CLINICAL TRAJECTORY REPRESENTATIONS FOR CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Analyzing and grouping typical patient trajectories is crucial to understanding their health state, estimating their prognosis, and determining optimal treatment. The increasing availability of electronic health records (EHRs) opens the opportunity to support clinicians in their decisions with machine learning solutions. We propose the Multi-scale Health-state Variational Auto-Encoder (MHealthVAE) to learn medically informative patient representations and allow meaningful subgroup detection from sparse EHRs. We derive a novel training objective to better capture health information and temporal trends into patient embeddings and introduce new performance metrics to evaluate the clinical relevance of patient clustering results.

1 INTRODUCTION

Time series is a common data modality in medical applications (Sun et al., 2020). Analyzing and clustering patient trajectories is crucial to understanding their health state, estimating their prognosis, and determining optimal treatment. In fact, doctors implicitly perform such analysis: given historical records, they match patients to the most similar cohort for the most suitable medicines and treatments. Our goal is to perform this patient subgroup analysis from a data-driven approach.

Clustering or learning representations of time series are well-studied tasks, with recent works on time series k-means (Astakhova et al., 2015), dynamic time warping (Giannoula et al., 2018), and deep learning-based methods (Ma et al., 2019b) for time series analysis. Still, medical time series pose additional challenges due to their often sparse nature, with many variables missing and evolving over different timescales (Sun et al., 2020). These characteristics of EHR challenge conventional representation learning methods, making it even harder to provide medically informative and clustering-friendly embeddings. Additionally, there are no well-established metrics to quantitatively evaluate patient clustering results and their medical meaningfulness.

Our contributions are as follows: 1) we propose the Multi-scale Health-state Variational Auto-Encoder (MHealthVAE), a novel scalable representation learning architecture for sparse medical time series; 2) we introduce the masked multi-scale reconstruction loss that helps learn medically informative embeddings; 3) we propose new clinically-relevant metrics to evaluate clustering results. Our results show that MHealthVAE improves clustering performance over prior work.

2 RELATED WORKS

Extensions of conventional clustering methods for time-series typically propose different similarity measures for the space of time series, including Euclidean distance (Javed et al., 2020) or dynamic time warping (Giannoula et al., 2018; Xing et al., 2010). Scalable variants of DTW include miniDTW (Cai et al., 2021) and fastDTW (Salvador & Chan, 2007). These metrics allow to cluster trajectories using conventional algorithms such as K-means (MacQueen, 1967), hierarchical clustering (Aghabozorgi et al., 2015; Das et al., 2008), or probabilistic approaches (Rigon et al., 2020). Although these methods are easy to implement, they are not designed to deal with data missingness and heterogeneous timescales of variation in multivariate time-series (Javed et al., 2020).

Deep-learning-based approaches are often more scalable in terms of sequence length and input dimensionality (Alqahtani et al., 2021). These models map time-series to lower-dimensional spaces for clustering. Unsupervised embeddings of time-series can be obtained with variational autoencoders (VAE) (Kingma & Welling, 2013; Fortuin et al., 2018) or recurrent networks such as LSTM (Staudemeyer & Morris, 2019), VaDER (de Jong et al., 2019) and transformers (Zerveas et al., 2020; Vaswani et al., 2017). Training can be further regularized with different clustering-friendly objec-

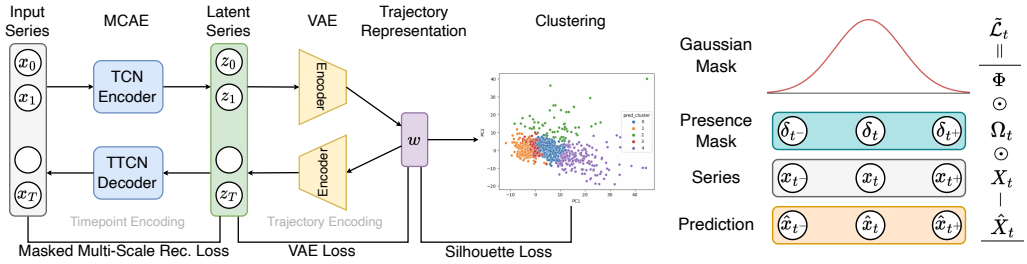
tives (Fortuin et al., 2018; Ma et al., 2019a), or by leveraging weak supervision with contrastive loss (Yèche et al., 2021). Such embedding-based methods are often efficient, flexible, and address the limitations of classical approaches but remain brittle to data missingness. In the following, we present an approach to obtain clustering-friendly and medically-informative representations of patient trajectories. Further details on related works in included in Appendix A.

3 METHODS

Consider the EHR of a group of patients where each record contains hundreds of distinct variables. We denote the acquired sequence for each patient as $X = \{x_1, \cdots, x_T\}, x_t \in \mathbb{R}^p$, where p is the number of variables. Data processing details are provided in Appendix C.1.

3.1 ARCHITECTURE DESIGN

The MHealthVAE consists of a Multi-scale Convolutional Auto-Encoder (MCAE) and a full-trajectory VAE. As illustrated in Figure 1a, this model performs a two-phase representation learning: first timepoint representation and then sequence representation. Clustering and downstream tasks are performed on these embeddings. Implementation details are included in Appendix C.2.



(a) MHealthVAE structure.

(b) Multi-scale reconstruction loss.

Figure 1: MHealthVAE workflow and its masked multi-scale reconstruction loss.

Timepoint Representation. The timepoint embedding learns a sequence of latent health states capturing the patient's current status, recent past, and future trends. The MCAE uses a dilated causal temporal convolutional network (TCN) as its encoder (Bai et al., 2018), mapping $X \in \mathbb{R}^{T \times p}$ to a sequence of timepoint representations $Z \in \mathbb{R}^{T \times q}$. The causal structure of this architecture ensures each timepoint representation z_t only accesses past information, allowing computation in an online manner at test time. In addition, the different kernels and the dilation patterns of the TCN architecture fit different weights and timescales of variation for each input variable.

The decoder of MCAE is a transpose TCN that maps each z_t to reconstruct a k-hour neighborhood centered around x_t, as illustrated in Figure 5c. By reconstructing this sequence $\{x_{t^-}, \cdots, x_{t^+}\}$, where $t^- = t - k/2$ and $t^+ = t + k/2$, from z_t, we enforce z_t to capture a recent history and to be predictive a near future. As a compact representation of the local patient state, the latent sequence $\{z_t\}_{t=0}^T$ can also be used in downstream tasks such as organ failure prediction (see Appendix D).

Sequence Representation and Clustering. Before clustering, a VAE maps the latent sequence $\{z_t\}_{t=0}^T$ to a full-trajectory representation $w \in \mathbb{R}^m$. We propose different regularizations to ensure this sequence-embedding space is medically meaningful and clustering-friendly. Clustering is then done in this space using K-means (MacQueen, 1967), assigning each patient a subgroup where we expect them to share high-level similarities such as length of stay (LOS), survival rate, and trajectory trends.

3.2 OBJECTIVE FUNCTION

Our training objective is designed to jointly train the two-stage representation learning pipeline for optimal clustering performance. We formulate our loss function as $\mathcal{L} = \mathcal{L}_{tcn} + \lambda \mathcal{L}_{vae} + \beta \mathcal{L}_{sil}$, where $\mathcal{L}_{tcn}$ is a masked multi-scale reconstruction loss, $\mathcal{L}_{vae}$ is the full trajectory ELBO, and $\mathcal{L}_{sil}$ is a clustering regularizer (all defined below). During training, hyperparameters $\{\alpha, \beta\}$ are scheduled to shift emphasis on different model elements. More details are included in Appendix C.3

The timestep representation learning model maps z_t to a sequence $\{\hat{x}_{t^-}, \ldots, \hat{x}_t, \ldots, \hat{x}_{t^+}\}$. We design our loss such that the reconstruction $\hat{x}_\tau$ is more accurate for a small $|\tau - t|$, by scaling the reconstruction loss of each term by a Gaussian weight $\Phi(\tau) = \mathcal{N}(\tau; t, \sigma)$. We set $6\sigma = k$ such

that the 99% percentile covers the full k-hour window. Additionally, to modulate for missingness in the trajectory, we also multiply our loss by a presence mask $\Omega(t) = \{\alpha \text{ if } x_t \text{ imputed}, 1 \text{ else}\}$, such that reconstructions of true and imputed values are scaled by 1 and hyperparameter $\alpha \in (0, 1]$ respectively. Overall, this defines our masked timepoint reconstruction loss $\mathcal{L}_{tcn}$ as follows, for a given patient trajectory:

$$\mathcal{L}_{tcn} = \sum_t \frac{1}{\Sigma_\Omega(t)} \sum_{\tau=-k/2}^{k/2} \Omega(t+\tau)\Phi(\tau)|x_{t+\tau} - \hat{x}_{t+\tau}(z_t)|^2 \tag{1}$$

where $\Sigma_\Omega(t) = \sum_\tau \Omega(t+\tau)$ is a normalizing term.

The full trajectory representation learning loss $\mathcal{L}_{vae}$ is the traditional ELBO loss (Kingma & Welling, 2013): $\mathcal{L}_{vae} = \mathbb{E}_{q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}(q_\phi(Z|X)||p_\theta(Z))$. We amend it such that the reconstruction part, $\log p_\theta(X|Z)$, is only computed on x_t and z_t for t less than the patient's LOS. This prevents the VAE from learning on padded time points.

Finally, to ensure the full-trajectory embedding to be clustering-friendly, we added an additional Silhouette loss for regularization. The Silhouette score quantifies how good a clustering result is based on intra-cluster and inter-cluster distances (Rousseeuw, 1987). Since higher Silhouette score means better clustering, the regularizor is defined to be $\mathcal{L}_{sil} = -\mathcal{S}$, where $\mathcal{S}$ is the Silhouette score.

4 Experimental Details

Datasets. Our experiments concern two datasets: one is the high-resolution ICU dataset (HiRID, Hyland et al. (2020)), where 12 most available and useful variables are selected for a patient cohort undergone cardiac surgeries. The other is a synthetic dataset that is designed to be of the same variables and the same missingness as HiRID. This dataset is constructed to be of 5 clusters where each subgroup share similar trends in variables. More details are included in Appendix B.

Comparison Baselines. Using K-Means, we compared MHealthVAE with some baseline models including clustering on the LOS, on the raw data, and using DTW. To demonstrate the impact of our novel loss function, we also studied MHealthVAE with plain reconstruction loss (denoted TCN+AE in Table 1a). We additionally experiment with recurrent models including LSTM, VaDER, and Transformers; however, due to the huge proportion of missingness (15% available), these models do not bring advantages in clustering. See Appendix D for more information.

Evaluation Metrics. Traditional clustering metrics such as K-means loss (Ma et al., 2019a) and Silhouette scores (Rousseeuw, 1987) are not inforamtive about patients' health states or prognosis. Hence, we introduce three metrics to evaluate patient clustering: 1) *LOS difference*: generally, patients' health state is correlated with their LOS; this metric quantifies LOS distribution differences across clusters. 2) *Survival Rate Difference*: survival rate is a crucial indicator of prognosis; this metric measures survival rate differences across clusters over a certain period τ. 3) *Trajectory Difference*: it characterizes how different trajectories of each cluster are. In the synthetic dataset, we also evaluated the clustering accuracy and pairwise clustering accuracy. More detailed definitions of these metrics are in Appendix C.4.

5 Experimental Results

5.1 Synthetic Dataset Clustering

As shown in Table 1a, only using LOS for clustering renders poor performance in all metrics except for LOS difference, which is also beyond the true metric. This shows LOS difference itself is not enough for evaluation although it is informative from a medical perspective. Clustering using pure DTW distance or VaDER gives relatively poor performance. With the employment of MHealthVAE architecture brings noticeable improvement of 8% compared to clustering on raw trajectories. Adding the multi-scale reconstruction loss, we are able to obtain almost full identification of all clusters. Note that except clustering on the LOS, Table 1a demonstrates good correlation between clustering accuracy and the three metrics that we proposed. It also shows the advantage of MHealthVAE in correctly grouping patient subclasses despite the high missingness.

5.2 Real-world Data Clustering

With hyper-parameter tuning (see Appendix D), we clustered patients in HiRID into 5 groups. As shown in Table 1b, clustering on LOS renders good metrics although at the cost of putting $> 99\%$ of

Table 1: Clustering performance results, in comparison to baselines based on DTW (Giannoula et al., 2018), TCN (Bai et al., 2018) and VaDER (de Jong et al., 2019).

(a) Synthetic dataset.

	Accuracy (%)	Pairwise Acc. (%)	Silhouette	LOS Diff.	Traj. Diff.
Ground Truth	-	-	0.20	5.53	122.09
LOS	48.6 ± 1.4	72.2 ± 0.1	-0.04 ± 0.01	$\mathbf{9.4} \pm \mathbf{0.01}$	58.7 ± 0.70
Raw	86.2 ± 4.7	84.6 ± 3.5	0.05 ± 0.02	4.2 ± 0.07	57.9 ± 2.8
DTW	48.2 ± 0.1	68.9 ± 0.1	0.01 ± 0.01	1.67 ± 0.01	118.98 ± 0.10
TCN + AE	94.0 ± 0.9	95.8 ± 1.2	0.14 ± 0.01	3.94 ± 0.29	112.98 ± 2.48
VaDER	41.6 ± 0.1	48.2 ± 0.1	-0.06 ± 0.01	3.79 ± 0.01	40.17 ± 0.13
MHealthAE (Ours)	$\mathbf{99.1} \pm \mathbf{0.1}$	$\mathbf{99.1} \pm \mathbf{0.2}$	$\mathbf{0.19} \pm \mathbf{0.01}$	5.27 ± 0.38	$\mathbf{118.25} \pm \mathbf{3.66}$

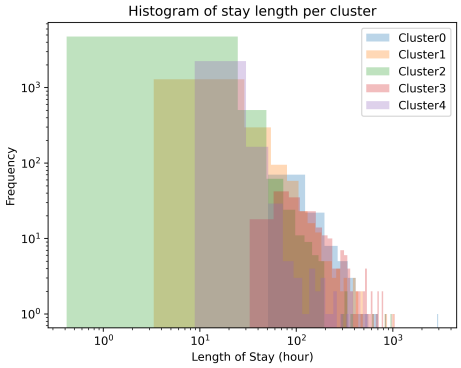
(b) HiRID dataset (Hyland et al., 2020).

	LOS Diff.	Silhouette	Surv. 3m (%)	Surv. 1y (%)	Surv. 5y (%)	Traj. Diff.
LOS	10.85	-	20.97	23.49	24.78	37.92
Raw	1.22	0.04	12.24	13.97	33.65	44.28
DTW	1.27	-	6.39	8.56	10.32	28.28
TCN + AE	0.2	-0.04	1.59	1.58	15.55	19.93
VaDER	0.03	0.65	1.01	1.31	1.24	2.23
MHealthAE (Ours)	3.71	0.014	18.47	20.91	**26.47**	**48.51**

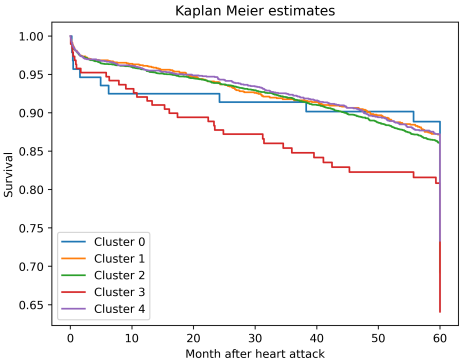
patients (LOS below 2 days) into one cluster. It is true that patients who stay longer than 3 days have higher death rates; however, this clustering fails to identify subgroups in the majority of patients. Excluding clustering on LOS, our model displays better statistics in all metrics.

Apart from good statistics, MHealthVAE renders subgroups of medical differences. Figure 2b demonstrates prognosis differences between groups: cluster 3 is the sickest group where patients are continuously lost in the first 5 years. Cluster 0 are ones that are initially sick yet recovered after 1 year. Although survival curves do not distinguish the other three healthier groups, inspections of mean cluster trends demonstrate differences in many variables: the trends in dimensions such as heart rate, lactate level, and arterial pressure agrees with the health conditions in each subgroup (see Appendix D). Additionally, each cluster also has slightly different LOS distributions (see Figure 2a). These all indicate that MHealthVAE indeed picked up health-state information in its embedding w.

Besides full-trajectory embedding being medically informative, the latent sequence $\{z_t\}_{t=0}^{T}$ also encodes health states: prediction of circulatory failure using $\{z_t\}_{t=0}^{T}$ renders higher accuracy, AUPRC, and recall compared to prediction using raw sequence (see Appendix D). Generally, our model is capable of inferring health-related information (survival rate included) even if the model is not explicitly given those inputs.



(a) Length of stay across clusters.



(b) Survival rate across clusters.

Figure 2: Cluster visualizations.

5.3 Conclusion

The MHealthVAE learns meaningful patient representations and clusters from medical time-series, even under high missingness and with heterogeneous timescales of variation. In future work, we hope to explore more distributions in the masked multi-scale reconstruction loss, and to disentangle trajectory representations in terms of human interpretable dimensions. This work should help build accurate, interpretable and reliable models for medical applications.

REFERENCES

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16–38, 2015. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2015.04.007. URL https://www.sciencedirect.com/science/article/pii/S0306437915000733.

Ali Alqahtani, Mohammed Ali, Xianghua Xie, and Mark W. Jones. Deep time-series clustering: A review. *Electronics*, 10(23), 2021. ISSN 2079-9292. doi: 10.3390/electronics10233001. URL https://www.mdpi.com/2079-9292/10/23/3001.

N. N. Astakhova, Liliya A. Demidova, and Evgeny V. Nikulchev. Forecasting method for grouped time series with the use of k-means algorithm. *ArXiv*, abs/1509.04705, 2015.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018. URL http://arxiv.org/abs/1803.01271.

Borui Cai, Guangyan Huang, Najmeh Samadiani, Guanghui Li, and Chi-Hung Chi. Efficient time series clustering by minimizing dynamic time warping utilization. *IEEE Access*, 9:46589–46599, 2021. doi: 10.1109/ACCESS.2021.3067833.

Swagatam Das, Ajith Abraham, and Amit Konar. Automatic clustering using an improved differential evolution algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1):218–237, 2008. doi: 10.1109/TSMCA.2007.909595.

Johann de Jong, Mohammad Asif Emon, Ping Wu, Reagon Karki, Meemansa Sood, Patrice Godard, Ashar Ahmad, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience*, 8(11), 11 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz134. URL https://doi.org/10.1093/gigascience/giz134. giz134.

Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Som-vae: Interpretable discrete representation learning on time series, 2018. URL https://arxiv.org/abs/1806.02199.

Alexia Giannoula, Alba Gutiérrez-Sacristán, Àlex Bravo, Ferran Sanz, and Laura I Furlong. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. *Scientific Reports*, 8, 03 2018. doi: 10.1038/s41598-018-22578-1.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3): 364–373, Mar 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0789-4. URL https://doi.org/10.1038/s41591-020-0789-4.

Ali Javed, Byung Suk Lee, and Donna M. Rizzo. A benchmark study on time series clustering. *Machine Learning with Applications*, 1:100001, 2020. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2020.100001. URL https://www.sciencedirect.com/science/article/pii/S2666827020300013.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Qianli Ma, Jiawei Zheng, Sen Li, and Gary W Cottrell. Learning representations for time series clustering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/1359aa933b48b754a2f54adb688bfa77-Paper.pdf.

Qianli Ma, Jiawei Zheng, Sen Li, and Gary W Cottrell. Learning representations for time series clustering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/1359aa933b48b754a2f54adb688bfa77-Paper.pdf.

J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.

Jangho Park, Juliane Muller, Bhavna Arora, Boris Faybishenko, Gilberto Pastorello, Charuleka Varadharajan, Reetik Sahu, and Deborah Agarwal. Long-term missing value imputation for time series data using deep neural networks, 2022. URL https://arxiv.org/abs/2202.12441.

Tommaso Rigon, Amy H. Herring, and David B. Dunson. A generalized bayes framework for probabilistic clustering, 2020. URL https://arxiv.org/abs/2006.05451.

Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987. doi: 10.1016/0377-0427(87)90125-7.

Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, oct 2007. ISSN 1088-467X.

Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019. URL https://arxiv.org/abs/1909.09586.

Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data, 2020. URL https://arxiv.org/abs/2010.12493.

Nenad Tomašev, Xavier Glorot, Jack Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton Baker, Kelly Peterson, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572:116, 08 2019. doi: 10.1038/s41586-019-1390-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

John Weldon, Tomas Ward, and Eoin Brophy. Generation of synthetic electronic health records using a federated gan, 2021. URL https://arxiv.org/abs/2109.02543.

Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, 12(1):40–48, nov 2010. ISSN 1931-0145. doi: 10.1145/1882471.1882478. URL https://doi.org/10.1145/1882471.1882478.

Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. 2021. doi: 10.48550/ARXIV.2106.05142. URL https://arxiv.org/abs/2106.05142.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning, 2020. URL https://arxiv.org/abs/2010.02803.

Type	Method	Objective	Modeling	Pros	Cons
Time Series Clustering	DTW	Evaluate time series similarities	Defined DTW metric on time series	Simple implementation	Slow computation
	VaDER	Learn a variational latent space for clustering	Use implicit imputation and VAE to obtain the latent space for clustering	Deal with missing values	Potential degradation if data is very sparse.
	DTCR	Learn a clustering friendly embedding	Use an auto-encoder augmented by fake-sample classification and K-Means loss	Clustering friendly latent space	The auxiliary classification may not be medically informative
Representation Learning	SOM-VAE	Learn a latent space that smoothly characterize health transitions.	Train auto-encoder that is regularized by self-organizing map.	Interpretable representation of patients	No explicit handling of missing values.
	NCL	Learn representation that boosts online monitoring	Used neighborhood contrastive learning for time point classification	Useful for event prediction	Representation not primarily targeted for clustering

Figure 3: Overview of works related to medical time series clustering and representation learning including dynamic time warping (Giannoula et al., 2018), VaDER (de Jong et al., 2019), DTCR (Ma et al., 2019b), SOM-VAE (Fortuin et al., 2018), and contrastive loss (Yèche et al., 2021).

A Additional related works

In section 2, we mentioned some works related to medical time series clustering. Figure 3 is a summarized analysis of the five most relevant methods. Despite their respective advantages or disadvantages, the performance of all methods above can be largely impaired by the presence of missing values. There are many ways that help impute missing values: forward filling, using additional imputing neural network (de Jong et al., 2019; Park et al., 2022), or taking presence features as additional input to the representation learning networks (Tomašev et al., 2019). Still, inherent missingness in medical data makes it difficult to evaluate the real-world performance of different imputation techniques; the downstream unsupervised tasks such as clustering makes it even harder to evaluate the algorithm as a whole. To some extent, existing methods lack effective means to deal with sparsity and high missingness in EHR.

B Dataset Details

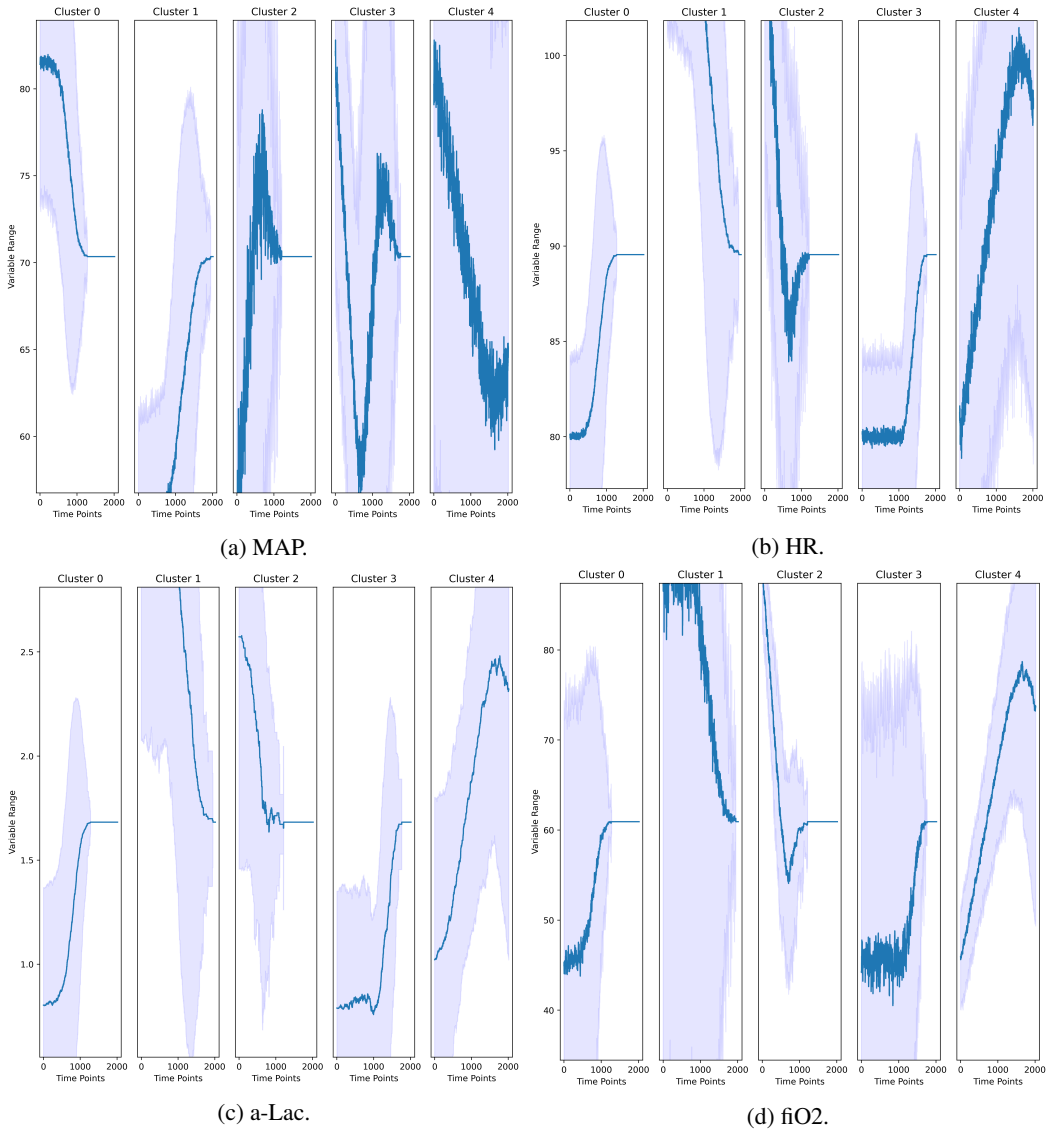
Table 2: Variables selected in HiRID dataset with their normal/abnormal range.

Variable Name	Abbreviation	Normal Range	Abnormal Range
Glasgow Comma Score	GCS	11 ± 2	5 ± 2
Heart Rate	HR	70 ± 10	110 ± 20
Mean Arterial Pressure	MAP	70 ± 10	50 ± 5
Carbon Monoxide	CO	5 ± 0.5	3 ± 0.3
Venous Oxygen Saturation	SVO2	0.7 ± 0.1	0.5 ± 0.2
Central Venous Pressure	ZVD	8 ± 2	15 ± 3
Arterial Lactate	a-Lac	1 ± 1	10 ± 5
Oxygen Saturation	SPO2	0.95 ± 0.05	0.8 ± 0.07
Respiratory Rate	Res Rate	16 ± 2	32 ± 5
Fraction of Inspired Oxygen	fiO2	0.21 ± 0.04	0.8 ± 0.2
Base Excess	a-BE	0 ± 2	-10 ± 5
Hemoglobin	Hb	14.5 ± 2	8 ± 4

HiRID Dataset. Our experiments are based on two datasets, a HiRID dataset of real patients' stays in ICU (Hyland et al., 2020). More specifically, we selected a subset of 10018 patients who had just undergone cardiac surgeries. Among all the variables of their EHR, we selected 12 variables that are most available and also most indicative of a patient's health conditions such as heart rate, SPO2, and Glasgow comma scale. More details on the selected variables are listed in Table 2. The length of stay (LOS) of these patients varies much, where healthier ones usually stay in ICU for no longer than 24 hours while sicker ones could stay for weeks.

Synthetic Dataset. In addition to running experiments on this real-world HiRID dataset, we constructed a synthetic dataset with the same variables as the HiRID dataset. Synthetic patients are

generated with the same LOS distribution; the same degree of missingness in each variable is preserved as in the real dataset. Moreover, by consulting medical experts, each variable's scale and variance reflect a patient's health conditions: we assign abnormally high or low values in each variable to indicate that a patient is suffering and assign normal values if a patient is in good condition (see Table 2). By constructing different transitions from healthy to unhealthy or vice versa, we designed 5 clusters of patients. More specifically, we designed cluster 0 to be a healthy group where patient LOS is shorter and trajectory starts within the normal range. Cluster 1 is the diseased group where LOS is higher and all variables are within the abnormal range (see Figure 4). Cluster 2 is a group that transitions from diseased to healthy, where patients' LOS is relatively short. Cluster 3 and 4 are patients who are initially healthy but got diseased later. Cluster 4 is different from cluster 3 in that the transition is slower and their health deterioration is severer. In Figure 4, we demonstrated the variable trends of 4 variables.



(a) MAP.

(b) HR.

(c) a-Lac.

(d) fiO2.

Figure 4: Synthetic dataset variable demo. Each cluster's mean trajectory and 95% confidence interval are plotted.

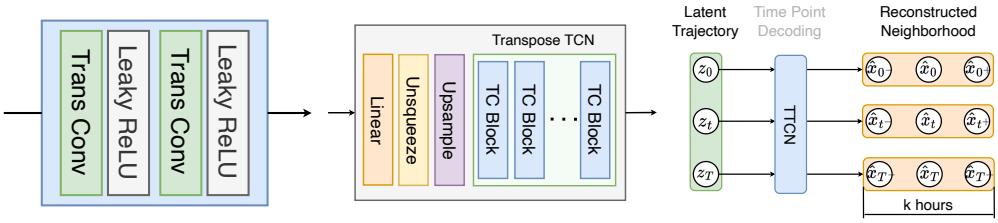
C IMPLEMENTATION DETAILS

C.1 DATA PREPROCESSING

The raw patient records require some elementary preprocessing to become usable for learning. Consider the EHRs of a group of patients where each record contains hundreds of distinct variables. Since these variables are of drastically different scales, they are first normalized per variable using the mean and variance computed on all available time points. As the measure of variables or test results could happen at arbitrary time, each variable are resampled temporally to 5 minute grids, where the data points that are missing are imputed by forward filling. Finally, these sequences are zero padded or cropped to a fixed length T of 7 days. Further data processing such as data augmentation (Weldon et al., 2021) and other types of imputation (de Jong et al., 2019) could be performed; we do not introduce them here.

C.2 ARCHITECTURE DETAILS

For the encoder structure of MCAE, we employed a 5-layer dilated causal temporal convolutional network as introduced by Bai et al. (2018). We set the dilation factor as 2 and a kernel size of 12. As for the decoder, we used a transpose TCN. The specific architecture is demonstrated in Figure 5. The basic building block is transpose convolutional block that consists of two transpose convolution operation with Leaky ReLU as activation (see Figure 5a). To upsample a latent embedding, Transpose TCN first use a linear layer to map time point representation z_t to larger dimensions. Then, it unsqueeze and upsample the output to be of same dimension as the original sequence $\{x_t\}_{t=0}^{T}$. 5-layer of transpose convolutional block follows, mapping the upsampled sequence to a reconstruction (see Figure 5b).



(a) Transpose conv block. (b) Transpose TCN structure. (c) Time point decoder pipeline.

Figure 5: Transpose TCN architecture

As for the full-trajectory VAE, we employ a simple 5-layer Bayesian VAE (Kingma & Welling, 2013). Note that the latent sequences $\{z_t\}_{t=0}^{T}$ are flattened before they are used by VAE.

C.3 TRAINING DETAILS

Hyperparameters. There are several hyper-parameters related to the model. The first is the presence mask parameter α that mask out the missing dimensions at different time point. For our experiment, by hyperparameter tuning on a designed synthetic dataset, we set $\alpha = 0.2$. We set $k = 24$ in the timepoint representation learning pipeline, so that each z_t is trained to reconstruct a 24-hour window $\hat{N}_{24}(x_t) \in \mathbb{R}^{\Delta T_{24} \times p}$ of the original time series. Another parameter is the Gaussian distribution used in loss calculation (see Equation (1)). By parameter tuning, the variance σ was set such that $6\sigma = k$ correspond to 24 hour.

Loss Computation. During training, the Silhouette loss is computed as part of the loss function. Note that this Silhouette score is not computed in for batches, but computed for the full dataset's clustering. This could be computationally intensive if the dataset size is large. To overcome this complexity, one could compute the expected Silhouette loss on random subsets of a fixed size. Since our HiRID dataset is relatively small, where we focused on 10,000 patients undergone cardiac surgeries, this loss computation is still feasible.

Loss Scheduling. The training is scheduled to be of two phases. In the first part, we jointly train MCAE and the full-trajectory VAE using the full loss function $\mathcal{L} = \mathcal{L}_{tcn} + \lambda\mathcal{L}_{vae} + \beta\mathcal{L}_{sil}$. This ensures that the latent sequence $\{z_t\}_{t=0}^{T}$ encodes health state information and could also be well-separated in full-trajectory embedding space. In the second phase, we fix the parameters in MCAE and only train the VAE. Here, to avoid weights of VAE being stuck in local minimas, we re-initialize the weights of VAE and train on loss $\mathcal{L}_{vae} + \beta\mathcal{L}_{sil}$.

C.4 Experimental Details

In Section 4, we mentioned three metrics that we used to evaluate the medically-meaningfulness of clustering results. We formally define these metrics as follows.

LOS Difference: By splitting the patients' LOS range into 50 equally sized bins, we are able to calculate the empirical LOS distribution of a cluster $c \in \mathcal{C}$. Denote this empirical LOS distribution as $\hat{P}_c$. We hence define the LOS difference as

$$D_{LOS}(\mathcal{C}) := \frac{2}{|\mathcal{C}| \times (|\mathcal{C}| - 1)} \sum_{c \neq c' \in \mathcal{C}} KL(\hat{P}_c || \hat{P}_{c'}) \tag{2}$$

Survival Rate Difference: Assume cluster $c \in \mathcal{C}$ has a survial rate up until time τ of $p_c^{(\tau)}$. The survival rate difference is defined as

$$D_{surv}(\mathcal{C}) := \text{median}_{c \neq c' \in \mathcal{C}} |p_c^{(\tau)} - p_{c'}^{(\tau)}| \tag{3}$$

Trajectory Difference: for each cluster c, a mean trajectory X_c is calculated. Then, the trajectory difference is defined as the trimmed mean over all $|X_c - X_{c'}|, c, c' \in \mathcal{C}, c \neq c'$. This metric reflects how far the clusters are separated from each other.

Pairwise Clustering Accuracy: given ground truth cluster labels, a pair of patients is pairwise correctly clustered if they are assigned to the same cluster when they have the same ground truth label; they are also correctly clustered if they are assigned different labels when their ground truth labels are different. The pairwise clustering accuracy is the proportion of correctly clustered patient pairs.

D Additional Results

D.1 HiRID Clustering Results

In Section 5, we mentioned that we implemented some recurrent structures including LSTM and Transformers. These results are not listed in the main text since recurrent networks cannot learn too much useful information in a dataset where 85% of data are forward-filled values. Since little variance is present in the trajectories, without proper incentives, LSTMs and Transformers do not render good clustering results under unsupervised settings. The complete result is listed in Table 3.

As mentioned in Section 5.2, the number of clusters, 5, used in HiRID patient clustering is a result of hyper-parameter tuning. Table 4 is a comparison of clustering metrics given different number of clusters. 5 clusters is better at most metrics.

As shown in Section 5.2, MHealthVAE performs well in discovering subgroups within the cardiac surgery patient cohort. A cluster population distribution is plotted in Figure 6a. Here, we see that MHealthVAE identifies two subgroups of smaller populations: cluster 0 and cluster 3 are the ones of sicker patients. For the healthier clusters, cluster 1, 2, and 4, although no noticeable differences were displayed from survival rate (see Figure 2b), we do notice distinct trends in their health-related variables. Take arterial lactate (a-Lac) as an example, Figure 7c illustrates different trends of a-Lac in the 5 identified subgroups. For cluster 0 and 3 (sicker patients), the lactate level is higher than any other subgroups, indicating that the patients are undergoing circulatory issues after their cardiac surgeries. This partially explains the high death rate seen in the two groups. For healthier groups, we notice lower lactate level, which is in concordance with their quick recovery and good prognosis outcomes. Other variables of different clusters also display trend differences as shown in Figure 7.

Table 3: Clustering performance results. DTW (Giannoula et al., 2018), TCN (Bai et al., 2018), VaDER (de Jong et al., 2019)

(a) Synthetic dataset.

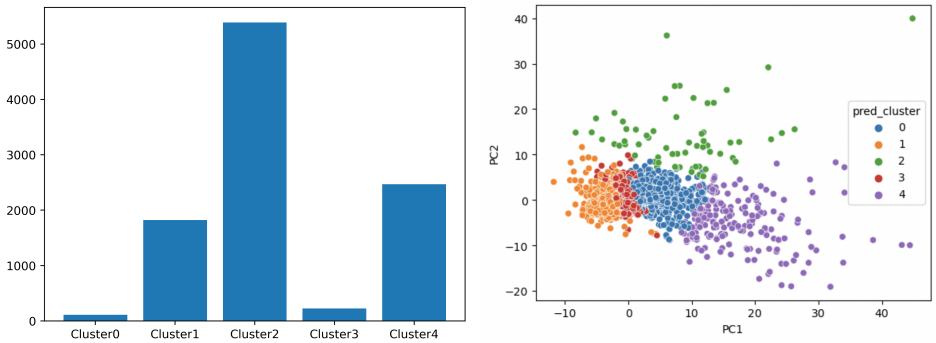
	Accuracy (%)	Pairwise Acc. (%)	Silhouette	LOS Diff.	Traj. Diff.
Ground Truth	-	-	0.20	5.53	122.09
LOS	48.6 ± 1.4	72.2 ± 0.1	-0.04 ± 0.01	$\mathbf{9.4 \pm 0.01}$	58.7 ± 0.70
Raw	86.2 ± 4.7	84.6 ± 3.5	0.05 ± 0.02	4.2 ± 0.07	57.9 ± 2.8
DTW	48.2 ± 0.1	68.9 ± 0.1	0.01 ± 0.01	1.67 ± 0.01	118.98 ± 0.10
AE Latent	86.7 ± 0.7	85.7 ± 1.7	0.07 ± 0.04	5.06 ± 0.47	108.5 ± 5.2
TCN + AE	94.0 ± 0.9	95.8 ± 1.2	0.14 ± 0.01	3.94 ± 0.29	112.98 ± 2.48
LSTM + AE	45 ± 2.9	50.9 ± 3.3	-0.08 ± 0.01	1.43 ± 1.47	33.37 ± 5.76
VaDER	41.6 ± 0.1	48.2 ± 0.1	-0.06 ± 0.01	3.79 ± 0.01	40.17 ± 0.13
Transformer + AE	97.8 ± 0.8	97.9 ± 1.3	0.17 ± 0.01	5.53 ± 0.04	117.25 ± 0.53
MHealthAE (Ours)	$\mathbf{99.1 \pm 0.1}$	$\mathbf{99.1 \pm 0.2}$	$\mathbf{0.19 \pm 0.01}$	5.27 ± 0.38	$\mathbf{118.25 \pm 3.66}$

(b) HiRID dataset (Hyland et al., 2020).

	LOS Diff.	Silhouette	Surv. 3m (%)	Surv. 1y (%)	Surv. 5y (%)	Traj. Diff.
LOS	10.85	-	20.97	23.49	24.78	37.92
Raw	1.22	0.04	12.24	13.97	33.65	44.28
DTW	1.27	-	6.39	8.56	10.32	28.28
AE Latent	0.45	0.018	6.56	7.26	5.05	46.09
TCN + AE	0.2	-0.04	1.59	1.58	15.55	19.93
LSTM + AE	0.03	0.33	0.87	1.07	3.69	1.53
VaDER	0.03	0.65	1.01	1.31	1.24	2.23
Transformer + AE	0.29	0.728	0.61	0.83	1.36	3.25
MHealthAE (Ours)	3.71	0.014	18.47	20.91	**26.47**	**48.51**

Table 4: Cluster performance for different number of clusters based on MHealthVAE.

N Cluster	LOS Diff.	Silhouette	Surv. 3m (%)	Surv. 1y (%)	Surv. 5y (%)	Traj. Diff.
3	1.63	**0.88**	16.25	16.43	16.51	13.61
4	**4.07**	0.76	10.73	11.84	20.16	2.24
5	3.71	0.014	**18.47**	**20.91**	**26.47**	**48.51**
6	2.80	0.44	11.76	11.58	12.19	28.73



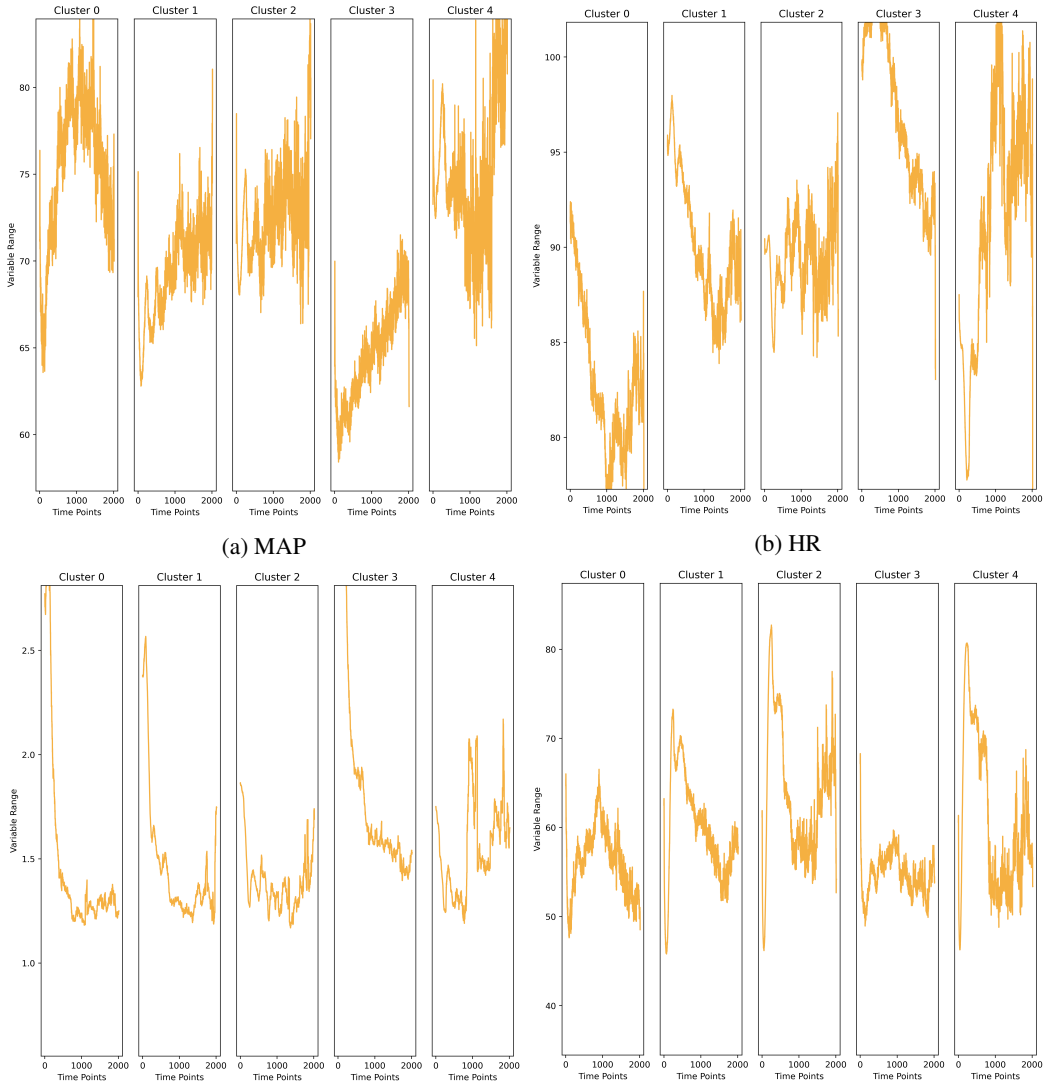
(a) Cluster population distribution.

(b) Visualization of full-trajectory embedding.

Figure 6: MHealthVAE cluster visualizations.

This proves that the model is indeed clustering patients in a medically meaningful and interpretable manner.

Figure 6b is a visualization of full trajectory embeddings of patients projected onto two PCA components. These clusters appear to be compact yet well separated in the latent space. This is a result of the two regularizers that we introduced: a KL regularization in $\mathcal{L}_{vae}$ for the compactness and a Silhouette loss for the well-separateness.



(a) MAP (b) HR

(c) a-Lac. (d) fiO2.

Figure 7: Visualization of common subgroups found by different methods. The four variables' cluster mean and 95% confidence intervals are plotted for the first 7 days (2016 recorded data points) within ICU stay.

D.2 Circulatory Failure Prediction

Table 5: Circularatory failure prediction using different input sequences.

	Test Acc. (%)	AUC	Recall	AUPRC	Ave. Precision
Original Series	92.03 ± 0.72	0.929 ± 0.003	0.667 ± 0.011	0.798 ± 0.003	0.664 ± 0.006
TCN	90.74 ± 0.55	0.936 ± 0.003	$\mathbf{0.729 \pm 0.013}$	0.822 ± 0.005	0.672 ± 0.011
LSTM	$\mathbf{96.23 \pm 0.37}$	0.500 ± 0.000	0.500 ± 0	0.500 ± 0.000	0.500 ± 0.000
MHealthVAE (Ours)	93.10 ± 0.56	$\mathbf{0.948 \pm 0.001}$	0.723 ± 0.025	$\mathbf{0.831 \pm 0.010}$	$\mathbf{0.709 \pm 0.005}$

In addition to clustering on HiRID dataset to identify subtly different patient subgroups, we performed a downstream task, circulatory failure prediction, on the latent trajectory $\{z_t\}_{t=0}^{T}$ to show that it is a more compressed representation of the original patient trajectories. We first trained three auto-encoders based on TCN, LSTM, and MHealthVAE to encode the original trajectories into latent

sequences (5 dimensions per timepoint). These sequences are passed to an MLP classifier to predict if circulatory failure is about to happen. We compared the results to classification using the original trajectory (12 variables per timepoint) as a baseline. A comparison of the classification results is shown in Table 5. Comparing the original series and encoded sequence, we see that MHealth-VAE encoded latent sequence brings an advantage in failure prediction in terms of all metrics. This shows that our encoded sequence is a more compressed representation of a patient's health condition without much loss of information.

Note that a circulatory failure is a rare event: only 4% of the time points are positive for failures. In particular, the MLP classifies every timepoint from the LSTM-encoded sequence as negative of failures, rendering the best accuracy and the worst AUPRC. By inspection, we noticed that the latent sequence embedded by LSTM is of little variation. We believe this is because of the huge forward-filling used in EHR imputation: since only 15% of the time points are available, the sequences are mostly filled with padding values, leaving little variation for the LSTM to learn. One can train recurrent structures including LSTM and Transformers better in a supervised setting; yet, in an unsupervised way, our MHealthVAE brings an embedding that is generic to downstream tasks.