provide more details on the definition of tasks for circulatory failure, respiratory failure and mortality from HiRID benchmark (Yèche et al., 2021) and decompensation and mortality from MIMIC-III benchmark (Harutyunyan et al., 2019). The details about the MIMIC-III and HiRID datasets, including the number of patients, endpoint definition, and statistics on annotated failure events and labels, are available in the corresponding papers that introduced these datasets: (Johnson et al., 2016) for MIMIC-III and (Faltys et al., 2021) for HiRID. The respective patient splits are provided in the corresponding benchmark papers: (Harutyunyan et al., 2019) for MIMIC-III and (Yèche et al., 2021) for HiRID.

HiRID benchmark tasks:

1. **Circulatory failure** is a failure of the cardiovascular system, detected in practice through elevated arterial lactate (> 2 mmol/l) and either low mean arterial pressure (< 65 mmHg) or administration of a vasopressor drug. Yèche et al. (2021) defines a patient to be experiencing a circulatory failure event at a given time if those conditions are met for 2/3 of time points in a surrounding two-hour window. Binary classification, dynamic prediction throughout stay
2. **Respiratory failure** is defined by Yèche et al. (2021) as a P/F ratio (arterial pO_2 over FIO_2) below 300 mmHg. This definition includes mild respiratory failure. As above, Yèche et al. (2021) consider a patient to be experiencing respiratory failure if 2/3 of timepoints are positive within a surrounding 2h window. Binary classification, dynamic prediction throughout stay
3. **Mortality** refers to the death of the patient. The label of the time-point 24 hours after ICU admission was set to 1 (positive) if the patient died at the end of the stay according to this field, and 0 (negative) otherwise, defining a binary classification problem to be solved once per stay. If the admission was shorter than 24 hours, no label was assigned to the patient.
4. **Patient phenotyping** is classifying the patient after 24h regarding the admission diagnosis, using the APACHE group II and IV labels[2].
5. **Remaining length of stay** is a regression task, continuous prediction of the remaining ICU stay duration.

MIMIC-III benchmark tasks:

1. **Decompensation** refers to the death of a patient in the next 24h. The event labels are directly extracted from the MIMIC-III (Johnson et al., 2016) metadata about the time of death of a patient.
2. **Mortality** refers to the death of a patient after 48 hours of observed ICU data. The event labels are directly extracted from the MIMIC-III (Johnson et al., 2016) metadata.
3. **Length of stay** is a prediction of the remaining time the patient will stay in the ICU.

MIMIC-III license is PhysioNet Credentialed Health Data License 1.5.0; HiRID — PhysioNet Contributor Review Health Data License 1.5.0.

A.2. Pre-processing

We describe the pre-processing steps we applied to both datasets, HiRID and MIMIC-III.

Imputation. Diverse imputation methods exist for ICU time series. For simplicity, we follow the approach of original benchmarks (Harutyunyan et al., 2019; Yèche et al., 2021) by using forward imputation when a previous measure existed. The remaining missing values are zero-imputed after scaling, corresponding to a mean imputation.

Scaling. Whereas prior work explored clipping the data to remove potential outliers (Tomašev et al., 2019), we do not adopt this approach as we found it to reduce performance on early prediction tasks. A possible explanation is that, due to the rareness of events, clipping extreme quantiles may remove parts of the signal rather than noise. Instead, we simply standard-scale data based on the training sets statistics.

Appendix B. Implementation details

B.1. Modality splitting

Organ splitting is detailed in Table 10 and Table 11. Splitting by variable type is provided in Table 12 and

[2] APACHE II and IV (Zimmerman et al., 2006; Knaus et al., 1985) are subsequent versions of the major illness severity score used in the ICU. They also introduce a patient grouping according to admission reason. We use an aggregate of these two groupings for this task (see also Yèche et al. (2021))

Table 3: **Hyperparameter search range** for mortality, MIMIC-III with Transformer(Vaswani et al., 2017) backbone. In **bold** are parameters selected by random search.

Hyperparameter	Values
Learning Rate	(1e-5, 3e-5, **1e-4**, 3e-4)
Drop-out	(0.0, 0.1, 0.2, 0.3, **0.4**)
Attention Drop-out	(0.0, 0.1, 0.2, **0.3**, 0.4)
Depth	(**1**, 2, 3)
Heads	(**1**, **2**, 4)
Hidden Dimension	(16, 32, **64**)
L1 Regularization	(**1e-3**, 1e-2, 1e-1, 1, 10)

Table 13. Both are obtained from metadata in HiRID and MIMIC-III datasets, which specify which organ or value type each variable belongs to. Measurement splitting is determined by whether the variable is numerical or categorical, and this can be found in the related dataset descriptions (Faltys et al., 2021; Johnson et al., 2016).

B.2. Training Setup

Training details. For all models, we set the batch size according to the available hardware capacity. We use `Nvidia RTX2080 Ti` GPUs with 11GB of GPU memory. Depending on the model size, dataset and task, we use between 1 to 8 GPUs in a distributed data-parallel mode. We early stopped each model training according to their validation loss when no improvement was made after 10 epochs.

Libraries. A full list of libraries and the version we used is provided in the `environment.yml` file. The main libraries on which we build our experiments are the following: pytorch 1.11.0 (Paszke et al., 2019), scikit-learn 0.24.1(Pedregosa et al., 2011), ignite 0.4.4, CUDA 10.2.89(NVIDIA et al., 2020), cudNN 7.6.5(Chetlur et al., 2014), gin-config 0.5.0 (gin-config Team, 2019).

Infrastructure. We follow all guidelines provided by `pytorch` documentation to ensure the reproducibility of our results. However, reproducibility across devices is not ensured. Thus we provide here the characteristics of our infrastructure. We trained all models on a 1 to 8 `Nvidia RTX2080 Ti` with a `Xeon E5-2630v4` CPU. Training took between 3 and 10 hours for a single run.

Architecture choices for the sequential backbone model. We used the same architecture and hyperparameters reported giving the best performance on circulatory failure, respiratory failure and decompensation in Yèche et al. (2022). For all other tasks from HiRID benchmark, we used the same architecture and hyperparameters reported in Yèche et al. (2021). For mortality, MIMIC-III benchmark we carried out our own random search on validation AUPRC performance. The exact parameters for this task are reported in Table 3.

Gradient Boosting We used the same architecture and hyperparameters reported giving the best performance on HiRID benchmark tasks in Yèche et al. (2021). For mortality and decompensation, MIMIC-III benchmark we carried out our own random search on validation AUPRC performance. The range of hyperparameters considered for the gradient boosting method, LightGBM framework[5] can be found in Table 4.

B.3. Embedding architectures

We follow MLP, ResNet and FT-Transformer implementation, described in Gorishniy et al. (2021). Architecture and hyperparameters investigated for each task are given in Table 5 for MLP and ResNet architectures and in Table 6 for FT-Transformer, along with the setting giving optimal validation performance in each case.

B.4. Concept aggregation

Embeddings from each concept are aggregated by taking the average of the multiple embeddings, concatenating them, of computed an attention-based aggregation. Hyperparamters investigated for each task for attention-based aggregation are given in Table 7.

Appendix C. Additional results and ablations

In this Section, we provide the additional results on other metrics, which support our conclusions from the Section 5.

[5] https://lightgbm.readthedocs.io/en/latest/

Task	Depth	Colsample_bytree [3]	Subsample [4]
Mortality	(3, **4**, 5, 6, **7**)	(0.33, **0.66**, 1.00)	(0.33, 0.66, **1.00**)
Decompensation	(3, 4, 5, **6**, 7)	(**0.33**, 0.66, 1.00)	(0.33, 0.66, **1.00**)

Table 4: Hyperparameter search range for LGBM. In **bold** are the parameters we selected using random search.

Table 5: **Embedding architecture and hyperparameter values** studied for each clinical prediction task for MLP and ResNet architectures. Best values, obtained by random search over the proposed grid, are highlighted in bold.

Dataset	HiRID			MIMIC-III	
Clinical prediction task	Circulatory Failure	Respiratory Failure	Mortality	Decompensation	Mortality
Embedding architecture	(**MLP**, ResNet)	(**MLP**, ResNet)	(**MLP**, ResNet)	(**MLP**, ResNet)	(**MLP**, ResNet)
Modality split	(none, organ, categorical, **type**)	(none, **organ**, categorical, type)	(none, **organ**, **categorical**, type)	(none, **organ**, categorical, type)	(**none**, organ, categorical, type)
Aggregation	(avg., **concat**., attention)	(avg., **concat.**, attention)	(**avg.**, concat., attention)	(avg., concat., **attention**)	(**avg.**, concat., attention)
Embedding depth	(1 2 **3** 4)	(1 **2** 3 4)	(**1** 2 3 4)	(1 2 **3** 4)	(**1** 2 3 4)
Embedding latent dim.	(8 16 **32** 64)	(8 16 **32** 64)	(8 16 32 **64**)	(8 16 32 **64**)	(8 16 **32** 64)
L1 regularization weight	(0 0.1 1 **10**)	(0 0.1 1 **10**)	(0 **0.1** 1 10)	(0 **0.1** 1 10)	(**0** 0.1 1 10)

Table 6: **Embedding architecture and hyperparameter values** studied for each clinical prediction task for FTT architecture. Best values, obtained by random search over the proposed grid, are highlighted in bold.

Dataset	HiRID			MIMIC-III	
Clinical prediction task	Circulatory Failure	Respiratory Failure	Mortality	Decompensation	Mortality
Modality split	(none, **organ**, categorical, type)	(none, **organ**, categorical, type)	(none, **organ**, categorical, type)	(**none**, organ, categorical, type)	(none, **organ**, categorical, type)
Aggregation	(avg., concat., **attention**)	(avg. concat., **attention**)	(avg., **concat.**, attention)	(avg., concat., **attention**)	(**avg.**, concat., attention)
FTT token dim	(32 **64**)	(32 **64**)	(16 32 **64** 128)	(16 32 **64** 128)	(16 32 **64** 128)
FTT depth	(**1** 2)	(**1** 2)	(**1** 2 3)	(1 **2** 3)	(**1** 2 3)
FTT heads	(1 **2** 3)	(1 2 **3**)	(1 **2** 3)	(1 2 **3**)	(1 2 **3**)

Table 7: **Hyperparameter values** studied for each clinical prediction task for attention-based aggregation. Best values, obtained by random search over the proposed grid, are highlighted in bold.

Dataset	HiRID			MIMIC-III	
Clinical prediction task	Circulatory Failure	Respiratory Failure	Mortality	Decompensation	Mortality
Agg. depth	(1 **2** 3)	(1 **2** 3)	(1 **2** 3)	(1 **2** 3)	(1 **2** 3)
Agg. heads	(1 2 **3**)	(1 2 **3**)	(1 2 **3**)	(1 **2** 3)	(**1** 2 3)

C.1. Comparison with unsupervised pretraining techniques

In Table 8 we provide a comparison with pretraining techniques followed by training MLPs on top of the pretrained representations to perform the downstream prediction tasks. For fair comparison with Yèche et al. (2021) we used a temporal convolutional network (TCN) as the backbone sequence architecture.

C.2. Additional performance benchmark results for embedding architectures

In this section, we additionally compare the previously described methods on all tasks. First, we report AUROC metric for the results, given in the Section 5, Table 1, see Table 9.

In addition to Table 2 we summarize our findings of possible concept-level architectures, feature groupings definitions, and aggregation methods on all other tasks in Table 14 - Table 20.

Table 8: Comparison with a set of unsupervised pretraining techniques on two MIMIC-III benchmark tasks: Decompensation and Length-of-stay. Semi-supervised approaches inlcude some labels to pretrain patient representations.

Task	Decompensation		Length-of-stay
Metric	AUPRC	AUROC	Kappa
Self-Supervised Pretraining			
SACL (Cheng et al., 2021)	29.3 ± 0.9	87.5 ± 0.4	40.1 ± 0.5
CLOCS (Kiyasseh et al., 2021)	32.2 ± 0.8	90.5 ± 0.2	43.0 ± 0.2
NCL (Yèche et al., 2021)	35.1 ± 0.4	90.8 ± 0.2	43.2 ± 0.2
Semi-Supervised Pretraining			
SCL (D) (Khosla et al., 2021)	32.1 ± 0.9	89.5 ± 0.3	41.8 ± 0.4
NCL (Yèche et al., 2021)	37.1 ± 0.7	90.9 ± 0.1	43.8 ± 0.3
Our Supervised Step-Wise Embedding Approach			
FTT embedding (Gorishniy et al., 2021)	**38.2** ± 0.4	90.9 ± 0.3	42.9 ± 0.6
FTT with organ grouping	**38.2** ± 0.5	**91.1** ± 0.3	**44.0** ± 0.3

C.3. Interpretability

Additional tasks. Additional results on circulatory failure prediction are shown in Figure 4 and in Figure 5. Average attention weights between different organ systems, highlight the importance of relevant groups of features in predicting the corresponding organ failure. We find that the cardiovascular and hematology organ systems show the highest relevance to predicting circulatory failure, confirming that variables related to heart function, blood pressure, and vascular dynamics may play a critical role. Overall, features and organ groups with highest attention weights correspond to important predictive variables from a clinical perspective as shown by Hyland et al. (2020).

Attention over time. We provide additional examples of attention pattern over time in Figure 6 and Figure 7, showing the insights gained from attention-based embedding methods in interpreting model behaviour.

Appendix D. Prior work: deep learning backbones for ICU data

As was mentioned, the performance gap between proposed deep learning methods and tree-based approaches remains significant (Yèche et al., 2021; Hyland et al., 2020). Some approaches have considered the use of additional data sources via fusion models (Husmann et al., 2022; Khadanga et al., 2019) to achieve comparable performance. In this section, we also aim to address several prior papers that have contributed to the development of backbone model architectures for supervised clinical time-series tasks. The SeFT model (Horn et al., 2020) treats observations as tuples of time value, observed variable value, and modality indicator. These tuples are concatenated and each individually passed through a linear layer to generate embeddings, which then are aggregated across the *entire time-series*. The RAINDROP model (Zhang et al., 2021) maps each observed variable to a high-dimensional space with an MLP and uses Graph Neural Networks to learn relevant relationships. Separate line of research explores self-supervised pre-training methodologies for clinical time-series representation learning (Tipirneni and Reddy, 2022; Labach et al., 2023). StraTS (Tipirneni and Reddy, 2022) represents the data in the same way as SeFT. The TESS model (Labach et al., 2023) considers time bins which are passed through an MLP. To summarize, SeFT and StraTS employ the same architecture, where the features interact within the whole time-series, and necessitates a specific data representation. Time-step level and group level interactions are not in the scope of these studies. Similar to SeFT, RAINDROP aggregates information across the entire time-series for the feature embeddings and employs an architecture not suited for online prediction

Table 9: **Performance benchmark for different embedding architectures,** measured through the receiver operating characteristic curve (AUROC). Mean and standard deviation are reported over five training runs.

Dataset	HiRID			MIMIC-III	
Clinical prediction task	Circulatory Failure	Respiratory Failure	Mortality	Decompensation	Mortality
LightGBM (Yèche et al., 2021)	**91.2** ± 0.1	**70.8** ± 0.1	90.5 ± 0.0	90.3 ± 0.1	84.2 ± 0.1
Deep learning backbone (DL)	90.5 ± 0.2	69.9 ± 0.4	90.7 ± 0.2	90.5 ± 0.1	86.1 ± 0.1
+ linear embedding (Yèche et al., 2021)	90.9 ± 0.1	**71.0** ± 0.2	90.8 ± 0.2	91.1 ± 0.1	85.8 ± 0.2
+ MLP embedding (Tomašev et al., 2021)	91.0 ± 0.1	70.7 ± 0.3	90.5 ± 0.1	91.1 ± 0.5	85.6 ± 0.1
+ ResNet embedding (Tomašev et al., 2019)	90.1 ± 0.3	69.5 ± 0.1	89.9 ± 0.2	90.7 ± 0.2	85.9 ± 0.2
+ FTT embedding (Gorishniy et al., 2021)	91.1 ± 0.1	70.0 ± 0.1	90.5 ± 0.2	**91.6** ± 0.1	85.8 ± 0.2
+ FTT with type grouping	91.0 ± 0.2	**70.6** ± 0.2	90.1 ± 0.3	**91.4** ± 0.1	**86.0** ± 0.2
+ FTT with organ grouping	**91.6** ± 0.03	**70.6** ± 0.4	**91.0** ± 0.3	**91.4** ± 0.1	**86.1** ± 0.2

tasks. TESS and StraTS focus on exploring the effects of self-supervised pre-training, which is distinct from the focus of our research.

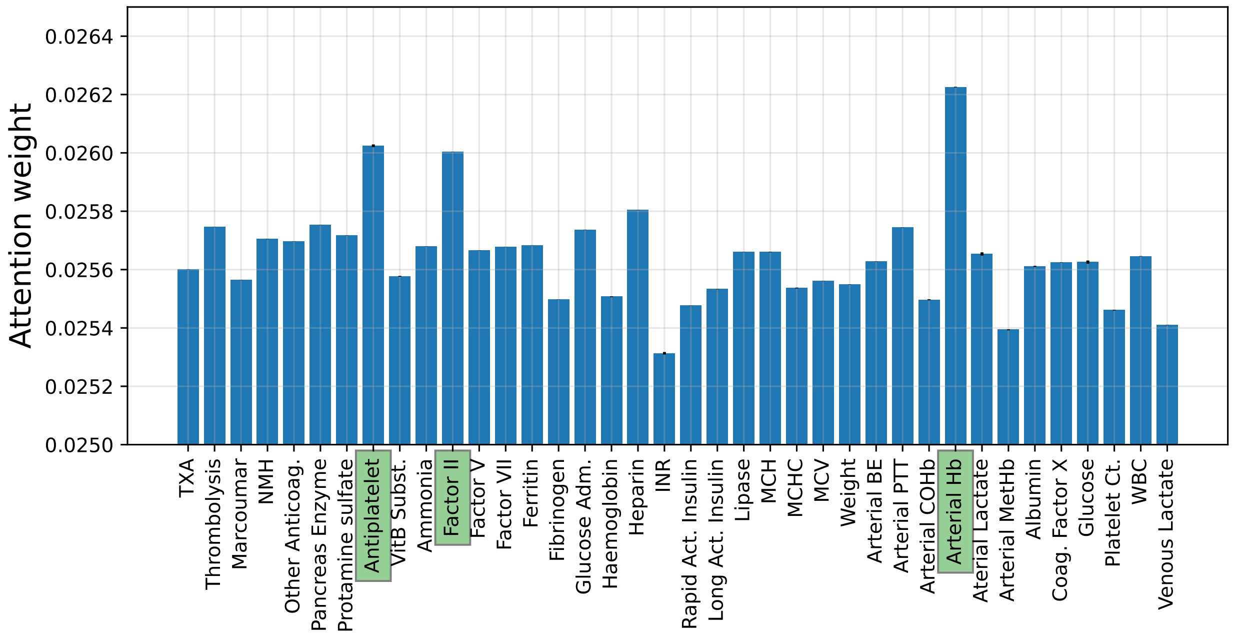


Figure 4: Within concept embedding (hematology system).

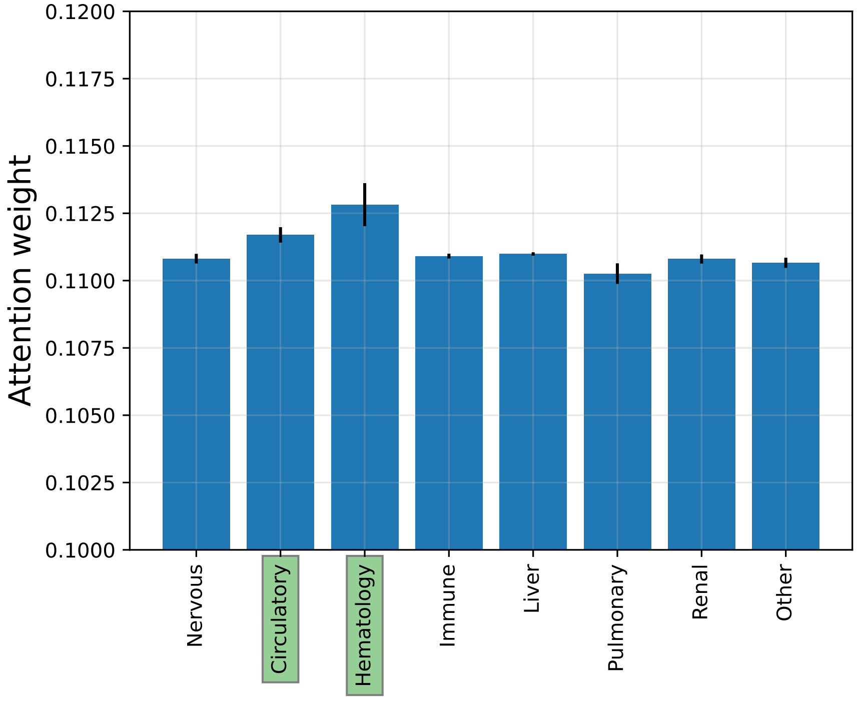
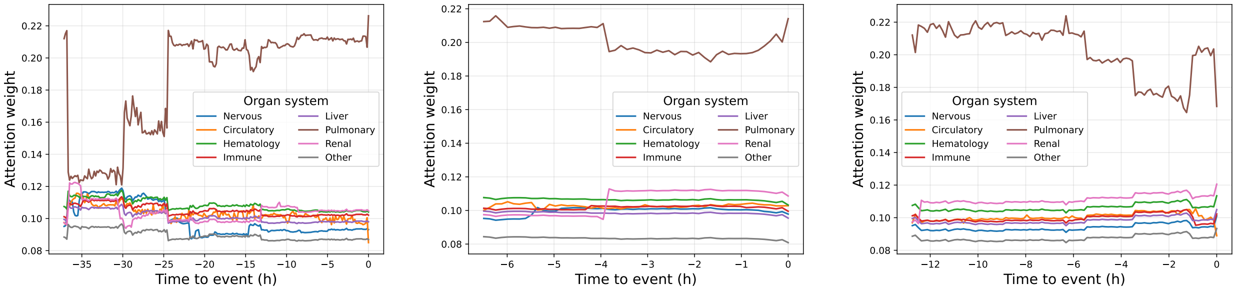
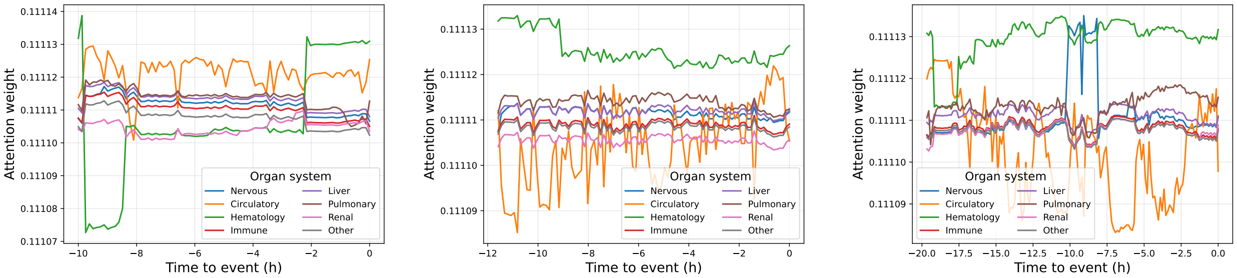


Figure 5: Between concept embeddings (organ systems).



(a) (b) (c)

Figure 6: **Attention patterns over time** in embeddings for clinical time-series for Respiratory failure prediction task. Example attention weights between different organ systems.



(a) (b) (c)

Figure 7: **Attention patterns over time** in embeddings for clinical time-series for Circulatory failure prediction task. Example attention weights between different organ systems.

Table 10: **Variable splitting by organ type,** obtained based on public metadata in the HiRID dataset. An intensive care physician was consulted to confirm the validity of these splits. Details on variable name and acronyms can be obtained from the respective datasets (Hyland et al., 2020).

Organ	Variable name
Central nervous system (Brain)	GCS Antwort, GCS Motorik, GCS Augenöffnen, RASS, ICP, TOF, Benzodiacepine, Alpha 2 Agonisten, Barbiturate, Propofol, Liquor/h, Nimodipin, Opiate, Non-opioide, NSAR, Ketalar, Peripherial Anesthesia, Antiepileptica, Anti delirant medi, Psychopharma, Muskelrelaxans, Anexate, Naloxon, Parkinson Medikaiton, pH Liquor, Laktat Liquor, Glucose Liquor
Circulatory system (Heart)	HR, T Central, ABPs, ABPd, ABPm, NIBPs, NIBPd, NIBPm, PAPm, PAPs, PAPd, PCWP, CO, SvO2(m), ZVD, ST1, ST2, ST3, Rhythmus, IN, OUT, Incrys, Incolloid, packed red blood cells, FFP, platelets, coagulation factors, norepinephrine, epinephrine, dobutamine, milrinone, levosimendan, theophyllin, vasopressin, desmopressin, vasodilatators, ACE Inhibitors, Sartane, Ca Antagonists, B-Blocker, Andere, Adenosin, Digoxin, Amiodaron, Atropin, Thyrotardin, Thyroxin, Thyreostatikum, Mineralokortikoid, Antihistaminka, Terlipressin, Troponin-T, creatine kinase, creatine kinase-MB, BNP, TSH, AMYL-S
Hematology (Blood)	Glucose Administration, Insuling Langwirksam, Insulin Kurzwirksam, Thrombozytenhemmer, Heparin, NMH, Others in Case of HIT, Marcoumar, Protamin, Anti Fibrinolyticum, Lysetherapie, Pankreas Enzyme, VitB Substitution, Weight, a-BE, a_COHb, a_Hb, a_Lac, a_MetHb, v-Lac, aPTT, Fibrinogen, FII, Factor V, Factor VII, factor X, INR, albumin, glucose, Ammoniak, Hb, total white blood cell count, platelet count, MCH, MCHC, MCV, Ferritin, Lipase
Immune system	Administriation of antibiotics, Administation of antimycotic, administration of antiviral, Antihelmenticum, Steroids, Enteral Feeding, steroids, non-steroids, Chemotherapie, Immunoglobulin, Immunsuppression, GCSF, C-reactive protein, procalcitonin, lymphocyte, Neutr, Segm. Neut., Stabk. Neut., BSR, Cortisol
Hepatic system (Liver)	ASAT, ALAT, bilirubine, total, Bilirubin, direct, alkaline phosphatase, gamma-GT
Pulmonary system (Lung)	SpO2, ETCO2, RR, supplemental oxygen, FIO2, Peep, Ventilator mode, TV, Spitzendruck, Plateaudruck, AWPmean, RR set, AiwayCode, Beh. Pulm. Hypertonie, a_pCO2, a_PO2, a_SO2, Zentral venöse sättigung, pH Drain, AMYL-Drainag
Renal system (Kidneys)	OUTurine/h, K-sparend, Aldosteron Antagonist, Loop diuretics, Thiazide, Acetazolamide, Haemofiltration, Parenteral Feeding, Kalium, Phosphat, Na, Mg, Ca, Trace elements, Bicarbonate, a_HCO3-, a_pH, K+, Na+, Cl-, Ca2+ ionizied, Ca2+ total, phosphate, Mg_lab, Urea, creatinine, urinary creatinin, urinary Na+, urinary urea

Table 11: **Variable splitting by organ type,** obtained based on public metadata in MIMIC-III dataset. An intensive care physician was consulted to confirm the validity of these splits. Details on variable name and acronyms can be obtained from the respective datasets (Hyland et al., 2020).

Organ	Variable name
Central nervous system (Brain)	Glascow coma scale eye opening, Glascow coma scale motor response, Glascow coma scale total, Glascow coma scale verbal response
Circulatory system (Heart)	Diastolic blood pressure, Heart Rate, Mean blood pressure, Systolic blood pressure, Temperature, Capillary refill rate
Hematology (Blood)	Glucose
Pulmonary system (Lung)	Fraction inspired oxygen, Oxygen saturation, Respiratory rate
Renal system (Kidneys)	pH

Table 12: **Variable splitting by data acquisition type,** obtained based on public metadata in the HiRID dataset.

Variable type	Variable name
Derived from raw data	ETCO2, OUTurine/h, IN, OUT, Incrys, Incolloid
Laboratory values	a-BE, a_COHb, a_Hb, a_HCO3-, a_Lac, a_MetHb, a_pH, a_pCO2, a_PO2, a_SO2, Zentral venöse sättigung, Troponin-T, creatine kinase, creatine kinase-MB, v-Lac, BNP, K+, Na+, Cl-, Ca2+ ionizied, Ca2+ total, phosphate, Mg_lab, Urea, creatinine, urinary creatinin, urinary Na+, urinary urea, ASAT, ALAT, bilirubine, total, Bilirubin, direct, alkaline phosphatase, gamma-GT, aPTT, Fibrinogen, FII, Factor V, Factor VII, factor X, INR, albumin, glucose, Ammoniak, C-reactive protein, procalcitonin, lymphocyte, Neutr, Segm. Neut., Stabk. Neut., BSR, Hb, total white blood cell count, platelet count, MCH, MCHC, MCV, Ferritin, TSH, AMYL-S, Lipase, Cortisol, pH Liquor, Laktat Liquor, Glucose Liquor, pH Drain, AMYL-Drainag
Monitored variables	HR, T Central, ABPs, ABPd, ABPm, NIBPs, NIBPd, NIBPm, PAPm, PAPs, PAPd, PCWP, CO, SvO2(m), ZVD, ST1, ST2, ST3, SpO2, ETCO2, RR, ICP, TOF, FIO2, Peep, Ventilator mode, TV, Spitzendruck, Plateaudruck, AWPmean, RR set
Observed variables	ZVD, Rhythmus, supplemental oxygen, GCS Antwort, GCS Motorik, GCS Augenöffnen, RASS, ICP, AiwayCode, Haemofiltration, Liquor/h, Weight
Treatment variables	packed red blood cells, FFP, platelets, coagulation factors, norepinephrine, epinephrine, dobutamine, milrinone, levosimendan, theophyllin, vasopressin, desmopressin, vasodilatators, ACE Inhibitors, Sartane, Ca Antagonists, B-Blocker, Andere, Adenosin, Digoxin, Amiodaron, Atropin, K-sparend, Aldosteron Antagonist, Loop diuretics, Thiazide, Acetazolamide, Administriation of antibiotics, Administation of antimycotic, administration of antiviral, Antihelmenticum, Benzodiacepine, Alpha 2 Agonisten, Barbiturate, Propofol, Glucose Administration, Insuling Langwirksam, Insulin Kurzwirksam, Nimodipin, Opiate, Non-opioide, NSAR, Ketalar, Peripherial Anesthesia, Steroids, Thrombozytenhemmer, Enteral Feeding, Parenteral Feeding, Heparin, NMH, Others in Case of HIT, Marcoumar, Protamin, Anti Fibrinolyticum, Kalium, Phosphat, Na, Mg, Ca, Trace elements, Bicarbonate, Antiepileptica, Anti delirant medi, Psychopharma, steroids, non-steroids, Thyrotardin, Thyroxin, Thyreostatikum, Mineralokortikoid, Antihistaminka, Chemotherapie, Lysetherapie, Muskelrelaxans, Anexate, Naloxon, Beh. Pulm. Hypertonie, Pankreas Enzyme, Terlipressin, Immunoglobulin, Immunsuppression, VitB Substitution, Parkinson Medikaiton, GCSF

Table 13: **Variable splitting by data acquisition type,** obtained based on public metadata in the MIMIC-III dataset.

Variable type	Variable name
Laboratory values	Glucose, pH
Monitored variables	Diastolic blood pressure, Heart Rate, Mean blood pressure, Systolic blood pressure, Temperature, Fraction inspired oxygen, Oxygen saturation, Respiratory rate
Observed variables	Glascow coma scale eye opening, Glascow coma scale motor response, Glascow coma scale total, Glascow coma scale verbal response, Capillary refill rate

Table 14: **Benchmarking analysis of embedding design choices** for *Respiratory failure prediction* on the HiRID dataset. Ablations on the default architecture: FTT (Gorishniy et al., 2021) with organ splitting and attention-based aggregation.

(*a*) Embedding architecture.

Architecture	AUPRC
None	59.5 ± 0.4
MLP	**60.6** ± 0.2
ResNet	58.2 ± 0.4
FTT	**60.7** ± 0.5

(*b*) Group aggregation.

Aggregation	AUPRC
Concatenate	**61.1** ± 0.1
Average	60.1 ± 0.3
Attention	**60.7** ± 0.2

(*c*) Feature grouping.

Grouping	AUPRC
None	59.8 ± 0.1
Variable type	**60.7** ± 0.1
Meas. type	**60.3** ± 0.3
Organ	**60.7** ± 0.5

Table 15: **Benchmarking analysis of embedding design choices** for *Mortality prediction* on the HiRID dataset. Ablations on the default architecture: FTT (Gorishniy et al., 2021) with organ splitting and attention-based aggregation.

(*a*) Embedding architecture.

Architecture	AUPRC
None	60.1 ± 0.3
MLP	60.3 ± 0.2
ResNet	57.8 ± 0.4
FTT	**61.6** ± 1.3

(*b*) Group aggregation.

Aggregation	AUPRC
Concatenate	**62.3** ± 1.9
Average	61.0 ± 0.7
Attention	61.6 ± 1.3

(*c*) Feature grouping.

Grouping	AUPRC
None	60.5 ± 0.6
Variable type	60.9 ± 0.2
Meas. type	61.6 ± 1.0
Learned	**62.3** ± 1.2
Organ	**61.6** ± 1.3

Table 16: **Benchmarking analysis of embedding design choices** for *Length-of-Stay* prediction on the HiRID dataset. Best performing model shown while fixing the specific variation and performing a random search over the others.

(*a*) Embedding architecture.

Architecture	MAE ↓
None	59.3 ± 0.6
MLP	56.9 ± 1.1
ResNet	57.3 ± 0.7
FTT	**54.0** ± 0.1

(*b*) Group aggregation.

Aggregation	MAE ↓
Concatenate	54.2 ± 0.2
Average	**54.0** ± **0.1**
Attention	**54.0** ± **0.1**

(*c*) Feature grouping

Grouping	MAE ↓
None	55.7 ± 0.1
Meas. type	54.4 ± 0.3
Organ	**54.0** ± 0.1

Table 17: **Benchmarking analysis of embedding design choices** for *Phenotyping* prediction on the HiRID dataset. Best performing model shown while fixing the specific variation and performing a random search over the others.

(*a*) Embedding architecture.

Architecture	*Bal.Acc* ↑
None	42.7 ± 1.5
MLP	39.5 ± 1.8
ResNet	43.3 ± 1.7
FTT	**46.5** ± 1.4

(*b*) Group aggregation.

Aggregation	*Bal.Acc* ↑
Concatenate	43.2 ± 0.9
Sum	**46.5** ± 1.4
Attention	41.8 ± 1.7

(*c*) Feature grouping

Grouping	*Bal.Acc* ↑
None	39.8 ± 2.6
Meas. type	43.6 ± 0.8
Organ	**46.5** ± 1.4

Table 18: **Benchmarking analysis of embedding design choices** for *Decompensation prediction* on MIMIC-III dataset. Ablations on the default architecture: FTT (Gorishniy et al., 2021) with organ splitting and attention-based aggregation.

(*a*) Embedding architecture.

Architecture	AUPRC
None	**38.7** ± 0.3
MLP	36.3 ± 0.3
FTT	38.0 ± 0.4

(*b*) Group aggregation.

Aggregation	AUPRC
Concatenate	36.2 ± 1.3
Average	37.4 ± 0.1
Attention	**38.0** ± 0.4

(*c*) Feature grouping.

Grouping	AUPRC
None	**38.7** ± 0.3
Variable type	34.8 ± 0.3
Meas. type	38.1 ± 0.2
Organ	38.0 ± 0.4

Table 19: Feature grouping.

Table 20: **Benchmarking analysis of embedding design choices** for *Mortality prediction* on MIMIC-III dataset. Ablations on the default architecture: FTT (Gorishniy et al., 2021) with organ splitting and attention-based aggregation.

(*a*) Embedding architecture.

Architecture	AUPRC
None	51.2 ± 0.8
MLP	51.3 ± 1.01
ResNet	50.6 ± 0.7
FTT	**51.8** ± 0.6

(*b*) Group aggregation.

Aggregation	AUPRC
Concatenate	51.9 ± 0.6
Average	**52.6** ± 0.6
Attention	51.8 ± 0.6

(*c*) Group aggregation.

Grouping	AUPRC
None	51.1 ± 0.5
Variable type	51.1 ± 0.7
Meas. type	51.4 ± 2.2
Organ	**51.8** ± 0.6