PhysioJEPA: Joint Embedding Representations of Physiological Signals for Real Time Risk Estimation in the Intensive Care Unit

Benjamin Fox BEN.FOX@ICAHN.MSSM.EDU
Dung Hoang JOLIE.HOANG@ICAHN.MSSM.EDU
Joy Jiang JOY.JIANG@ICAHN.MSSM.EDU
Pushkala Jayaraman PUSHKALA.JAYARAMAN@MSSM.EDU
Ankit Parekh ANKIT.PAREKH@MSSM.EDU
Girish N. Nadkarni GIRISH.NADKARNI@MOUNTSINAI.ORG
Ankit Sakhuja ANKIT.SAKHUJA@MSSM.EDU
The Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Self-supervised learning of multi-modal, high-frequency physiological signals is largely unexplored, despite its potential for critical care applications. We present PhysioJEPA, a Joint Embedding Predictive Architecture (JEPA) designed for multi-modal physiological signals from critical care bedside monitoring devices. PhysioJEPA learns representations from 30-minute segments of physiological signals from three channels: arterial blood pressure, electrocardiography lead II, and photoplethysmography. Trained on over 10.7 million minutes of data from 4,282 intensive care unit stays (N=2,631 patients) in the Medical Information Mart for Intensive Care-III (MIMIC-III) Waveform Database, the learned, frozen representations of PhysioJEPA can be used to estimate 5-minute risk of hypotension (AUROC = 0.83 [Confidence Interval or CI 0.83-0.84]) and shock index (AUROC = 0.95 [0.95-0.96]), with comparable performance to a self-supervised Patch Time Series Transformer framework (AUROC = 0.87 [0.86-0.87] and 0.96 [0.96-0.96]), better performance compared to another JEPA physiological signal model, ECG-JEPA (AUROC = 0.73 [0.72-74] and 0.92 [0.92-0.93]), and better performance compared to a supervised convolutional model (AUROC = 0.78 [0.78-0.78] and 0.95 [0.95-0.95]). Notably, it can generalize to an independent healthcare system (AUROC = 0.78 [0.78-0.78] and 0.92 [0.92-0.93]) better than all comparison models. These results suggest that self-supervised JEPA representation learning is a promising approach for multi-modal bedside monitoring signal data.

Keywords: self-supervision, transformer, JEPA, physiological signals, multi-modal, multichannel, electrocardiography, photoplethysmography, critical care, hypotension, shock

Data and Code Availability The MIMIC-III Waveform Database Matched Subset is publicly available upon request from https://physionet.org/content/mimic3wdb-matched/1.0/. Code is available at https://github.com/benmfox/PhysioJEPA. The Mount Sinai Bedmaster Dataset is not currently available.

Institutional Review Board (IRB) The MIMIC-III Waveform Database Matched Subset is publicly available. The Mount Sinai Bedmaster Dataset was approved for use under IRB by the Icahn School of Medicine at Mount Sinai (STUDY-20-00338).

1. Introduction

Physiological time series data contain information-rich health data, capturing continuous measurements of vital signs, organ function, and metabolic processes that provide real-time insight (Orphanidou, 2019; Rooney and Clermont, 2023). From electrocardiograms (ECG) and electroencephalograms (EEG) in hospital settings to continuous glucose monitoring and wearable sensor data in ambulatory care, these high-frequency signals contain complex temporal patterns that reflect underlying pathophysiological processes. The ability to effectively model and learn from such time series data is crucial for enabling early detection of adverse events and supporting clin-

ical decision-making (Orphanidou and Wong, 2017; Rooney and Clermont, 2023).

Following the advancements of the attention mechanism (Vaswani et al., 2017) and subsequent work with bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) and chat generative pretrained transformers (ChatGPT) (Brown et al., 2020), self-supervised representation learning has transformed approaches to modeling complex, high-dimensional data across domains. In healthcare specifically, foundation models trained on large-scale datasets have demonstrated remarkable capabilities in medical imaging (Zhou et al., 2023), cardiac magnetic resonance imaging videos (Kim et al., 2024), clinical notes (Wornow et al., 2023; Huang et al., 2019; Singhal et al., 2022), and multimodal medical data (Moor et al., 2023a,b). However, physiological time series present unique challenges that distinguish them from other medical data modalities: they are inherently temporal and often multivariate with complex interchannel dependencies, characterized by high sampling frequencies that generate massive datasets.

Recent developments in self-supervised learning have demonstrated that Joint Embedding Predictive Architectures (JEPA) can learn superior representations by predicting in the embedding space rather than reconstructing raw input data, particularly in imaging and video data (Assran et al., 2023; Bardes et al., 2024). Unlike traditional masked autoencoding approaches that reconstruct values or contrastive approaches that compare augmented embeddings, JEPA uses a context encoder to embed random parts of an input (i.e., patches of an image, intervals of a time series). Then, a predictor network uses these context embeddings to estimate other, non-overlapping parts of the input in the embedding space that have been encoded by a separate target encoder. This target encoder, updated via exponential weighted moving average, generates ground-truth embeddings for comparison. This embedding-space prediction framework is hypothesized to avoid noisy pixel reconstruction while learning higher-level representations, leading to more robust and generalizable features (Assran et al., 2023). For physiological time series, this approach is especially compelling as it circumvents the reconstruction of inherently noisy signal data while potentially capturing complex temporal and cross-channel dependencies that are critical for clinically relevant tasks. Thus, we hypothesize that JEPA offers better representation learning for multi-modal physiological signals by modeling predictive relationships within the embedding space and across signal channels, rather than reconstructing inputs or contrasting samples' embeddings.

In this work, we introduce PhysioJEPA, a JEPA-based model for multi-modal physiological signals from bedside monitoring devices in intensive care units (ICU). PhysioJEPA learns representations from 30-minute segments of three-channel bedside monitoring data (arterial blood pressure [ABP], ECG lead II, and photoplethysmography [PPG]) through self-supervised training. This approach enables accurate risk estimation of critical care outcomes via task-specific non-linear probing (i.e., fitting a supervised classifier on top of frozen representations).

Our work makes the following key contributions:

- We present the first application of JEPA to bedside monitoring data, extending JEPA principles to handle 30-minute segments of three-channel physiological time series (ABP, ECG, PPG) sampled at 125 Hz.
- We develop channel-specific mask token prediction strategies, employ rotary positional embeddings, and use depthwise convolutions for tokenization, enabling the model to learn distinct representations for each physiological signal type over time while capturing cross-channel dependencies.
- We show that PhysioJEPA is effective in critical care risk estimation for 5-minute hypotension risk and 5-minute shock index risk, matching or outperforming supervised and self-supervised comparison models.
- We demonstrate that PhysioJEPA can robustly generalize to an external dataset without retraining, outperforming performance of supervised and self-supervised comparison models and establishing the potential for real-world ICU deployment.

2. Related Work

2.1. Time Series Representation Learning

Self-supervised representation learning for time series has seen significant development, with frameworks ranging from contrastive learning (Yue et al., 2022) to masked autoregressive models like Patch Time Series Transformer (PatchTST) (Nie et al., 2023). In healthcare specifically, current approaches

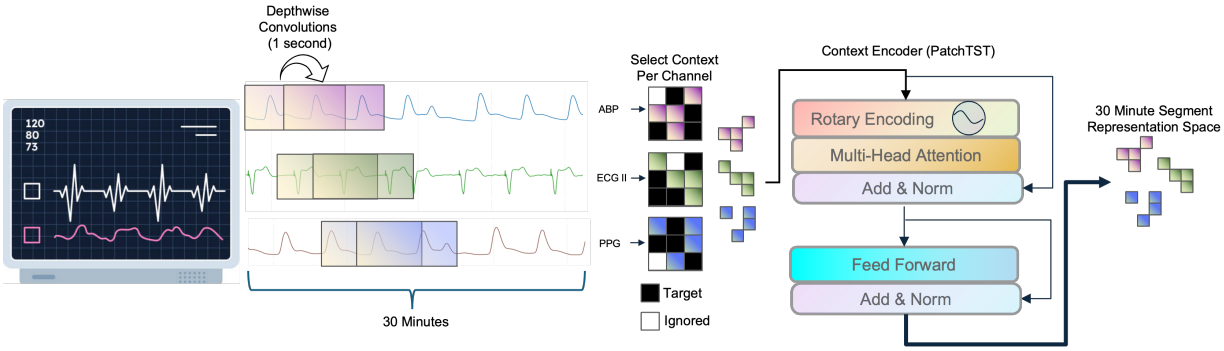

Figure 1: PhysioJEPA Context Encoder. Bedside monitoring signal channels (ABP, ECG lead II, and PPG) are extracted in 30-minute segments sampled at 125 Hz and tokenized into 1-second patches. Random patches are selected by context and target masks, which are encoded using a Patch Time Series Transformer (PatchTST) encoder (Nie et al., 2023).

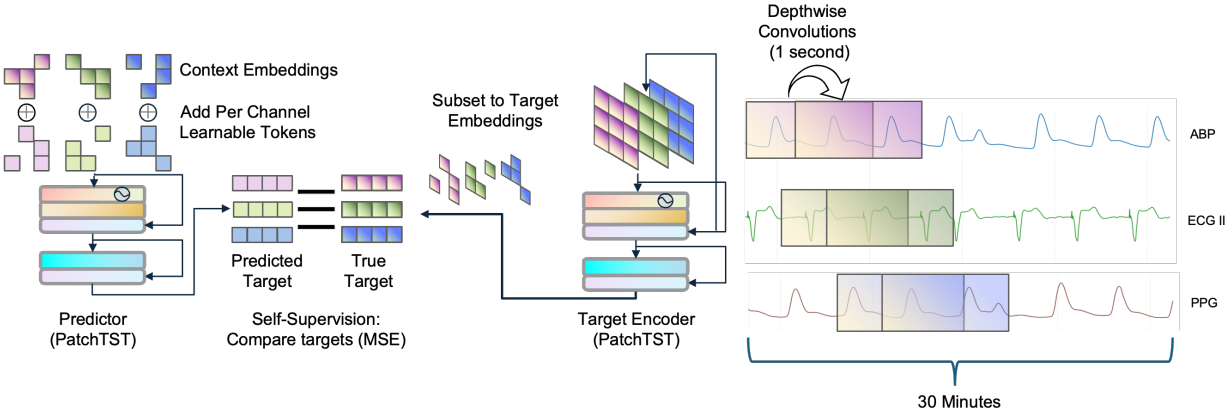

Figure 2: PhysioJEPA Predictive Architecture and Target Encoder. Channel specific mask tokens are added to the context embeddings to indicate which patches the predictor must reconstruct, while the target encoder processes the full segment and subsets embeddings at the target locations. Prediction is optimized by minimizing mean squared error between predicted and target embeddings.

for physiological signals have made important advances but still typically rely on shorter or lower frequency input sequences, single-channel or single-modality analysis, or domain-specific architectures that limit their generalizability across different physiological monitoring scenarios (Xu et al., 2024; Foumani et al., 2024; Kotoge et al., 2024; Chien et al., 2022; Ding et al., 2024; Zhang et al., 2022; Lutsker et al., 2024). This is a limitation, given that many clinically relevant patterns in physiological data manifest over extended time horizons and require integration of information across multiple signal types.

More recent works including SleepFM (Thapa et al., 2024), NormWear (Luo et al., 2024), PFTSleep (Fox et al., 2025), and wearable work by Merrill and Althoff (2022), Narayanswamy et al. (2024), and Xu et al. (2025) build representation models for higher frequency, multi-modal data streams in the sleep or wearable domains; however, none have been developed for or applied to bedside monitoring data.

2.2. Joint Embedding Predictive Architectures

Building on the foundational JEPA framework (Assran et al., 2023), recent work has explored extensions beyond the original image domain applications. Video-JEPA has demonstrated effectiveness for temporal sequence understanding by adapting the embedding-space prediction paradigm to handle temporal dynamics in video data (Bardes et al., 2024). In the physiological signal domain, studies have applied JEPA principles to ECG analysis, showing promising results on 10-second, 250-500Hz, 8 or 12-lead ECG segments for feature prediction and downstream classification tasks (Kim, 2024; Weimann and Conrad, 2025). Further, Signal-JEPA applies dynamic spatial

attention to improve generalization across datasets, showing promise for spatially distributed signals such as EEG (Guetschel et al., 2024). These applications highlight JEPA's particular advantages for noisy signal modalities; however, these applications to physiological data have been limited to short-duration, single-modality scenarios.

2.3. Physiological Signal Analysis for Critical Care

Traditional approaches to ICU risk estimation have relied heavily on aggregated features from electronic health records, signals, or vital signs (Maheshwari et al., 2021; Yoon et al., 2020; Moghadam et al., 2021; Cherifa et al., 2021; Kapral et al., 2024; Hatib et al., 2018; Jian et al., 2025). While these methods achieve reasonable performance on single tasks, they fail to leverage the rich temporal dynamics present in continuous bedside monitoring. Others have taken advantage of the continuous data for supervised classification tasks (Lee et al., 2021; Moon et al., 2024; Jeong et al., 2024; Jo et al., 2022; Sundrani et al., 2023). To our knowledge, our work is the first to apply self-supervised learning to bedside monitoring signal data for multi-outcome critical care risk estimation.

3. Methods

3.1. Data Extraction

We used the Medical Information Mart for Intensive Care-III (MIMIC-III) Waveform Database Matched Subset (Johnson et al., 2015, 2016). Waveform data with the ABP, ECG lead II, and PPG signal channels were extracted and stored in `zarr` (Miles et al., 2024) format. Also, we extracted adult ICU waveform signal data from an external institution for evaluation provided by the Mount Sinai Health System. The MIMIC-III data was split into 95% training and 5% validation for pretraining.

3.2. Data Segmentation and Normalization

3-channel signal data are segmented into 30-minute, non-overlapping windows and resampled to $f_s = 125\,\text{Hz}$. 30-minute segments are excluded from training if 20% or more of a single channel's values are constant or null values. Following, null values are linearly interpolated. Each channel $x_c(t)$ is then normalized with inter-quartile range normalization, as described by Brink-Kjaer et al. (2022).

3.3. PhysioJEPA Architecture

The framework consists of three components:

1. **Context encoder:** A PatchTST transformer (Nie et al., 2023) processes a masked subset of input patches to produce context encodings $\mathbf{h}_c$.
2. **Predictor:** A smaller PatchTST transformer predicts target encodings of masked tokens $\hat{\mathbf{h}}_t$.
3. **Target encoder:** An exponential weighted moving average copy of the context encoder produces target encodings $\mathbf{h}_t$ from the full input, the output of which is subset and compared to that of the Predictor.

Context and target encoders use 3-layer PatchTST encoders with 8 attention heads, dimension 512, and feedforward sizes of 2048. The predictor implements 2 layers with 4 heads and a dimension of 256.

3.3.1. Channel-Specific Mask Tokens

For target token prediction using context encodings, channel-specific learnable mask tokens $\mathbf{c}_i, \quad i = 1, \ldots, C$ are repeated t times (the number of targets to predict), concatenated, and *shuffled* with context tokens. The predictor then outputs $\hat{\mathbf{h}}_{t,i}^{(c)}$ at each masked position i for each channel c. This result is compared to the target encoder output and optimized with a mean squared error loss function.

3.3.2. Patch Embedding with Depthwise Convolutions

Each 30-minute segment is tokenized into 1-second patches via a depthwise convolution per channel, equivalent to PatchTST's channel independent patching and linear tokenization procedure (Nie et al., 2023):

$$h_{p,c,d} = \sum_{i=0}^{k-1} w_{c,d}[i]\, x_c[p \cdot k + i] \; + b_{c,d}, \quad d = 1, \ldots, D. \tag{1}$$

where $x_c \in \mathbb{R}^{125}$ is the channel c indexed at patch p of size k (the kernel size). Each channel c has D convolutional kernels $w_{c,d} \in \mathbb{R}^k, d = 1, \ldots, D$ with bias $b_{c,d}$. This helps learn patch-wise features into an embedding vector $Z \in \mathbb{R}^{c \times d \times p}$ where $d = 512$ and p is the number of tokens. Following this, the batch and channel dimensions are flattened together prior to input into the transformer, creating a tensor

of shape $bs * c \times 1800 \times 512$, where 1800 is the number of tokenized 1-second patches from the 30-minute segment.

3.3.3. Rotary Positional Embeddings

For temporal encoding, rotary positional embeddings (RoPE) were applied (Su et al., 2021) to each patch token. For token i with embedding dimension d, each embedding vector $Z \in \mathbb{R}^d$ is split into two subvectors $\mathbf{h}_j = [h_{2j}, h_{2j+1}], \quad j = 1, \ldots, \frac{d}{2} - 1$. Each even-odd pair in token i is then rotated by an angle $\theta_j = 10000^{-2j/d}$. This is applied independently to each token $\phi_{p,j} = p \cdot \theta_j$ with:

$$\tilde{\mathbf{h}}_{p,j} = \begin{bmatrix} \cos(\phi_{p,j}) - \sin(\phi_{p,j}) \\ \sin(\phi_{p,j}) + \cos(\phi_{p,j}) \end{bmatrix} \mathbf{h}_{p,j}. \tag{2}$$

Effectively, this step captures relative and absolute positional embeddings across and within patches.

3.3.4. Masking and Training Parameters

For training, target masks select 10–30% of patches at random from each flattened batch-channel dimension. Context masks select 10%–40% of the remaining patches. The representation framework was trained for 100 epochs with a one-cycle learning rate scheduler and AdamW optimizer. After training, the context encoder inputs tokenized signal data $Z_{in} \in \mathbb{R}^{c \times d \times p}$ and generates representations $Z_{out} \in \mathbb{R}^{c \times d \times r}$, where r is the learned representation for channel c with dimension d.

3.4. Non-Linear Probing Architecture

After representation learning, non-linear probing classifiers are trained with learned, frozen representations for each task. The context embedding tensor $Z \in \mathbb{R}^{c \times r \times d}$ is passed into an attentive classifier (Assran et al., 2023; Bardes et al., 2024) with flattened batch and channel dimensions. A single learned query vector $q \in \mathbb{R}^d$ is utilized to extract relevant encoding information from the context encoders representations Z. Queries, keys, and values are calculated $Q = qW_q$, $K = ZW_k$, and $V = ZW_v$ for each attention head. Following, attention scores and weights are generated:

$$\mathbf{a_{score}} = \frac{QK^T}{\sqrt{d}}, \quad \mathbf{a_{weights}} = softmax(a_{score}). \tag{3}$$

A final, pooled representation vector is calculated $\tilde{z} = a_{weights}V \in \mathbb{R}^{1 \times d}$. The pooled representation $\tilde{\mathbf{z}}$ is reshaped to $bs \times c \cdot d$ and passed through a final linear layer for binary classification.

Notably, the weights of the context encoder from PhysioJEPA are frozen during this process. Attentive classifiers used 4 attention heads. Classifiers are trained for 20 epochs with a one-cycle learning rate scheduler and the AdamW optimizer.

3.4.1. Augmentation Techniques

During non-linear probing, common augmentation techniques were employed to 30-minute input segments, including random noise, channel dropout, and mixup (Zhang et al., 2018) to better handle class imbalance.

3.5. Supervised and Self-Supervised Comparison Models

For comparison to a supervised model, we trained a fully supervised convolutional classifier presented by Wang et al. (2016) for each task. Additionally, for comparison to other representation learning frameworks, we trained an equivalent PatchTST (Nie et al., 2023) encoder via masked autoregression and another JEPA based signal model, ECG-JEPA (Kim, 2024). For implementation details, see Appendix A.1.

3.6. Clinical Tasks

3.6.1. 5-Minute Hypotensive Risk Estimation

For the first task, we estimated risk of having a hypotensive event at a 5-minute forecast horizon. Hypotensive risk was chosen due to its common occurrence in ICU patients (60-75% develop hypotension (Terwindt et al., 2022)) and the critical need for improved proactive monitoring. A hypotension threshold was defined based on mean arterial pressure (MAP) of ≤ 65 mmHg or systolic blood pressure (SBP) ≤ 90 mmHg at each minute.

A positive hypotensive event was defined as five consecutive minutes below the hypotension thresholds. Thus, a hypotensive event at a 5 minute forecast was derived with data from the 5 to 10 minute interval ahead of the end of the input signal. MAP and SAP were calculated using peak detection algorithms adapted from PhysioNet's wfdb package (Goldberger et al., 2000; Moody et al., 2022). Hypotensive events longer than 5 minutes were treated as a single event, ensuring at least 5 minutes between multiple events for the same patient. Non-hypotensive patients were those with no hypotensive events throughout their entire ICU stay. Data was split into training (80%), validation (10%), and testing (10%) using a proportional, subject-wise data splitter. Additionally, during training a weighted sampler was used to present

	Hypotension				Shock Index			
Data Split	N ICU Stays	N Patients	Positive Events	Negative Events	N ICU Stays	N Patients	Positive Events	Negative Events
MIMIC-III Train	3280	2060	44897 (0.04)	1015429 (0.96)	3264	2062	53897 (0.05)	1087321 (0.95)
MIMIC-III Validation	381	237	5601 (0.05)	116307 (0.95)	335	216	6882 (0.07)	96229 (0.93)
MIMIC-III Test	341	229	4435 (0.04)	106273 (0.96)	413	245	8526 (0.05)	149215 (0.95)
Mount Sinai Bedmaster Dataset	99	99	2638 (0.03)	82626 (0.97)	98	98	1952 (0.05)	37089 (0.95)

Table 1: Hypotension and shock index event statistics for training, validation, testing, and external test sets.

the model with more positive cases, given the high class imbalance.

3.6.2. 5-Minute Shock Index Risk Estimation

For the second task, we estimated risk of elevated shock index (SI) at a 5-minute forecast horizon. SI was chosen as it serves as an important marker for shock and mortality (Koch et al., 2019). SI was computed as the ratio of heart rate to SBP:

$$\text{SI} = \frac{\text{HR}}{\text{SBP}}, \quad \text{Positive class if SI} \geq 0.9 \tag{4}$$

A shock index event was defined as five consecutive minutes above the 0.9 threshold. Heart rate was derived from the ABP signal channel using PhysioNet's peak detection algorithms (Goldberger et al., 2000; Moody et al., 2022). Normal SI values range from 0.5 to 0.7, and higher values are more predictive of adverse outcomes (Cannon et al., 2009). Data was split into training (80%), validation (10%), and testing (10%) using a proportional label, subject-wise data splitter. Again, a weighted sampler was used during training to account for class imbalance.

3.7. External Validation

To assess generalizability, we conducted external validation on the Mount Sinai Bedmaster Dataset collected from 100 randomly selected patient ICU stays (per task) from 6 separate adult ICUs between 2019 and 2024.

3.8. Evaluation

We evaluated PhysioJEPA, PatchTST, ECG-JEPA, and supervised convolutional models for estimating 5-minute risks of hypotension and shock index using the MIMIC-III test set and external Mount Sinai Bedmaster Dataset. Performance metrics included area under the receiver operating characteristic curve (AUROC), average precision, F1, recall, specificity, and sensitivity at 90% and 95% specificity (Sens@90%Spec, Sens@95%Spec, respectively). The latter two measures could be particularly valuable in clinical settings, which require high specificity to reduce false alarms.

4. Results

Of the 5,660 stays from the MIMIC-III Waveform Database with the required three signal channel data (ABP, ECG lead II, PPG), 1,378 were not used for representation learning due to 20% or more constant or NaN values in a single channel (within every 30-minute segment). We found that the majority of the removed samples had $\geq$ 75% constant or NaN values (total of 1,050 samples). Thus, samples removed were majority constant or NaN values. Overall, PhysioJEPA was trained with 356,903 30-minute segments (total of 10,707,090 minutes) across 4,282 ICU stays (N=2,631 patients) with 3-channel signal data. Demographics are shown in Appendix C Table 3. Training was stopped after 100 epochs, and the last model was used for non-linear probing. An overview of the architecture is shown in Figures 1 and 2. Compute resources are detailed in Appendix B.

4.1. Risk Estimation Results

Dataset statistics of both hypotension and SI are reported in Table 1.

Performance of Supervised Convolutional Models: Two fully supervised convolutional classifiers achieved AUROC scores of 0.778 (95% bootstrapped confidence interval [CI]: 0.771–0.784) for 5-minute hypotension (Figure 3A) and 0.950 (95% CI: 0.948–0.952) for 5-minute shock index risk estimation on the held-out test set (Figure 4A).

Performance of ECG-JEPA Models: Non-linear probing with frozen ECG-JEPA encoder representations achieved AUROC scores of 0.729 (95% CI: 0.721–0.738) for 5-minute hypotension (Figure 3A) and 0.923 (95% CI: 0.921–0.925) for 5-minute shock index risk estimation on the held-out test set (Figure 4A).

Performance of PatchTST Models: Non-linear probing with frozen PatchTST encoder representations achieved AUROC scores of 0.867 (95% CI: 0.861–0.871) for 5-minute hypotension (Figure 3A) and 0.956 (95% CI: 0.955–0.958) for 5-minute shock index risk estimation on the held-out test set (Figure 4A).

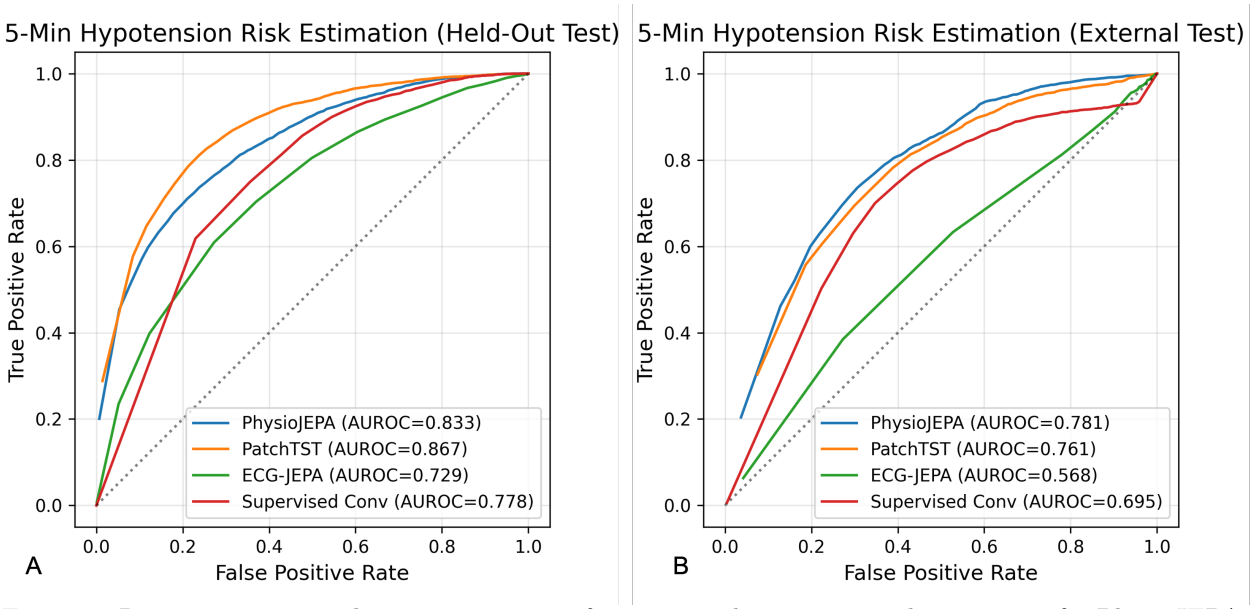

Figure 3: Receiver operating characteristic curves for 5-minute hypotension risk estimation for PhysioJEPA, PatchTST, ECG-JEPA, and the fully supervised convolutional model for held-out (A) and external (B) test sets.

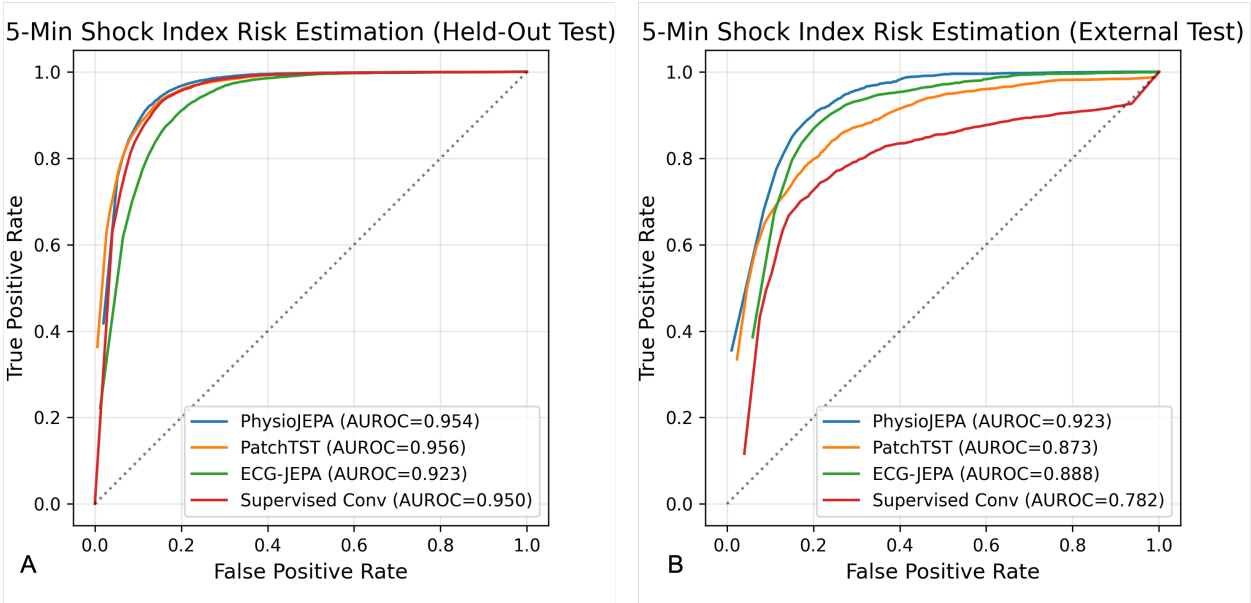

Figure 4: Receiver operating characteristic curves for 5-minute shock index risk estimation for PhysioJEPA, PatchTST, ECG-JEPA, and the fully supervised convolutional model for held-out (A) and external (B) test sets.

Performance of PhysioJEPA Models: Nonlinear probing with frozen PhysioJEPA encoder representations was similar to or outperformed comparison models, achieving AUROC scores of 0.833 (95% CI: 0.825–0.838) for 5-minute hypotension risk (Figure 3A) and 0.954 (95% CI: 0.952–0.955) for 5-minute shock index risk estimation (Figure 4A). For details of other performance metrics of all models, see Table 2.

At clinically relevant specificity thresholds, PhysioJEPA achieved matched or superior performance on the held-out test set (Table 2): 55.2% sensitivity at 90% specificity and 40.7% at 95% specificity for 5-minute hypotension risk estimation, compared to

Condition	Dataset	Model	AUROC	Avg Precision	F1	Recall	Specificity	Sens@90%Spec	Sens@95%Spec
Hypotension	Held-Out	PhysioJEPA	0.833 (0.825 - 0.838)	0.264 (0.254 - 0.277)	**0.122 (0.119 - 0.125)**	0.924 (0.916 - 0.931)	**0.446 (0.443 - 0.449)**	0.552 (0.538 - 0.566)	0.407 (0.391 - 0.421)
		PatchTST	**0.867 (0.861 - 0.871)**	**0.267 (0.257 - 0.282)**	0.105 (0.102 - 0.108)	0.979 (0.974 - 0.984)	0.304 (0.302 - 0.307)	**0.576 (0.565 - 0.616)**	**0.429 (0.416 - 0.442)**
		ECG-JEPA	0.729 (0.721 - 0.738)	0.110 (0.104 - 0.116)	0.077 (0.075 - 0.079)	**1.000 (1.000 - 1.000)**	0.000 (0.000 - 0.000)	0.322 (0.310 - 0.335)	0.146 (0.135 - 0.157)
		Supervised Conv	0.778 (0.771 - 0.784)	0.140 (0.133 - 0.148)	0.082 (0.080 - 0.085)	0.998 (0.997 - 0.999)	0.073 (0.071 - 0.074)	0.422 (0.406 - 0.435)	0.296 (0.282 - 0.310)
	Mount Sinai	PhysioJEPA	**0.781 (0.773 - 0.788)**	**0.098 (0.093 - 0.104)**	**0.084 (0.081 - 0.086)**	0.949 (0.939 - 0.956)	**0.340 (0.337 - 0.343)**	**0.339 (0.319 - 0.356)**	**0.204 (0.191 - 0.218)**
		PatchTST	0.761 (0.753 - 0.769)	0.085 (0.081 - 0.090)	0.075 (0.072 - 0.078)	0.955 (0.947 - 0.963)	0.251 (0.248 - 0.254)	0.303 (0.287 - 0.324)	0.000 (0.000 - 0.000)
		ECG-JEPA	0.568 (0.560 - 0.577)	0.039 (0.037 - 0.041)	0.060 (0.058 - 0.062)	**0.986 (0.981 - 0.991)**	0.018 (0.017 - 0.019)	0.153 (0.059 - 0.168)	0.063 (0.056 - 0.072)
		Supervised Conv	0.695 (0.686 - 0.705)	0.057 (0.054 - 0.060)	0.064 (0.062 - 0.067)	0.917 (0.908 - 0.927)	0.151 (0.149 - 0.154)	0.213 (0.197 - 0.225)	0.069 (0.056 - 0.074)
Shock Index	Held-Out	PhysioJEPA	0.954 (0.952 - 0.955)	0.474 (0.465 - 0.485)	0.308 (0.303 - 0.313)	0.981 (0.978 - 0.983)	0.749 (0.747 - 0.751)	**0.881 (0.875 - 0.889)**	0.706 (0.697 - 0.715)
		PatchTST	**0.956 (0.955 - 0.958)**	**0.595 (0.584 - 0.607)**	**0.391 (0.385 - 0.396)**	0.941 (0.936 - 0.945)	**0.836 (0.834 - 0.838)**	0.873 (0.866 - 0.880)	**0.743 (0.734 - 0.760)**
		ECG-JEPA	0.923 (0.921 - 0.925)	0.364 (0.355 - 0.370)	0.217 (0.213 - 0.221)	**0.986 (0.984 - 0.989)**	0.595 (0.593 - 0.598)	0.726 (0.716 - 0.735)	0.437 (0.425 - 0.447)
		Supervised Conv	0.950 (0.948 - 0.952)	0.555 (0.546 - 0.565)	0.296 (0.292 - 0.300)	0.978 (0.974 - 0.980)	0.735 (0.733 - 0.738)	0.852 (0.846 - 0.859)	0.676 (0.665 - 0.687)
	Mount Sinai	PhysioJEPA	**0.923 (0.917 - 0.927)**	**0.396 (0.381 - 0.410)**	0.226 (0.219 - 0.234)	0.969 (0.963 - 0.976)	0.652 (0.648 - 0.657)	**0.716 (0.699 - 0.738)**	0.356 (0.335 - 0.374)
		PatchTST	0.873 (0.864 - 0.882)	0.312 (0.295 - 0.333)	**0.244 (0.235 - 0.254)**	0.862 (0.847 - 0.877)	**0.727 (0.723 - 0.731)**	0.672 (0.642 - 0.691)	**0.501 (0.477 - 0.520)**
		ECG-JEPA	0.888 (0.882 - 0.894)	0.231 (0.218 - 0.242)	0.142 (0.137 - 0.146)	**0.988 (0.982 - 0.992)**	0.371 (0.365 - 0.375)	0.584 (0.565 - 0.607)	0.000 (0.000 - 0.000)
		Supervised Conv	0.782 (0.771 - 0.794)	0.162 (0.152 - 0.171)	0.167 (0.160 - 0.174)	0.840 (0.828 - 0.853)	0.569 (0.564 - 0.572)	0.522 (0.498 - 0.548)	0.202 (0.185 - 0.218)

Table 2: Comprehensive performance comparison of PhysioJEPA, PatchTST, ECG-JEPA, and the fully supervised convolutional model on hypotension and shock index risk estimation tasks across the held-out MIMIC-III test set and external Mount Sinai Bedmaster Dataset. Best performance for each metric is shown in bold (95% bootstrapped confidence intervals).

PatchTST performance of 57.6% and 42.9%, ECG-JEPA performance of 32.2% and 14.6%, and the supervised convolutional model performance of 42.2% and 29.6%, respectively. For shock index, PhysioJEPA achieved 88.1% and 70.6% sensitivity at 90% and 95% specificity thresholds, similar to PatchTST results of 87.3% and 74.3%, and outperforming ECG-JEPA results of 72.6% and 43.7% and supervised convolutional classifier results of 85.2% and 67.6%, respectively.

4.2. External Validation Results

External validation on 100 randomly selected patient-ICU stays (for each task) from the Mount Sinai Bedmaster Dataset demonstrated strong cross-site generalizability without retraining. The external test set comprised data from 6 different adult ICUs. Demographics are shown in Appendix C Table 4. 3 samples were excluded due to our missingness constraints as detailed in the Methods. 5-minute hypotension risk estimation using the PhysioJEPA architecture and classifier trained on MIMIC-III data achieved an AUROC of 0.781 (95% CI: 0.773–0.788), compared to 0.761 (95% CI: 0.753–0.769) for PatchTST, 0.568 (95% CI: 0.560–0.577) for ECG-JEPA, and 0.695 (95% CI: 0.686–0.705) for the supervised convolutional classifier (Figure 3B). Similarly, 5-minute shock index risk estimation achieved an AUROC of 0.923 (95% CI: 0.917–0.927), compared to 0.873 (95% CI: 0.864–0.882) for PatchTST, 0.888 (95% CI: 0.882–0.894) for ECG-JEPA, and 0.782 (95% CI: 0.771–0.794) for the supervised convolutional classifier (Figure 4B).

At clinically relevant specificity thresholds, PhysioJEPA achieved consistent high performance across both tested thresholds, 33.9% and 20.4% sensitivity at 90% and 95% specificity for 5-minute hypotension risk estimation, compared to PatchTST performance of 30.3% and 0.00%, ECG-JEPA performance of 15.3% and 6.30%, and the supervised convolutional model performance of 21.3% and 6.90%, respectively. For shock index, PhysioJEPA achieved 71.6% and 35.6% sensitivity at 90% and 95% specificity thresholds, compared to PatchTST results of 67.2% and 50.1%, ECG-JEPA results of 58.4% and 0.00%, and supervised convolutional model results of 52.2% and 20.2%, respectively (Table 2).

5. Discussion and Conclusion

PhysioJEPA demonstrates that JEPA-based self-supervised learning effectively captures complex multi-modal physiological signal relationships for critical care risk estimation. The strong performance on both tasks establishes the potential for real-world ICU deployment and cross-site adaptability. Compared to PatchTST, which has been established as a multi-modal physiological signal representation model (Fox et al., 2025), PhysioJEPA's primary advantage lies in improved performance on the external test set in terms of AUROC and average precision across both tasks. For more clinically relevant thresholds, PhysioJEPA outperforms PatchTST for hypotension and shock index risk estimation at 90% specificity. At 95% specificity, PhysioJEPA outperforms PatchTST for hypotension risk estimation, but under performs on the shock index outcome. Compared to the supervised convolutional model and ECG-JEPA, which has been established as a single modality JEPA-based physiological signal representation model (Kim, 2024), PhysioJEPA consistently matches or outperforms across both test sets for hypotension and shock index risk estimation. In general, PhysioJEPA maintains consistently high performance across both risk estimation tasks, in comparison to other supervised and self-supervised

models. This warrants further investigation into additional risk estimation outcomes to distinguish how performance across these models vary.

Clinical Impact: The 5-minute risk estimation horizon provides actionable early warning capabilities for preventing adverse events. PhysioJEPA's performance demonstrates the clinical value of self-supervised representation learning. The model relies solely on bedside monitoring data, enabling deployment across diverse ICU environments without requiring electronic health record integration.

Technical Contributions: Our work extends JEPA to long-duration, high frequency, multi-modal physiological time series and adds improvements to better handle signals independently (depthwise convolutions, rotary embeddings, per channel masked tokens, and flattened batch-channel dimensions for attentive classification). This adds to other physiological signal JEPA approaches to handle multi-modal signals, instead of only ECG (Kim, 2024) or EEG (Guetschel et al., 2024). Additionally, PhysioJEPA is to our knowledge, the first self-supervised representation learning model to be developed for ICU bedside monitoring data. Our results show that the embedding-space prediction method can learn relevant features of physiological signal data and generalize equally or better to unseen datasets for predictive tasks, compared to PatchTST, ECG-JEPA, and a fully supervised convolutional method. Furthermore, given that the representations are generated without being fine-tuned to specific tasks, they may be applicable to other critical care outcomes, which merits further investigation.

Comparison to Existing Approaches: Our work adds to the growing literature on hypotensive risk estimation by introducing a self-supervised framework that learns directly from raw bedside signals. Prior approaches using raw bedside signals trained fully supervised models to estimate hypotensive risk (Lee et al., 2021; Moon et al., 2024; Jeong et al., 2024; Jo et al., 2022). Lee et al. (2021), Moon et al. (2024), and Jo et al. (2022) used multi-modal signals to estimate 3-, 5-, 10-, and/or 15-minute hypotensive risk on the VitalDB (https://vitaldb.net/) database, without external validation. Jeong et al. (2024) trained a fully supervised model using VitalDB with ECG, PPG, capnography, bispectral index, and non-invasive ABP to estimate intraoperative hypotension at a 5-minute forecast with good performance on an external test set (AUROC: 0.833 (95% CI, 0.830–0.836)). In contrast, PhysioJEPA learns task-agnostic features by learning representations directly from multi-modal bedside signals via self-supervision and demonstrates competitive performance across internal and external datasets. To our knowledge, we are among the first to predict shock index using a self-supervised representation learning approach for raw multi-modal physiological data.

Limitations: First, the model currently requires invasive ABP monitoring, which may not be available for all patients or institutions. Second, the evaluation was conducted on two specific critical care outcomes; broader validation across additional clinical endpoints is needed to establish general applicability. Third, the 30-minute input window needs to be investigated further along with the missingness threshold of 20%, as the tradeoff between these two elements could be optimized to reduce noise and better represent longer-term physiological patterns. Fourth, the external validation was conducted on a relatively small subset of patients from a single additional institution, and broader multi-site validation would strengthen generalizability claims. Fifth, optimal masking ratios for context and targets was not explored. Adding more context (like in PatchTST or ECG-JEPA) could improve learned representations and performance on downstream tasks. Finally, an additional transformer-based supervised classifier method could be explored to better understand performance, given both PhysioJEPA, PatchTST, and ECG-JEPA encoded signals with a transformer backbone.

Future Work: Evaluating non-invasive risk estimation by removing the ABP channel would significantly expand the patient population that could benefit from this approach. Exploring longer forecast horizons (10-15 minutes) and varying input segment lengths could optimize the balance between early warning capability and prediction accuracy. Exploration of alternative preprocessing techniques, integration of additional physiological channels, masking ratios, and adoption of task-specific fine-tuning strategies or different classifier heads could further improve performance. Comparative evaluation against other self-supervised learning paradigms, such as a contrastive learning method, could further validate the effectiveness of joint embedding predictive architectures for physiological time series data. Additionally, applying the learned representations to other critical care outcomes such as sepsis onset, respiratory failure, or cardiac arrest would demonstrate broader utility.

Acknowledgments

This work was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL175992 and the National Center for Advancing Translational Sciences of the National Institutes of Health under award numbers TL1TR004420. Further, this work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. apr 13 2023. doi: 10.48550/arXiv.2301.08243. URL http://arxiv.org/abs/2301.08243. arXiv:2301.08243 [cs].

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video, February 2024. URL http://arxiv.org/abs/2404.08471. arXiv:2404.08471 [cs].

Andreas Brink-Kjaer, Eileen B. Leary, Haoqi Sun, M. Brandon Westover, Katie L. Stone, Paul E. Peppard, Nancy E. Lane, Peggy M. Cawthon, Susan Redline, Poul Jennum, Helge B. D. Sorensen, and Emmanuel Mignot. Age estimation from sleep studies using deep learning predicts life expectancy. *npj Digital Medicine*, 5(1):1–10, July 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00630-9. URL https://www.nature.com/articles/s41746-022-00630-9. Publisher: Nature Publishing Group.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Chad M. Cannon, Carla C. Braxton, Mendy Kling-Smith, Jonathan D. Mahnken, Elizabeth Carlton, and Michael Moncure. Utility of the shock index in predicting mortality in traumatically injured patients. *The Journal of Trauma*, 67(6): 1426–1430, December 2009. ISSN 1529-8809. doi: 10.1097/TA.0b013e3181bbf728.

Ményssa Cherifa, Yannet Interian, Alice Blet, Matthieu Resche-Rigon, and Romain Pirracchio. The Physiological Deep Learner: First application of multitask deep learning to predict hypotension in critically ill patients. *Artificial Intelligence in Medicine*, 118:102118, August 2021. ISSN 0933-3657. doi: 10.1016/j.artmed.2021.102118. URL https://www.sciencedirect.com/science/article/pii/S0933365721001111.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked Auto-encoder for EEG Representation Learning, October 2022. URL http://arxiv.org/abs/2211.02625. arXiv:2211.02625 [eess].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. may 24 2019. doi: 10.48550/arXiv.1810.04805. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J Lee, Cynthia Rudin, and Xiao Hu. Siamquality: A convnet-based foundation model for imperfect physiological signals. arXiv preprint arXiv:2404.17667, 2024. https://arxiv.org/abs/2404.17667.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing Self-supervised EEG

Representation Through Informative Masked Inputs, June 2024. URL http://arxiv.org/abs/2402.17772. arXiv:2402.17772 [eess].

Benjamin Fox, Joy Jiang, Sajila Wickramaratne, Patricia Kovatch, Mayte Suarez-Farinas, Neomi A Shah, Ankit Parekh, and Girish N Nadkarni. A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. *Sleep*, page zsaf061, March 2025. ISSN 0161-8105. doi: 10.1093/sleep/zsaf061. URL https://doi.org/10.1093/sleep/zsaf061.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), June 2000. ISSN 1524-4539. doi: 10.1161/01.cir.101.23.e215. URL http://dx.doi.org/10.1161/01.cir.101.23.e215.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-jepa: Towards seamless cross-dataset transfer through dynamic spatial attention. arXiv preprint arXiv:2403.11772, 2024. https://arxiv.org/abs/2403.11772.

Feras Hatib, Zhongping Jian, Sai Buddi, Christine Lee, Jos Settels, Karen Sibert, Joseph Rinehart, and Maxime Cannesson. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology*, 129(4):663–674, October 2018. ISSN 1528-1175. doi: 10.1097/ALN.0000000000002300.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Heejoon Jeong, Donghee Kim, Dong Won Kim, Seungho Baek, Hyung-Chul Lee, Yusung Kim, and Hyun Joo Ahn. Prediction of intraoperative hypotension using deep learning models based on non-invasive monitoring devices. *Journal of Clinical Monitoring and Computing*, 38(6):1357–1365, December 2024. ISSN 1573-2614. doi: 10.1007/s10877-024-01206-6. URL https://doi.org/10.1007/s10877-024-01206-6.

Zhongping Jian, Xianfu Liu, Karim Kouz, Jos J. Settels, Simon Davies, Thomas W. L. Scheeren, Neal W. Fleming, Denise P. Veelo, Alexander P. J. Vlaar, Michael Sander, Maxime Cannesson, David Berger, Michael R. Pinsky, Daniel I. Sessler, Feras Hatib, and Bernd Saugel. Deep learning model to identify and validate hypotension endotypes in surgical and critically ill patients. *British Journal of Anaesthesia*, 134 (2):308–316, February 2025. ISSN 0007-0912, 1471-6771. doi: 10.1016/j.bja.2024.10.048. URL https://www.bjanaesthesia.org/article/S0007-0912(24)00712-8/fulltext. Publisher: Elsevier.

Yong-Yeon Jo, Jong-Hwan Jang, Joon-myoung Kwon, Hyung-Chul Lee, Chul-Woo Jung, Seonjeong Byun, and Han-Gil Jeong. Predicting intraoperative hypotension using deep learning with waveforms of arterial blood pressure, electroencephalogram, and electrocardiogram: Retrospective study. *PLOS ONE*, 17(8):e0272055, August 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0272055. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272055. Publisher: Public Library of Science.

Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-III Clinical Database, 2015. URL https://physionet.org/content/mimiciii/1.4/. [Online; accessed 2024-03-06].

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-III, a freely accessible critical care database. *Scientific Data*, 3 (1):160035, may 24 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. number: 1 publisher: Nature Publishing Group.

Lorenz Kapral, Christoph Dibiasi, Natasa Jeremic, Stefan Bartos, Sybille Behrens, Aylin Bilir, Clemens Heitzinger, and Oliver Kimberger. Development and external validation of temporal fusion transformer models for continuous intraoperative blood pressure forecasting. *eClinicalMedicine*, 75, September 2024. ISSN 2589-5370. doi: 10.1016/j.eclinm.2024.102797. URL https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(24)00376-6/fulltext. Publisher: Elsevier.

Sehun Kim. Learning General Representation of 12-Lead Electrocardiogram with a Joint-Embedding Predictive Architecture, December 2024. URL http://arxiv.org/abs/2410.08559. arXiv:2410.08559 [cs].

Vladislav Kim, Lisa Schneider, Soodeh Kalaie, Declan O'Regan, and Christian Bender. Heartmae: Advancing cardiac mri analysis through optical flow guided masked autoencoding. *Proceedings of Machine Learning Research*, 259:594–609, 2024.

Erica Koch, Shannon Lovett, Trac Nghiem, Robert A Riggs, and Megan A Rech. Shock index in the emergency department: utility and limitations. *Open Access Emergency Medicine : OAEM*, 11: 179–199, August 2019. ISSN 1179-1500. doi: 10.2147/OAEM.S178358. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6698590/.

Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. SplitSEE: A Splittable Self-supervised Framework for Single-Channel EEG Representation Learning, October 2024. URL http://arxiv.org/abs/2410.11200. arXiv:2410.11200 [cs].

Solam Lee, Hyung-Chul Lee, Yu Seong Chu, Seung Woo Song, Gyo Jin Ahn, Hunju Lee, Sejung Yang, and Sang Baek Koh. Deep learning models for the prediction of intraoperative hypotension. *British Journal of Anaesthesia*, 126(4):808–817, April 2021. ISSN 0007-0912. doi: 10.1016/j.bja.2020.12.035. URL https://www.sciencedirect.com/science/article/pii/S0007091221000027.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals (normwear). arXiv preprint arXiv:2412.09758, 2024. https://arxiv.org/pdf/2412.09758.

Guy Lutsker, Gal Sapir, Anastasia Godneva, Smadar Shilo, Jerry R. Greenfield, Dorit Samocha-Bonet, Shie Mannor, Eli Meirom, Gal Chechik, Hagai Rossman, and Eran Segal. From Glucose Patterns to Health Outcomes: A Generalizable Foundation Model for Continuous Glucose Monitor Data Analysis. aug 20 2024. doi: 10.48550/arXiv.2408.11876. URL http://arxiv.org/abs/2408.11876. arXiv:2408.11876 [cs, q-bio].

Kamal Maheshwari, Sai Buddi, Zhongping Jian, Jos Settels, Tetsuya Shimada, Barak Cohen, Daniel I. Sessler, and Feras Hatib. Performance of the Hypotension Prediction Index with non-invasive arterial pressure waveforms in non-cardiac surgical patients. *Journal of Clinical Monitoring and Computing*, 35(1):71–78, feb 1 2021. ISSN 1573-2614. doi: 10.1007/s10877-020-00463-5.

Mike A. Merrill and Tim Althoff. Self-supervised Pretraining and Transfer Learning Enable Flu and COVID-19 Predictions in Small Mobile Sensing Datasets. jun 2 2022. doi: 10.48550/arXiv.2205.13607. URL http://arxiv.org/abs/2205.13607. arXiv:2205.13607 [cs].

Alistair Miles, jakirkham, M. Bussonnier, Josh Moore, Dimitri Papadopoulos Orfanos, Davis Bennett, David Stansby, Joe Hamman, James Bourbeau, Andrew Fulton, Gregory Lee, Ryan Abernathey, Norman Rzepka, Zain Patel, Mads R. B. Kristensen, Sanket Verma, Saransh Chopra, Matthew Rocklin, AWA BRANDON AWA, Max Jones, Martin Durant, Elliott Sales Andrade, Vincent Schut, raphael dussin, Shivank Chaudhary, Chris Barnes, Juan Nunez-Iglesias, and shikharsg. zarr-developers/zarr-python: v3.0.0-alpha, jun 12 2024. URL https://zenodo.org/records/11592827. DOI: 10.5281/zenodo.11592827.

Mina Chookhachizadeh Moghadam, Ehsan Masoumi, Samir Kendale, and Nader Bagherzadeh. Predicting hypotension in the ICU using noninvasive physiological signals. *Computers in Biology and Medicine*, 129:104120, feb 1 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2020.104120.

George Moody, Tom Pollard, and Benjamin Moody. Wfdb software package, 2022. URL https://physionet.org/content/wfdb/10.7.0/.

Jeong-Hyeon Moon, Garam Lee, Seung Mi Lee, Jiho Ryu, Dokyoon Kim, and Kyung-Ah Sohn. Frequency Domain Deep Learning With Non-Invasive Features for Intraoperative Hypotension Prediction. *IEEE Journal of Biomedical and Health Informatics*, 28(10):5718–5728, October 2024. ISSN 2168-2208. doi: 10.1109/JBHI.2024.3403109. URL https://ieeexplore.ieee.org/document/10535187.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec,

Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 4 2023a. ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4. publisher: Nature Publishing Group.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. *Proceedings of Machine Learning Research*, 225:353–367, 2023b.

Girish Narayanswamy et al. Scaling wearable foundation models. arXiv preprint arXiv:2410.13638, 2024. https://arxiv.org/abs/2410.13638.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. mar 5 2023. doi: 10.48550/arXiv.2211.14730. URL http://arxiv.org/abs/2211.14730. arXiv:2211.14730 [cs].

Christina Orphanidou. A review of big data applications of physiological signal data. *Biophysical Reviews*, 11(1):83–87, February 2019. ISSN 1867-2469. doi: 10.1007/s12551-018-0495-3. URL https://doi.org/10.1007/s12551-018-0495-3.

Christina Orphanidou and David Wong. Machine Learning Models for Multidimensional Clinical Data. In Samee U. Khan, Albert Y. Zomaya, and Assad Abbas, editors, *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, pages 177–216. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58280-1. doi: 10.1007/978-3-319-58280-1_8. URL https://doi.org/10.1007/978-3-319-58280-1_8.

Sydney R. Rooney and Gilles Clermont. Forecasting algorithms in the ICU. *Journal of Electrocardiology*, 81:253–257, November 2023. ISSN 0022-0736. doi: 10.1016/j.jelectrocard.2023.09.015. URL https://www.sciencedirect.com/science/article/pii/S0022073623002248.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. URL https://arxiv.org/abs/2104.09864.

Sameer Sundrani, Julie Chen, Boyang Tom Jin, Zahra Shakeri Hossein Abad, Pranav Rajpurkar, and David Kim. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. *npj Digital Medicine*, 6(1):60, April 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00803-0. URL https://www.nature.com/articles/s41746-023-00803-0. Publisher: Nature Publishing Group.

Lotte E. Terwindt, Jaap Schuurmans, Björn J. P. van der Ster, Carin A. G. C. L. Wensing, Marijn P. Mulder, Marije Wijnberge, Thomas G. V. Cherpanath, Wim K. Lagrand, Alain A. Karlas, Mark H. Verlinde, Markus W. Hollmann, Bart F. Geerts, Denise P. Veelo, and Alexander P. J. Vlaar. Incidence, Severity and Clinical Factors Associated with Hypotension in Patients Admitted to an Intensive Care Unit: A Prospective Observational Study. *Journal of Clinical Medicine*, 11(22):6832, nov 18 2022. ISSN 2077-0383. doi: 10.3390/jcm11226832. PMID: 36431308 PMCID: PMC9696980.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals. may 27 2024. doi: 10.48550/arXiv.2405.17766. URL http://arxiv.org/abs/2405.17766. arXiv:2405.17766 [cs, eess].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. dec 5 2017. doi: 10.48550/arXiv.1706.03762. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs] version: 5.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. *CoRR*, abs/1611.06455, 2016. URL http://arxiv.org/abs/1611.06455.

Kuba Weimann and Tim O.F. Conrad. Self-supervised pre-training with joint-embedding

predictive architecture boosts ecg classification performance. *Computers in Biology and Medicine*, 196:110809, 2025. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2025.110809. URL https://www.sciencedirect.com/science/article/pii/S0010482525011606.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. Ehrshot: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. dec 11 2023. doi: 10.48550/arXiv.2307.02028. URL http://arxiv.org/abs/2307.02028. arXiv:2307.02028.

Maxwell A. Xu, Alexander Moreno, Hui Wei, Benjamin M. Marlin, and James M. Rehg. REBAR: Retrieval-Based Reconstruction for Time-series Contrastive Learning, October 2024. URL http://arxiv.org/abs/2311.00519. arXiv:2311.00519 [cs].

Maxwell A. Xu, Girish Narayanswamy, et al. Lsm-2: Learning from incomplete wearable sensor data. arXiv preprint arXiv:2506.05321, 2025. https://arxiv.org/abs/2506.05321.

Joo Heung Yoon, Vincent Jeanselme, Artur Dubrawski, Marilyn Hravnak, Michael R. Pinsky, and Gilles Clermont. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. *Critical Care*, 24(1): 1–9, 12 2020. ISSN 1364-8535. doi: 10.1186/s13054-020-03379-3. number: 1 publisher: BioMed Central.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8980–8987, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. apr 27 2018. doi: 10.48550/arXiv.1710.09412. URL http://arxiv.org/abs/1710.09412. arXiv:1710.09412 [cs].

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. oct 15 2022. doi: 10.48550/arXiv.2206.08496. URL http://arxiv.org/abs/2206.08496. arXiv:2206.08496 [cs].

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 10 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x. publisher: Nature Publishing Group.

Appendix A. Supplemental Methods

A.1. Supervised Classifier, PatchTST, and ECG-JEPA Implementation Details

Two fully supervised convolutional classifiers were trained to estimate the two outcomes for comparison to PhysioJEPA. Each used a three-layer convolution with standard parameters (dimensions: 128, 256, and 128; kernel sizes: 7, 5, and 3). Classifiers were trained for 20 epochs with a one-cycle learning rate scheduler and AdamW optimizer.

Two other representation learning frameworks were also trained for comparison to PhysioJEPA. PatchTST (Nie et al., 2023) has been utilized in other multi-modal physiological signal representation models for Sleep (Fox et al., 2025), thus making it a good self-supervised, multi-modal comparison. Like the PhysioJEPA context encoder, PatchTST was built with identical tokenization, positional encoding, and encoder parameters. Contrary to PhysioJEPA, masking was performed prior to tokenization, as detailed in the PatchTST architecture. A "target" masking ratio of 10% to 30%, equivalent to PhysioJEPA, was used and the model was trained to recreate the values from these masked out patches with a linear layer per channel and mean squared error loss function. This self-supervised training technique is commonly known as masked autoregression. PatchTST was trained for 100 epochs with a one-cycle learning rate scheduler, and AdamW optimizer.

For comparison to a JEPA representation framework, we trained ECG-JEPA (Kim, 2024) with our multi-modal input signals. ECG-JEPA was originally designed for ECG signals (a single modality) and employs a cross pattern attention mechanism to learn relationships among channels. Encoder size and number of heads were identical to PhysioJEPA with sinusoidal positional encodings. A target mask ratio of 10% to 30% of the total number of patches was utilized, equivalent to PhysioJEPA. All other patches not selected as targets were used as the context per the original ECG-JEPA implementation. ECG-JEPA was trained for 100 epochs with a one-cycle learning rate scheduler and AdamW optimizer. A smooth L1 loss function was used for optimization similar to the original implementation.

Following self-supervised learning, two additional attentive classifiers were trained for each task for each framework as described in the Methods section using the representations generated from PatchTST and ECG-JEPA.

Appendix B. Compute Resources

Self-supervised models were trained with two Nvidia H100 NVLink GPUs and 16 cores. Classifier models were trained with one Nvidia H100 NVLink GPU and 16 cores.

Appendix C. Demographic Characteristics of Datasets

		Hypotension	Non-hypotension	Shock Index $\geq$ 0.9	Shock Index $<$ 0.9
	Total Patients	1634	892	1578	945
Gender	Female	715 (43.8%)	369 (41.4%)	700 (44.4%)	385 (40.7%)
	Male	919 (56.2%)	523 (58.6%)	878 (55.6%)	560 (59.3%)
Age	18-39	92 (5.6%)	112 (12.6%)	136 (8.6%)	68 (7.2%)
	40-59	427 (26.1%)	321 (36.0%)	457 (29.0%)	295 (31.2%)
	60-79	769 (47.1%)	357 (40.0%)	712 (45.1%)	408 (43.2%)
	80+	276 (16.9%)	84 (9.4%)	221 (14.0%)	137 (14.5%)
	Unknown	70 (4.3%)	18 (2.0%)	52 (3.3%)	37 (3.9%)

Table 3: Age and gender of patients from the MIMIC-III dataset separated by outcome.

		Hypotension	Non-hypotension	Shock Index $\geq$ 0.9	Shock Index $<$ 0.9
	Total Patients	63	36	69	29
Gender	Female	36 (57.1%)	15 (41.7%)	38 (55.1%)	14 (48.3%)
	Male	27 (42.9%)	21 (58.3%)	31 (44.9%)	15 (51.7%)
Age	18-39	10 (15.9%)	11 (30.6%)	12 (17.4%)	3 (10.3%)
	40-59	18 (28.6%)	7 (19.4%)	17 (24.6%)	10 (34.5%)
	60-79	28 (44.4%)	13 (36.1%)	33 (47.8%)	10 (34.5%)
	80+	7 (11.1%)	5 (13.9%)	7 (10.1%)	6 (20.7%)

Table 4: Age and gender of patients from the Mount Sinai Bedmaster Dataset separated by outcome. Data was collected across 6 different adult ICUs.