Learning Unsupervised Representations for ICU Timeseries

Addison Weatherhead ADDISON.WEATHERHEAD@MAIL.UTORONTO.CA
University of Toronto & Vector Institute, The Hospital for Sick Children

Robert Greer ROBERT.GREER@SICKKIDS.CA
The Hospital for Sick Children

Michael-Alice Moga MICHAEL-ALICE.MOGA@SICKKIDS.CA
The Hospital for Sick Children

Mjaye Mazwi MJAYE.MAZWI@SICKKIDS.CA
The Hospital for Sick Children

Danny Eytan BILIARY.COLIC@GMAIL.COM
The Hospital for Sick Children

Anna Goldenberg ANNA.GOLDENBERG@UTORONTO.CA
University of Toronto & Vector Institute, The Hospital for Sick Children

Sana Tonekaboni STONEKABONI@CS.TORONTO.EDU
University of Toronto & Vector Institute, The Hospital for Sick Children

Abstract

Medical time series like physiological signals provide a rich source of information about patients' underlying clinical states. Learning such states is a challenging problem for ML but has great utility for clinical applications. It allows us to identify patients with similar underlying conditions, track disease progression over time, and much more. The challenge with medical time series however, is the lack of well-defined labels for a given patient's state for extended periods of time. Collecting such labels is expensive and often requires substantial effort. In this work, we propose an unsupervised representation learning method, called TRACE, that allows us to learn meaningful patient representations from time series collected in the Intensive Care Unit (ICU). We show the utility and generalizability of these representations in identifying different downstream clinical conditions and also show how the trajectory of representations over time exhibits progression toward critical conditions such as cardiopulmonary arrest or circulatory failure.

Data and Code Availability This paper uses the HiRID ICU dataset (Hyland et al., 2020), which is available on the PhysioNet repository [1]. The proposed model is also evaluated on an ICU dataset from The Hospital for Sick Children in Toronto Canada, which unfortunately is not publicly available. The code implementation of this work and all the experiments are made available here: https://github.com/Addison-Weatherhead/TRACE

1. Introduction

In the Intensive Care Unit (ICU), patients require continuous, close monitoring via devices that record and keep track of patient vital signs at all times. This results in large amounts of high frequency signals, including Electrocardiogram (ECG), Heart Rate (HR), respiratory rate (RR), arterial blood pressure (ABP), etc. The clinicians tending to the patients at the bedside use these time series data, together with ancillary information, to form a mental model of the patient state and guide treatments. However, the sheer volume of information generated from these signals can impose a cognitive load on clinicians leading to inefficient patient care. At the same time this complex data source presents an opportunity as a rich source of information and an ideal substrate for Machine Learning (ML) models. As a result, there's been an increase in research focused on developing data-driven ML models with promising results to assist ICU professionals in their practice and ultimately help improve delivery of care (Ong et al., 2012; Suresh

1. https://physionet.org/content/hirid/1.1.1/

et al., 2017; Razavian et al., 2016; Hyland et al., 2020; Gutierrez, 2020).

The physiological signals collected at the ICU are multivariate and highly non-stationary. They capture changes in the patient's underlying health condition over time and are therefore an informative source of data for ML models and clinicians alike. For both, learning informative representations of the underlying patient state in a lower dimensional encoding space may help with better understanding and modeling of the time series. Such lower dimensional representations of patient state resonate with the underlying physiological processes that generate the recorded signals and can serve as an ideal substrate to identify and track changes that happen in a patient's health condition by examining the trajectory of encodings over time. Moreover, these encodings can also be used for a variety of important downstream tasks without needing a complex model (Bengio et al., 2013). Examples of such tasks range from tracking patient trajectory throughout the disease course to identify need for interventions, diagnosis and prognostication, assessing treatment response, triaging patients according to severity, and even uncovering hitherto unknown distinct classes in heterogeneous clinical entities such as septic shock or acute respiratory distress syndrome.

The biggest challenge in learning general representations for rich time series datasets is lack of well-defined labels. Obtaining labels for the patient's state for extended periods of signals is expensive and often impractical; the underlying physiological state can be unknown or there can be lack of agreement between even experienced clinicians. This motivates using unsupervised representation learning frameworks for encoding information. Additionally, health datasets are often plagued by severe class imbalance, with a tiny subset of patients experiencing a particular clinical condition. In these situations, unsupervised methods are more favorable because they provide a more robust solution and are less prone to learning features only relevant to the dominant class (Liu et al., 2021).

In this work, we introduce an unsupervised representation learning framework that learns the underlying state of patients over time using high-frequency physiological signals collected at the bedside ICU. Our work builds on a previously developed Temporal Neighborhood Coding (TNC) framework (Tonekaboni et al., 2020) with substantial improvements to ensure that it is appropriate for the streaming signals in the ICU. First, we incorporate an encoder architecture that can handle long term dependencies as well as missing observations and measurements that are taken with different rates. Second, we introduce a novel way of identifying the temporal neighborhoods that is significantly more efficient and performs better in the presence of missing observations. With this new technique, we can also alleviate for sampling bias in the contrastive objective, without the need for an additional hyper-parameter or prior knowledge of the signal. Through a number of evaluations, we assess the usability and generalizability of our learned representations. We demonstrate that our approach encodes the informative parts of the signal in a lower dimensional representation, and that these encodings can be used for a number of downstream tasks such as predicting in-hospital mortality, cardiopulmonray arrest, and patient diagnostic, without the need of complex models. We further demonstrate that by tracking the representations over time, we can identify how high risk conditions evolve and appear in a patient's physiology. This is one of the major benefits of learning representations for complex medical time series, as it provides an opportunity to explore patient states and disease trajectories in a tangible lower dimensional space.

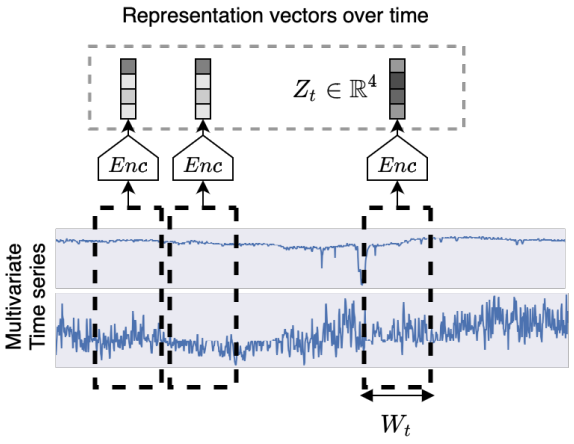

Figure 1: Overview of representation learning for time series window. Each encoder learns a representation vector Z_t for a window of time series W_t. The representations of consecutive windows over time, show the temporal evolution of the underlying states.

2. Related Work

ML for ICU ML has shown great utility in the ICU where an abundance of high frequency data is being collected for patients (Gutierrez, 2020). Many tools are designed to assist clinicians in identifying high-risk individuals in need of care. For instance models have been developed to predict intervention onset (Ghassemi et al., 2017; Suresh et al., 2017), disease onset (Razavian et al., 2016), circulatory failure (Hyland et al., 2020), cardiopulmonary arrest (Tonekaboni et al., 2018; Ong et al., 2012), sepsis (Nemati et al., 2018; Henry et al., 2015), or even unplanned unit transfer (Wellner et al., 2017). All these tools use information embedded in the medical time series in order to predict downstream events.

Representation learning in medicine: In healthcare, learning representation of temporal data can be very helpful for understanding patients' underlying health conditions. Most existing approaches for learning representations however are designed for specific downstream tasks. This means the representations are optimized to learn encodings that improve prediction of specific events (Choi et al., 2016a,b; Fiterau et al., 2017). Such methods improve performance significantly by extracting the informative parts of the rich and complex medical time series into lower dimensional encodings, but cannot be used as general representations. These representations are used for a variety of applications like identifying similar disease mechanisms (Lasko et al., 2013; Schulam et al., 2015), modelling disease progression (Wang et al., 2014; Alaa and van der Schaar, 2019), and multitask learning (Harutyunyan et al., 2019).

Unsupervised representation learning for time series: Recently there has been an increase in research on unsupervised representation learning methods, specifically designed for time series data. These methods belong to multiple general categories; for instance methods that use reconstruction objective for training, similar to Variational Auto Encoders (VAEs) that are commonly used in audio signals (Chorowski et al., 2019; Amiriparian et al., 2017). Other methods use measure of similarity to train the encoders (Lei et al., 2017; Ma et al., 2019; Madiraju et al., 2018), and more recent methods use different types of contrastive objectives for training (Oord et al., 2016; Franceschi et al., 2019; Tonekaboni et al., 2020; Hyvarinen and Morioka, 2016). All of these methods have shown great success on a variety of time series data.

3. Method

In this section, we introduce our unsupervised representation learning framework, called TRACE, designed to learn the underlying patient's states using medical timeseries collected in the ICU. We augment the TNC framework in a number of ways to improve the efficiency of the method and to make it suitable for the ICU setting.

3.1. Notation

Physiological signals collected from bedside monitors are in the form of multivariate time series. We represent each multivariate time series sample as $X^{(i)} \in \mathbb{R}^{D\times T}$, where i is the sample index, D is the number of features, and T is the time length of the sample. Note that the length of the time series can vary among samples depending on the patient's length of stay in the ICU. To deal with missing measurements, for each sample i, we define a binary mask $M^{(i)} \in \mathbb{R}^{D\times T}$ the same size as $X^{(i)}$ that indicates which input entries are measured (indicated by 1) and which one are missing (indicated by 0). For notational simplicity, the sample index (i) will be dropped throughout this paper.

We denote $W_t \in \mathbb{R}^{D\times\delta}$ to be a window of time series from sample X, centered at time t and with length δ. Our goal is to learn the representation $Z_t \in \mathbb{R}^L$ for each window, where L represents the size of the encoding and condenses the information in the window into a lower dimensional representation($L \ll D \times \delta$). Learning the representations of consecutive windows over time will allow us to track the state progression.

3.2. Background

Here, we provide a brief description of the TNC framework. At the heart of TNC is an encoder (Enc), typically a deep neural network, which takes a reference window W_t of time series and generates a vector representation $Z_t \in \mathbb{R}^L$, where L is the size of the encoding. The objective function (Eq. 1) is partly a contrastive learning objective that trains the signal encoder with a Discriminator ($\mathcal{D}$) that identifies representations of similar windows.

$$\mathcal{L}_{TNC} = -\mathbb{E}_{W_t \sim X}\Big[\mathbb{E}_{W_l \sim N_t}[\log \mathcal{D}(Z_t, Z_l)] + \mathbb{E}_{W_k \sim \bar{N}_t}\Big[(1-w)\log\left(1-\mathcal{D}(Z_t, Z_k)\right) + w \log \mathcal{D}(Z_t, Z_k)\Big]\Big] \quad (1)$$

This semantic similarity between windows is determined by the *temporal neighborhood* around a window W_t, which is defined as the region where the signals are relatively stationary and are therefore assumed to be generated from the same underlying state. The Augmented Dickey-Fuller (ADF) statistical test is used to determine the stationarity in the original TNC framework. Furthermore, the loss is weighted according to principles from Positive/Unlabeled (PU) learning to account for the potential sampling bias in the contrastive objective. This is to compensate for the fact that the negative samples, drawn from outside of the neighborhood, could in fact be similar to the reference window (in seasonal time series, for example). Further details on the framework can be found in Tonekaboni et al. (2020).

3.3. Proposed method: TRACE TempoRal AutoCorrelation Encoding - ICU

The TNC method provides a reasonable solution for learning representations for non-stationary time series. The method is based on the smoothness assumption of signals which holds well for the physiological signals, as changes in patient states often happen gradually. However, the method has a number of shortcomings that limit its usability in many settings, including the ICU. For instance, one of the biggest challenges of real world high-frequency time series, especially clinical data, is missingness and noise in the signals. The ADF test that is used to define the neighborhood range in the TNC framework is very sensitive to such artifacts and fails to find the appropriate size for a neighborhood, hindering the overall performance of the encoder. In our proposed TRACE method we have addressed a number of such issues resulting in a more efficient method with fewer hyperparameter to tune.

$$\mathcal{L} = -\mathbb{E}_{W_t \sim X}\Big[\mathbb{E}_{W_l \sim N_t}[\log(\mathcal{D}(Enc(W_t), Enc(W_l)))] + \mathbb{E}_{W_k \sim \bar{N}_t}[\log(\mathcal{D}(Enc(W_t), Enc(W_k)))]\Big] \quad (2)$$

We use the simplified contrastive objective in Eq. 2 to train our models. The encoder is a dilated causal convolutional neural network (Franceschi et al., 2020) that handles mutltivariate time series of variable length. The filters are exponentially dilated meaning deeper layers have filters that have been stretched out, leading to a larger receptive field, while making sure the output at time t is only reliant on time series data up to time t. When generating an encoding for a window W_t, the encoder is also fed the missing data mask to incorporate that information into its learning. The discriminator is a simple single layer MLP, that takes in a pair of encodings and estimated the probability of those encodings belonging to the same temporal neighbourhood.

Given that selecting good quality positive and negative samples is key in contrastive learning, we explain how we robustly define a temporal neighborhood for positive samples and the non-neighboring region for negative samples below. We also describe our novel approach for determining the optimal representation size L for every time series dataset as part of the learning process.

Defining more robust Temporal Neighborhoods: The temporal neighborhood determines the distribution of the positive and negative samples for the contrastive objective, therefore it is an integral part of both the TNC and our framework. If a neighbourhood is too narrow, many positive samples will have overlaps in time, and therefore the encoder would only learn trivial similarities, encoding information from the overlap and failing to generalize. On the other hand, if the neighborhood is too big, it would span over multiple underlying states, and the encoder would learn to assume all neighborhoods to be similar, failing to distinguish between distinct patient states. The ADF test used by TNC fails in timeseries with missing measurements or irregular sample intervals, therefore the test will very often return the smallest possible neighbourhood range in these scenarios.

To alleviate this issue, we propose a test based on the absolute value of the autocorrelation score

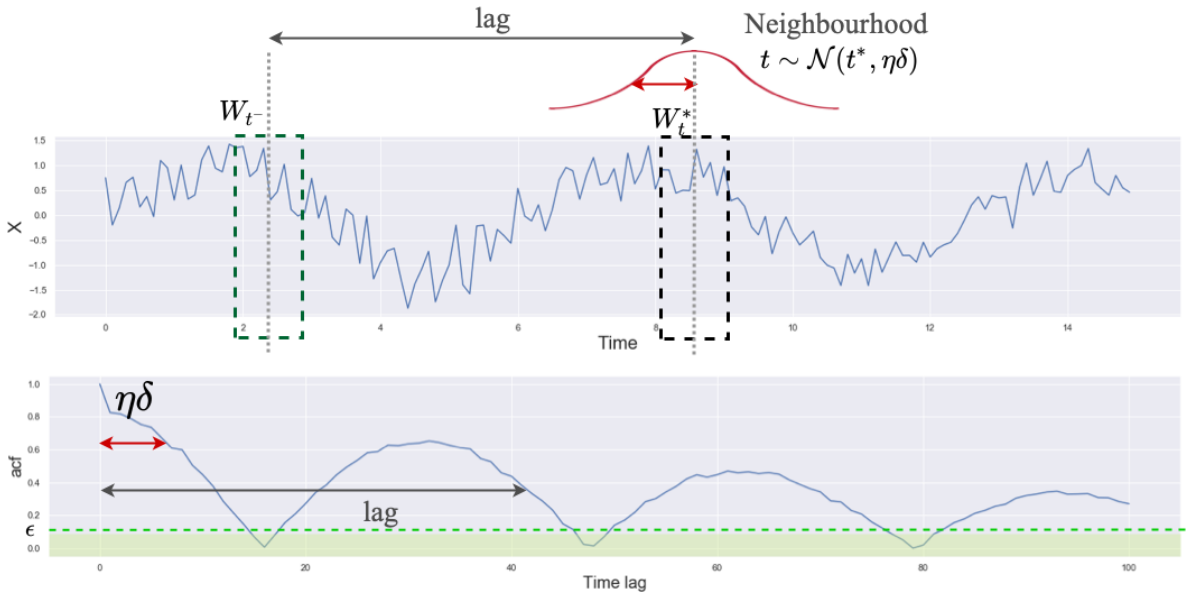

Figure 2: Autocorrelation test for identifying the neighbouring and non-neighbouring regions. The neighbourhood range is determined by the lag after which the acf score drops below the accepted threshold (0.6). The non-neighbouring samples are drawn from times that are further than 2 standard deviation apart from the center of the reference window, and also have a correlation smaller than ϵ.

$acf(\Delta t)$ for determining the neighborhood regions. The score is computed using the Pearson correlation between observations at different times, as a function of the time lag Δt. Since the neighborhood represents the stationary region of a time series, we assume that the measurements should be correlated for time lags smaller than the neighbourhood range η. When selecting the neighborhood around a window W_t, we compute the autocorrelation values for the sample that contains W_t. We find the smallest lag for which autocorrelation is smaller than the threshold (we assume it to be 0.6 throughout this paper), and then expand the neighborhood until it encompasses time series up to that lag to determine the neighbourhood range η. Correlated regions are assumed to contain samples with similar underlying states. In the evaluation section we show that the autocorrelation test can result in a better performance for the encoder than the ADF test. Another major benefit of this test over ADF is that it is far more computationally efficient. Calculating the ADF score can be slow and it becomes a bottleneck for training as it needs to be estimated for every single window. However, with the autocorrelation test, we only need to compute the $acf(\cdot)$ once for each sample in the training. This significantly reduces the number of computations required during training [2].

Overcoming sampling bias: Sampling bias is a common issue with contrastive learning methods. It is introduced when negative samples that are drawn randomly from the dataset are similar to the reference sample, and it will substantially impact the learning framework's performance (Chuang et al., 2020; Saunshi et al., 2019). For time series for instance, in the presence of seasonality, signals can exhibit similar behaviours at distant times (outside of the neighborhood region-$\bar{N}$) and therefore should not be considered as negative samples. TNC adjusts for this bias by introducing a weighting parameter (w in Eq. 1) that represents the probability of having samples similar to W_t in $\bar{N}$ and can be approximated using prior knowledge of the underlying state distribution or tuned as a hyperparameter. By using the autocorrelation, our method controls for the bias without the need for introducing an extra hyper-parameter or prior knowledge of the data distribution. We impose an additional constraint over the non-neighboring re-

2. The original TNC framework takes $\sim 100\times$ longer to train using the stattools implementation of the ADF test, as compared to TRACE. Training done on a T4 Nvidia GPU.

gion $\bar{N}_t$ to ensure the negative samples drawn from this region are not generated from the same underlying state as the reference window.

By default the negative samples are sampled from the set of all windows with distance larger than Δ from the reference, where Δ is determined as the second standard deviation of the neighborhood distribution. To remove negative samples that are similar to the reference, we check the correlation for those samples and drop the ones that have high correlation when estimating the second expectation in our objective function (Eq. 2). Since acf scores are measured for each sample once, this check can be done easily by looking at the correlation score function, as explained in Figure 2 for a simple time series example. Let's assume the reference window is centered at time t^*, and we randomly sample a negative window W_{t^-} centered at t^-. We need to check the ACF scores for the lag regions of $|t^- - t^*|$ and if the correlation is higher than our accepted threshold ϵ, W_{t^-} will be ignored. To simplify this idea in the objective, we define the non-neighbouring region for W_{t^*} ($\bar{N}_{W_{t^*}}$) as the set of t^- where $|t^- - t^*| > 2 * \eta * \delta$ and $acf(|t^- - t^*|) < \epsilon$.

Learning the optimal dimension of the encoding: A challenge and an open question in representation learning is determining the appropriate dimension size for representations. A larger encoding size will give the representations more power to encode information, however, it will also become prone to learning spurious details and as a result miss out on interpretability. On the other, if the selected size is too small, the encoder will not be able to encode enough information about the time series and will therefore lose on generalizability. One of the novelties of TRACE is that it automatically finds this balance and learn the optimal size for the encoding. We achieve this by pruning highly correlated encoding dimensions during the training process. Our learning algorithm is an iterative process of training the encoder and pruning correlated dimensions and we repeat this process until convergence is reached. In each iteration, to ensure that the encoder can recover after reducing the encoding size, we limit the number of dimensions that are removed to a single dimension, and continue training until we have reached convergence.

4. Experiments

We have evaluated the performance of TRACE on two different ICU datasets, with different patient populations and characteristics, as well as sample sizes and signal resolutions. In general, evaluating unsupervised methods is challenging as we often don't have access to well-defined labels for the latent states. We measure the usability and generalizability of the representations for different clinical tasks. We demonstrate that the representations summarize the informative parts of the signals and can be used for a number of downstream tasks, from prediction of a cardiac arrest to classifying the diagnostic groups.

4.1. Baselines

In our experiments, we compare the performance of our approach with various baselines. A number of them are unsupervised representation learning methods specifically designed for time series that have shown great utility across different datasets, and the others are models trained in a supervised fashion, using labels for the downstream task. Below are more details on each baseline:

1. TNC: The Temporal Neighbourhood Coding framework introduced in (Tonekaboni et al., 2020) and described in detail above.

2. CPC: Contrastive Predictive Coding (CPC) is an unsupervised representation learning framework introduced by van den Oord et al. (2019). This method uses predictive coding principles to train an encoder using a predictive contrastive loss.

3. Triplet-Loss [T-Loss]: This framework introduced by Franceschi et al. (2020) uses a time-based triplet loss objective for training encoders for time series samples. The triplet loss objective ensures similar time series have similar representations by minimizing the pairwise distance between positive and negative samples.

4. Supervised (End-to-end) [E2E]: This baseline uses the same encoder architecture as the other baselines, but trains its parameter end-to-end for the downstream task. This means, the encodings extracted here are specific to the classification labels and not necessarily a general encoding of the patient health state.

5. Supervised (Raw data) [Raw]: This baseline demonstrates the performance of the same downstream model as other baselines, applied on the raw data, as opposed to the learned representations. It can be used as a reference to estimate how much of the performance comes from the downstream model, and how much is the result of high-quality encodings.

4.2. HiRID Dataset

Data description and processing We paraphrase the description of the dataset from Hyland et al. (2020): High time Resolution ICU Data set (HiRID), is a freely accessible critical care dataset containing data relating to more than 33 thousand adult patient admissions to the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland. This dataset has a uniquely high time resolution of each entry every two minutes and includes information on a total of 712 routinely collected physiological variables, diagnostic test results and treatment parameters.

To show the usability of the learned representations using our approach, we evaluate how generalizable these representations are for identifying a variety of patient complications. We demonstrate that the representations can be used to train simple models that predict the diagnostic group code upon admission to the ICU, the risk of in-hospital mortality, and the risk of circulatory failure in ICU patients. Furthermore, the representations show the state progression as patients condition evolves; for instance when approaching circulatory failure. In all our evaluations TRACE is constantly amongst the top performing methods, regardless of the downstream task, further proving its generalizability. In the Appendix A.1 we explain in more details the inclusion criterion, the processing steps and the clinical variables used for each of the different experiments.

12 hour in-hospital mortality The learnt representations over time encode patients underlying health state into a lower dimensional space. Using such representations, we can train a simple model to predict the probability of in-hospital mortality. We train an single layer LSTM model on 4 days of patient representations, equivalent to 40 encodings, to predict the probability of mortality in the future 12 hours from the end of that time window. Table 1 summarizes the performance of our model TRACE and the different baselines on this task. Our method outperforms the rest of the unsupervised baselines, and is the closest in performance to a supervised model that is trained end-to-end for learning to predict mortality from the data. Increasing the complexity of the classifier model that predicts mortality from the representation brings the downstream performance closer to the supervised baselines. However, in this evaluation, our objective is the asses the quality of the representations, so we have chosen a simple model in order to minimize the impact of the classifier's learning capacity in our evaluation. Given that in-hospital mortality happens in less than 10% of our dataset, we have chosen to report the results using the AUROC, AUPRC, and the F1 score to be more reflective of the true performance of the model. In the Appendix A.4 we have included the Receiver-Operator curve and the Precision-Recall curves for further reference.

	HiRID Mortality		
Model	AUROC	AUPRC	F1
TRACE	0.80 ±0.03	0.48 ±0.03	0.59 ±0.08
TNC	0.77 ±0.04	0.32 ±0.02	0.41 ±0.20
CPC	0.62 ±0.00	0.15 ±0.00	0.46 ±0.06
T-Loss	0.64 ±0.00	0.19 ±0.01	0.44 ±0.19
Raw data	0.60 ±0.02	0.16 ±0.00	0.54 ±0.02
E2E	0.97 ±0.01	0.66 ±0.07	0.71 ±0.03

Table 1: Performance of all baselines in predicting in-hospital mortality within 12 hours

Clinical diagnostic groups To show that the representations are general enough to be used for different tasks, we further evaluate our baselines on a clinical diagnosis classification task. For this task we use the APACHE diagnostic groups (Acute Physiology and Chronic Health Evaluation) that is one of several ICU scoring systems for assessing severity-of-disease. It uses several key laboratory and physiological measurements to estimate patient prognosis and when employed additional information regarding the primary diagnostic group (upon admission) is collected by the clinicians. Examples of common diagnostic groups, as coded by the treating physicians, are sepsis, cardiovascular, metabolic/endocrinology, trauma, neurological and others. We show that our representations, derived using only physiological measurements and several laboratory values, can be used to classify patients into their admission diagnostic group. Thus, using representations derived from signals from the first 24 hours after admission, we train

a recurrent model to classify the patients' diagnostic groups. As shown by the results in Table 2, TRACE achieves a classification performance close to the supervised models, improving over most unsupervised baselines. This shows that the representations summarize parts of the signal that are informative and predictive of a variety of events.

	Diagnostic Groups	
Model	Accuracy (%)	AUROC
TRACE	21.3 ± 0.9	0.61 ± 0.01
TNC	19.0 ± 0.8	0.57 ± 0.01
CPC	13.3 ± 3.9	0.57 ± 0.01
T-Loss	19.8 ± 0.5	0.62 ± 0.00
Raw data	25.3 ± 0.8	0.72 ± 0.01
E2E	25.0 ± 0.5	0.72 ± 0.01

Table 2: The performance of all baselines on classifying the clinical diagnostic groups. Since this is multi-class classification, AUROC is computed using the 'One vs Rest' strategy.

Circulatory failure As another test, we evaluated our representations to see if they can be used to predict circulatory failure in patients. Circulatory failure is a common condition among critically ill patients, and while dangerous, its effects can be reversed if caught early enough. We use our representations to estimate the risk of a circulatory failure (as defined by Hyland et al. (2020)) in patients over time. Similar to the mortality prediction task, we take a sequence of representation and train an RNN to predict the risk of circulatory failure. In this case, each positive sample is a sequence of encodings derived from 2 days prior up to 3 hours prior to failure. As shown by the results in Table 3, TRACE performs quite well compared to the other unsupervised methods, however it does fall behind the supervised methods.

The learned representations are not only informative for predicting circulatory failure, but can also be used to estimate the risk over time. Figure 3 shows how the pattern of encodings (second row) are used by the RNN to estimate the risk of circulatory failure over time (third row). We can see a sudden change in representation roughly 10 hours prior to failure, which is accompanied by volatility and eventual increase in predicted risk. Presenting a low dimensional continuous representation that captures the health state of the patient is of great clinical utility as it allows the clinicians to not only identify patients that are deteriorating early, but also to intervene proactively and assess the adequacy of the interventions. The 2D projection of the encodings are also shown on Figure 4 for a different patient experiencing circulatory failure. Each data point represents the encoding of a window, and the color indicates how many hours before the failure the window was. We clearly see that as the patient approached circulatory failure, the underlying state changes, and states that represent high-risk signals that are closer to the failure, cluster separately from the representations of states farther from the failure.

	HiRID Circulatory Failure		
Model	AUROC	AUPRC	F1
TRACE	0.73 ±0.04	0.29 ±0.05	0.56 ±0.04
TNC	0.74 ±0.02	0.24 ±0.01	0.53 ±0.01
CPC	0.66 ±0.01	0.17 ±0.00	0.52 ±0.00
T-Loss	0.69 ±0.02	0.22 ±0.01	0.55 ±0.02
Raw data	0.87 ±0.02	0.54 ±0.03	0.69 ±0.00
E2E	0.86 ±0.02	0.51 ±0.04	0.66 ±0.02

Table 3: Performance of all baselines in predicting circulatory failure

4.3. High-frequency physiological signal from a pediatric hospital ICU

Data description and processing We have evaluated our approach on a dataset from the pediatric ICU of the Hospital for Sick Children in Toronto, Canada. This dataset consists of high-resolution physiological signals collected from bedside monitors over the course of 5 years. The measurements include vitals such as heart rate, respiratory rate and ventilator measures such as the end-tidal CO2, and are measured up to 12 samples per minute pervasively for all patients throughout their stay. More information about this dataset is provided in Appendix A.2. A subset of the patients in this cohort experienced in unit cardio pulmonary arrest (CPA), for which we have the labels. We used these labels to evaluate the quality of our learned representation in multiple ways.

Cardio Pulmonary Arrest (CPA) prediction For this experiment, we measure how well representations identify patients who experience a cardiopulmonary arrest. To do so we train an RNN to estimate the risk of CPA, using the representations of

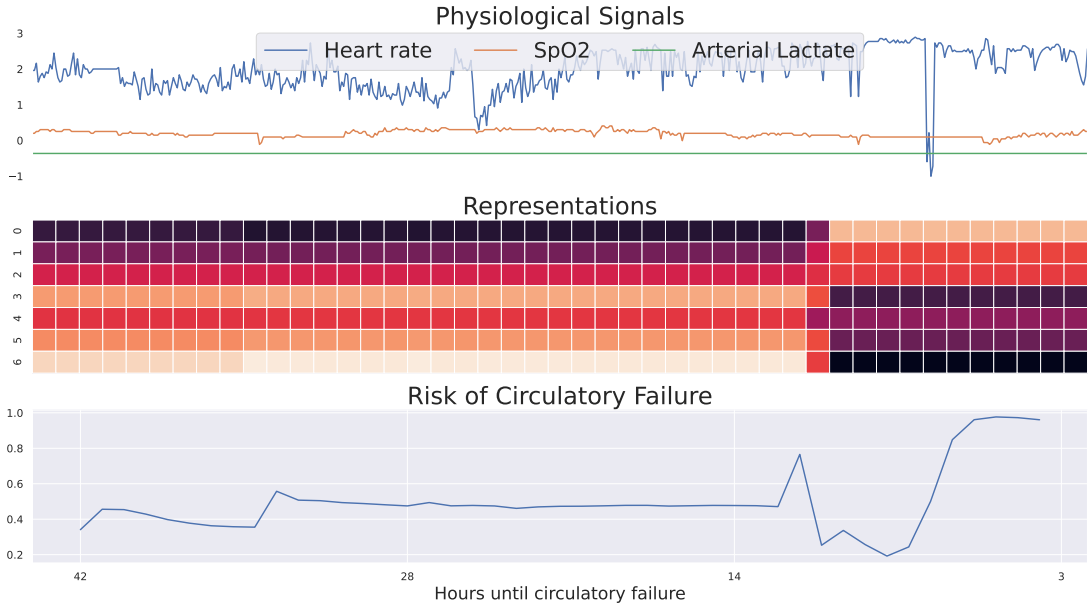

Figure 3: Representations of a patient who experiences circulatory failure during their stay. As the time of failure approaches, the patient transitions to a state associated with high risk.

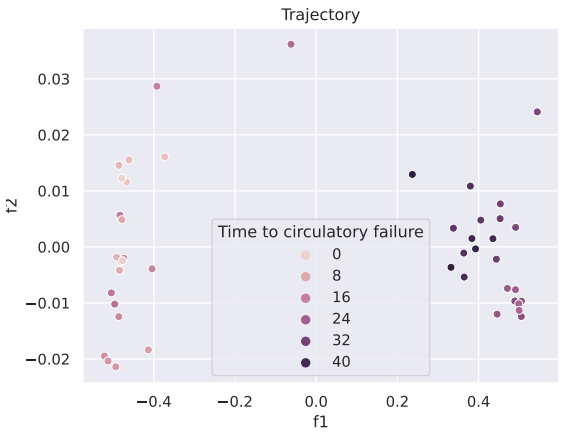

Figure 4: Trajectory of patient embedding approaching circulatory failure. Each data point in the plot represents the representation of a window of time series, and the color indicates how long before the circulatory failure the window was. Notice that as the states approach the circulatory failure, they shift from one cluster to another.

the 7 hours prior. CPA is a rare event that happens in the ICU, and is only present in about 2% of of the data samples. Table 4 demonstrates the performance results for our model in comparison to our supervised and unsupervised baselines. TRACE outperforms all unsupervised baselines for this task, and also performs better than a supervised model trained on the raw data. This shows that the significance of representation learning becomes more significant with increased frequency and complexity of the time series data.

	Cardiopulmonary Arrest		
Model	AUROC	AUPRC	F1
TRACE	0.98 ±0.00	0.60 ±0.05	0.77 ±0.01
TNC	0.96 ±0.01	0.48 ± 0.03	0.72 ±0.02
CPC	0.75 ±0.01	0.09 ±0.03	0.54 ±0.00
T-Loss	0.75 ±0.01	0.11 ±0.02	0.54 ±0.00
Raw data	0.67 ±0.06	0.08 ±0.04	0.51 ±0.24
E2E	0.97 ±0.01	0.66 ±0.07	0.71 ±0.03

Table 4: Performance of all baselines in predicting cardiopulmonary arrest

Another benefit of learning the underlying representation of signals over time is that it allows us

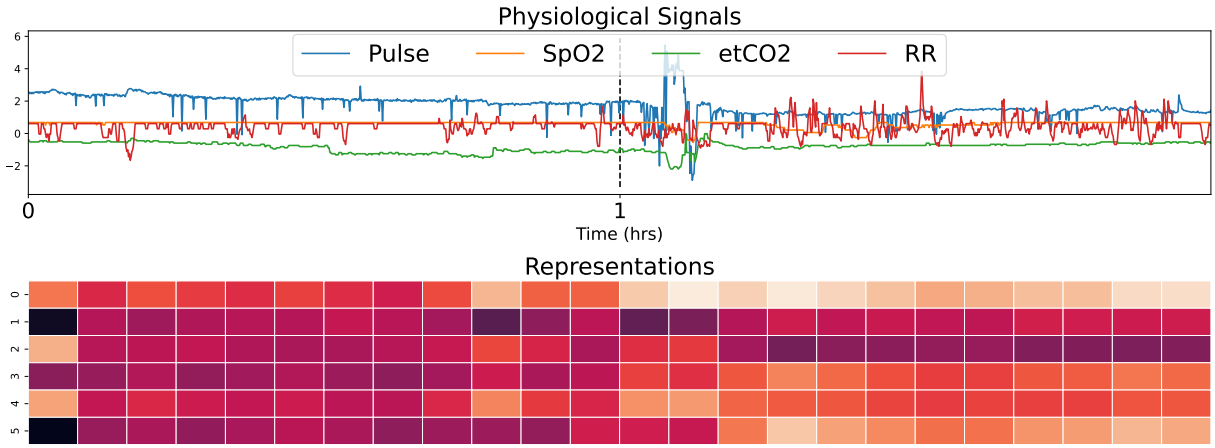

Figure 5: Physiological signals and the learned representations of states over time for a patient experiencing cardiopulmonary arrest. The top panel shows a subset of signals (normalized) over 2 hours. The bottom panel shows the 6 dimensional representations generated for each window of time by our encoder. This individual experiences cardiac arrest at the 1 hour mark. The encodings have a distinct state prior to arrest, which changes as the arrest approaches, and then settles in a different state after the arrest.

to better understand the health state trajectory and to identify abrupt changes in patient condition. We investigated this by looking at the patterns of signal representations as they approach the CPA. Figure 5 shows the normalized physiological signals (top panel) and the representations (bottom panel) for an individual experiencing arrest. The patient experiences arrest at hour 1, indicated by the vertical dashed line in the top graph. The heatmap presents the 6 dimensional encodings for consecutive windows of signal over time. As the patient approaches the cardiac arrest, the pattern of the representation starts changing as early as 30 minutes before the event. This shows one of the major benefits of representations learning for clinical time series that by tracking the representations over time we can see the change in the underlying state of patients and these kinds of insights can lead to early detection and intervention of severe cases. In Figure 5 we can also see that after resuscitation, the patient settles into a different state that is different from the state prior to the arrest.

To further show how patient states evolve over time, in Figure 6 we also demonstrate the 2-dimensional projection of encodings (over time) for an individual that experiences CPA. Each data point in the scatter plot is the representation of a window, and the color indicates the time to the arrest. We can clearly see the evolution of states as the windows approach the arrest, indicated by the lighter points.

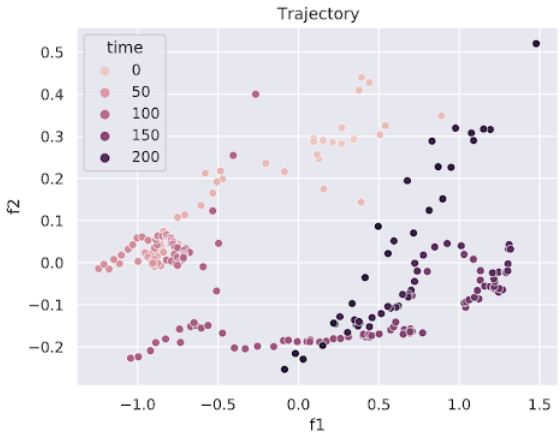

Figure 6: 2-dimensional projection of the representations of a patient over time, who experiences a cardiac arrest. Each datapoint in the scatter plot is the representation of a window of the physiological signals, and the color indicate how long before the arrest the window is. We can see the evolution of state as it approaches a critical event like an arrest.

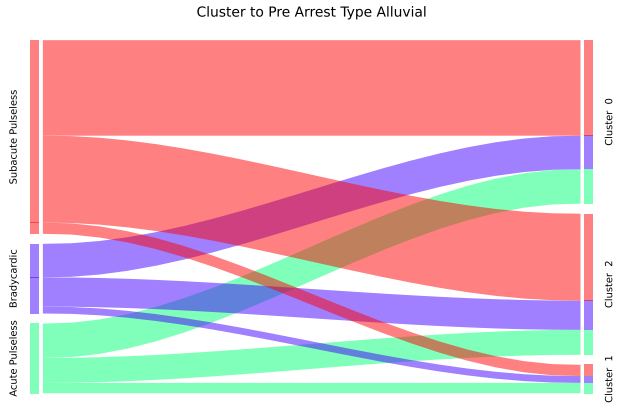

Figure 7: Alluvial Visualization to show how the discovered underlying states correspond to different subcategories of CPA.

Identifying sub-categories of CPA from the representation patterns In the previous evaluation, we show that signs of a cardiac arrest can be observed in the representations over time. But further investigation into the representations shows that these patterns can be different across individuals. Erez et al. (2021) has shown that individuals exhibit distinct physiologic patterns prior to in-hospital cardiac arrest. Based on these patterns, they have defined 3 categories of cardiac arrest, namely Bradycardic, Subacute Pulseless, and Acute Pulseless CPA. Using their definition, we evaluate whether such categories can be identified using our representations.

Using hierarchical clustering, we cluster the representations of pre-arrest signals into 3 states. Then we assess how correlated our states are with the predefined sub-categories as shown in Figure 7. Clusters 0 and 2 appear to have high correlation with the Subacute Pulseless type of arrest. Bradycardic and Acute Pulseless however appear to be more uniformly distributed between the 3 clusters.

5. Conclusion

Medical time series provide a complex but rich source of information for patients. We have shown in this paper that using the right unsupervised representation technique we can learn the underlying state of patients using their physiological signals in a representation vector. The representations summarize the informative parts of the signal into a lower dimensional space that can be used to train models for a number of downstream tasks and tracking them over time will help us monitor disease progressions in patients. Representation learning approaches can also have a potential in knowledge discovery to help uncover the underlying states in patients. In this work, we have made an attempt in overcoming a number of shortcomings in the existing methods to introduce an unsupervised approach that is suitable for medical time series with all its complexities like non-stationarity, missing observations, and irregular sampling rate.

Institutional Review Board (IRB)

This work has IRB approval for collection of data at the pediatric ICU of the Hospital for Sick Children (reference number: 100068499).

Acknowledgments

This work is generously supported by the Canadian Institute of Health Research (CIHR) and the Natural Science and Engineering Research Council (NSERC). Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and sponsors of the Vector Institute [3].

References

Ahmed Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. 2019.

Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. Sequence to sequence autoencoders for unsupervised representation learning from audio. In *Proc. of the DCASE 2017 Workshop*, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016a.

3. www.vectorinstitute.ai/partners

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016b.

Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020.

Ely Erez, Mjaye L Mazwi, Alexandra M Marquez, Michael-Alice Moga, and Danny Eytan. Hemodynamic patterns before inhospital cardiac arrest in critically ill children: An exploratory study. *Critical care explorations*, 3(6), 2021.

Madalina Fiterau, Suvrat Bhooshan, Jason Fries, Charles Bournhonesque, Jennifer Hicks, Eni Halilaj, Christopher Ré, and Scott Delp. Shortfuse: Biomedical time series representations in the presence of structured information. In *Machine Learning for Healthcare Conference*, pages 59–74. PMLR, 2017.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems*, pages 4652–4663, 2019.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series, 2020.

Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

Guillermo Gutierrez. Artificial intelligence in the intensive care unit. *Annual Update in Intensive Care and Emergency Medicine 2020*, pages 667–681, 2020.

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.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

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.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6), 2013.

Qi Lei, Jinfeng Yi, Roman Vaculin, Lingfei Wu, and Inderjit S Dhillon. Similarity preserving representation learning for time series clustering. *arXiv preprint arXiv:1702.03584*, 2017.

Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance, 2021.

Qianli Ma, Jiawei Zheng, Sen Li, and Gary W Cottrell. Learning representations for time series clustering. In *Advances in Neural Information Processing Systems*, pages 3776–3786, 2019.

Naveen Sai Madiraju, Seid M Sadat, Dimitry Fisher, and Homa Karimabadi. Deep temporal clustering: Fully unsupervised learning of time-domain features. *arXiv preprint arXiv:1802.01059*, 2018.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46(4):547, 2018.

Marcus Eng Hock Ong, Christina Hui Lee Ng, Ken Goh, Nan Liu, Zhi Xiong Koh, Nur Shahidah, Tong Tong Zhang, Stephanie Fook-Chong, and Zhiping Lin. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the

modified early warning score. *Critical Care*, 16(3): R108, 2012.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine Learning for Healthcare Conference*, pages 73–100, 2016.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337, 2017.

Sana Tonekaboni, Mjaye Mazwi, Peter Laussen, Danny Eytan, Robert Greer, Sebastian D Goodfellow, Andrew Goodwin, Michael Brudno, and Anna Goldenberg. Prediction of cardiac arrest from physiological signals in the pediatric icu. In *Machine Learning for Healthcare Conference*, pages 534–550, 2018.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 85–94, 2014.

Ben Wellner, Joan Grand, Elizabeth Canzone, Matt Coarr, Patrick W Brady, Jeffrey Simmons, Eric Kirkendall, Nathan Dean, Monica Kleinman, and Peter Sylvester. Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. *JMIR medical informatics*, 5(4):e8680, 2017.

Appendix A. Appendix

A.1. HiRID dataset

Note: This dataset was acquired through Physionet credentialing. Only authors on this paper who were given credentialing to use this dataset, had read permissions on the saved data. If you wish to use this dataset, you must request for access via physionet.

HiRID is a public dataset containing data relating to almost 34 thousand patient admissions to the Department of Intensive Care Medicine of the Bern University Hospital, Switzerland (ICU). The dataset contains de-identified demographic information and a total of 681 routinely collected physiological variables, diagnostic test results and treatment parameters for patients admitted during the period from January 2008 to June 2016. Data is stored with a uniquely high time resolution of one entry every two minutes. To reduce missingness and have a consistent frequency in measurements of physiological signals, we use the processing of data (imputed stage) as described in the original publication. This data has been down sampled to 5 minutes, and imputed. They also merge many variables into *meta variables*, description of which can be found on their website [4].

For training all baselines, we took the imputed stage for the patients who have at least 2 days of data recorded. Overall, samples with more than 40% missing measurements were excluded from the datasets. All other samples are left imputed so the length of the signals are a multiple of 4 days. The mask channel is used to indicate these imputations.

For the **Mortality Prediction** experiment, we again pick patients with at least 2 days of data, and a discharge status not recorded as null. We remove the 12 hours prior to discharge, and make our predictions based on the information available up to that point in time. The total number of samples are 6700 for training, 1300 for validation and 1600 for test, with a positive to negative ratio of ~ 0.10.

For the **Diagnostic group** (Apache Prediction Data), we take the first 24 hrs of data for the exact same patients, and store their apache groups assigned to them. We take the apache IV code if available, otherwise we take the apache II code. The total number of samples are 6100 for training and 1500 for test. We categorize patients with no apache code into the 'other' category. We then drop categories with less than 200 patients. The apache groups included are as follows: Cardiovascular, Pulmonary, Gastrointestinal, Neurological, Sepsis, Trauma, Metabolic/Endocrinology, Hematology, Other, Surgical Cardiovascular, Surgical Respiratory, Surgical, Gastrointestinal, Surgical Neurological, Surgical Trauma, Surgical Urogenital, Surgical Gynecology, Surgical Orthopedics, Urogenital, Surgical others, and Intoxication.

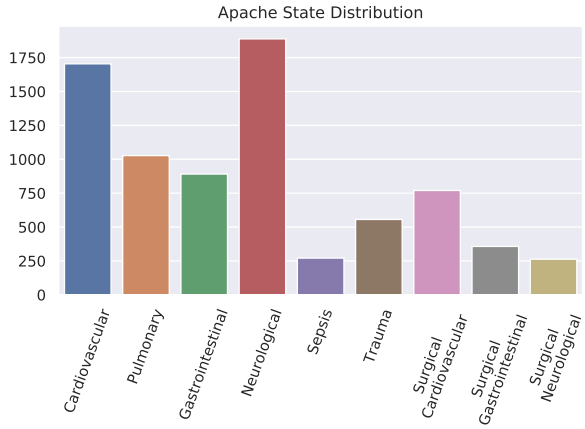

Figure A.8: Distribution of Apache States after removing low count categories

For the **Circulatory Failure** Prediction Data, we use the definition of circulatory failure in Hyland et al. (2020): "A patient is defined as being in circulatory failure if (1) arterial lactate is elevated ($\geq$ 2 mmol l–1), and (2) either mean arterial pressure (MAP) $\leq$ 65 mmHg, or the patient is receiving vasopressors or inotropes". We identified the vasoporessors in the dataset as being doses of Noradrenalin or Vasopressin and inotropes as being doses of Adrenalin, Dobutrex, Corotrop, or Simdax. We note all places the patients were administered an inotrope or vassopressor, and consider the next hour as them being on an inotrope or vassopressor. These pharmaceuticals informations are available as part of the dataset. [5]

For patients that do experience circulatory failure, we take the last 2 days of data before they first experience it. For patients that do not experience circulatory failure, we take a random 2 day period during their stay and store this.

4. https://hirid.intensivecare.ai/data-details

5. https://docs.google.com/spreadsheets/d/1MjihfhyXX4dwni8Fxy3Ji5RCvSvnhipDCyjYo_6rixY/edit?usp=sharing

Signal	Description	Sample Rate
HR	Heart rate derived from ECG	5 seconds
RR	Resp Respiratory rate	5 seconds
Pulse	Pulse rate	5 seconds
SPO2	Oxygen saturation level	5 seconds
etCO2	End-tidal CO2	5 seconds
NBPm	Non-invasive blood pressure	1-60 minutes
NBPd	Non-invasive blood pressure	1-60 minutes
NBPs	Non-invasive blood pressure	1-60 minutes

Table A.5: Description of high-frequency ICU signals

A.2. High-frequency data from pediatric ICU dataset

This dataset contains thousands of hours of physiological signals. For each patient, we first remove the last 5 minutes prior to arrest (if the patient does experience an arrest), and then left pad all signals so that the total length is an integer multiple of 7 hours. Then it is broken into 7 hour samples. Samples with missingness greater than 60% (i.e. samples where >= 60% of time steps were either padded data or had no observed value for any signal at that time step) were removed. Note it is possible for a single patient's data to be broken into multiple samples. Then, the data was forward imputed (meaning the last observed value for a given feature is put in place of unobserved values for that feature).

Each feature in the test set were normalized by subtracting the mean of that signal across all samples, and dividing by the standard deviation of that signal across all samples. Note that these statistics were only computed for observed values, not imputed values. The same process was done for the train/validation set jointly (that is the feature mean and standard deviation are computed across train and validation samples).

There were 5428 samples for training, and 143 of those were positive samples (samples for which an arrest happens at the end).

A.3. Experiment details

Refer to the CausalCNNEncoder class in tnc/models.py for our encoder architecture. The hyperparameters chosen for the HiRID model was in_channels=36, channels=4, depth=1, reduced_size=2, encoding_size=10, kernel_size=2, window_size=12. During pruning, the initial encoding size of 10, was reduced down to 6. The linear classifier trained on top of the encoder for the downstream tasks was an LSTM with hidden state of size 8. Please refer to the circulatory_failure_prediction.py for more details, as well as hyperparameters for the classifiers like learning rate, epochs, etc.

	Pediatric ICU	HiRID
ϵ	0.1	0.1
D (Number of features)	10	18
Representation size	8	10
Pruned Representation size	6	6

Table A.6: Caption

The hyperparameters chosen for the pediatric ICU model was in_channels=20, channels=8, depth=2, reduced_size=30, encoding_size=8, kernel_size=3, window_size=60. During pruning, the initial encoding size of 8, was reduced down to 6. The linear classifier trained on top of the encoder for the downstream tasks was an LSTM with hidden state of size 8. Please refer to the train_linear_classifier function in tnc/tnc.py for more details, as well as hyperparameters for the classifier like learning rate, epochs, etc.

The HiRID encoder was trained with the Adam optimizer with learning rate 0.00005, weight decay of 0.0005, and for 150 epochs.

The pediatric ICU encoder was trained with the Adam optimizer with learning rate 0.0007, weight decay of 0.0001, and for 150 epochs.

A.4. Supplementary plots

A.4.1. Performance plots:

For models predicting rare events such as mortality and cardio pulmonary arrest, performance evaluation can be challenging. Metrics like AUROC and AUPRC provide some insight on the performance of these models but to get a better understanding, we have included plots of the PRC and ROC curves for Mortality prediction on the HiRID dataset. The model numbers in the captions refer to the classifiers we trained on top of the single encoder (we had cross validation of 3 for the classifiers). We also include the loss curve of the encoder trained for HiRID.

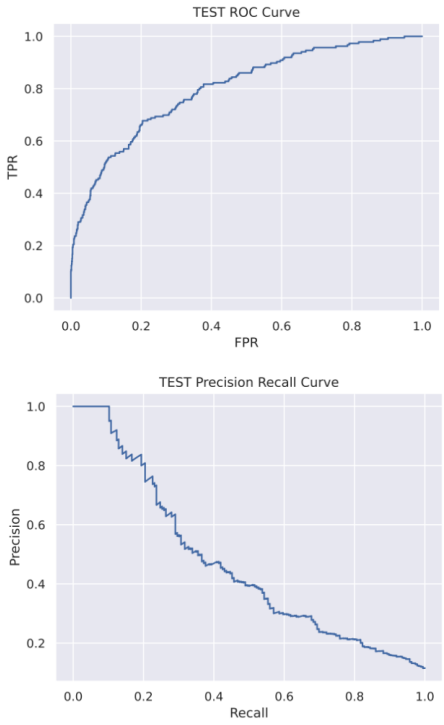

Figure A.9: Mortality Precision Recall Curve and Receiver Operator Characteristic Curve for Model 1

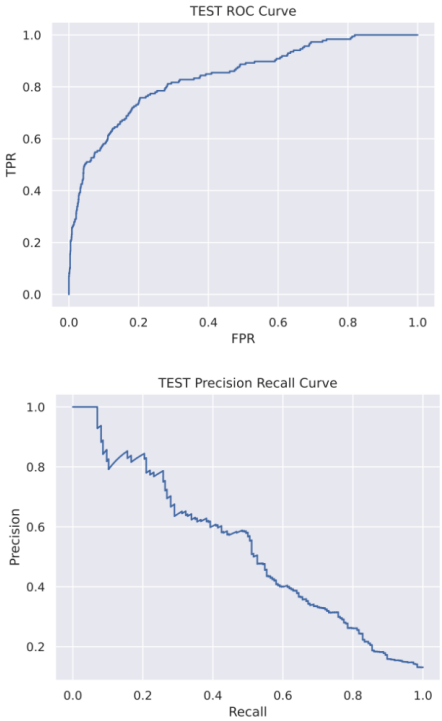

Figure A.10: Mortality Precision Recall Curve and Receiver Operator Characteristic Curve for Model 2

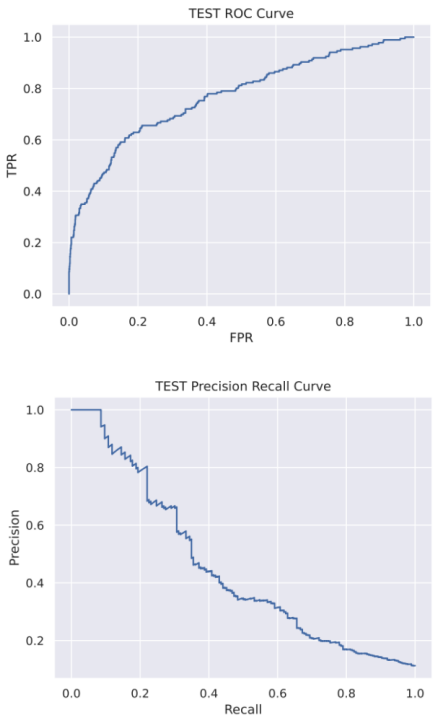

Figure A.11: Mortality Precision Recall Curve and Receiver Operator Characteristic Curve for Model 3

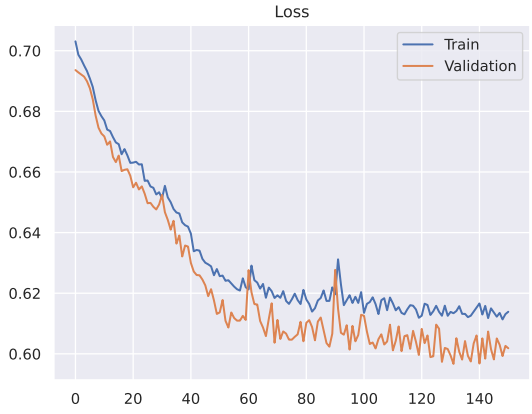

Figure A.12: Loss for our encoder trained on the HiRID data. A couple of the loss spikes caused by dimension pruning can be observed (at epochs 60 and 90)