Foundation Models for Hemodynamic Time Series: A New Paradigm in Cardiovascular Data Modeling

Khuong Vo* **Arash Shokouhmand** **Arghavan Arafati** **Cristhian Potes**
Advanced Patient Monitoring, Becton Dickinson

Abstract

Hemodynamic waveforms encode rich physiological signals essential for cardiovascular assessment, but scalable interpretation has been limited by the need for labeled data and costly imaging. Leveraging approximately 34,000 hours of finger-cuff and arterial blood pressure waveforms from about 12,000 subjects, collected using Edwards Lifesciences ClearSight and FloTrac devices, we develop a transformer-based foundation model that learns robust representations of cardiovascular dynamics. Trained with self-supervised learning, the model achieves sample-efficient performance, matching state-of-the-art benchmarks while using only 30% of labeled data for detecting aortic stenosis and reduced left ventricular ejection fraction. To our knowledge, this is the first foundation model trained solely on blood pressure waveforms for cardiovascular disease screening.

1 Introduction

The modeling of hemodynamic waveforms represents a fundamental pillar in cardiovascular assessment, as these continuous pressure traces contain rich physiological information that extends far beyond simple systolic and diastolic values. Blood pressure waveforms encode critical details about cardiac contractility, vascular compliance, wave reflection patterns, and arterial stiffness – parameters that are essential for understanding the underlying pathophysiology of cardiovascular disease and guiding optimal therapeutic interventions [1, 2]. In recent years, foundation models have triggered a paradigm shift in healthcare [3, 4], moving beyond task-specific algorithms to versatile, adaptable systems trained on massive and diverse datasets. These models leverage self-supervised learning to develop rich representations that can be rapidly adapted to new clinical tasks with minimal additional training. This capability is particularly valuable in healthcare, where labeled data is often scarce, expensive to obtain, and subject to privacy constraints.

In the context of hemodynamic monitoring, foundation models can transform raw physiological signals into complex temporal and morphological biomarkers, facilitating early detection of cardiovascular deterioration, personalized risk stratification, and precise hemodynamic optimization strategies. In this work, we introduce a novel hemodynamic foundation model (Figure 1) inspired by recent advances in large-scale time-series modeling [5, 6, 7, 8] to address two of the most challenging cardiovascular conditions: aortic stenosis [9] and reduced left ventricular ejection fraction [10]. Uniquely, our approach performs disease detection using only continuous, noninvasive arterial pressure waveforms obtained via the Edwards Lifesciences Clearsight finger-cuff, eliminating the need for costly, operator-dependent modalities such as echocardiography or cardiac MRI. By enabling automated, real-time waveform analysis at the point of care, this approach promises rapid, cost-effective screening and early intervention to mitigate the progression of cardiovascular dysfunction.

*Correspondence to: Khuong Vo <anthony.vo@bd.com>.

Learning from Time Series for Health Workshop, 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

2 Methodology

Our goal is to train a foundational model for blood pressure waveforms by designing an encoder $F : \mathbb{R}^T \to \mathbb{R}^Q$ that converts a fixed-length univariate time series $\mathbf{x} \in \mathbb{R}^T$ into a latent vector of dimension Q. During the pre-training phase, we utilize a large unlabeled dataset X_0 to learn task-agnostic representations. Then, in the downstream phase, given a labeled dataset (X, Y), we generate embeddings $\mathrm{Z} = \{F(\mathbf{x}) \mid \mathbf{x} \in \mathrm{X}\}$, which are used to train a classifier $h : \mathbb{R}^Q \to \{1, \ldots, K\}$.

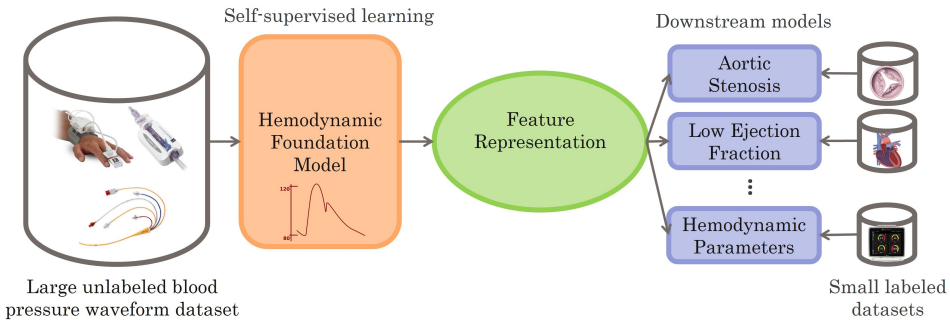

Figure 1: Hemodynamic foundation modeling.

2.1 Pressure-Aware Tokenization

Following the principle of NuTime [8], we normalize each window to stabilize gradient propagation, while separately embedding the mean and standard deviation to preserve physiologically meaningful pressure values. Given a univariate pressure waveform $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, we divide it into N non-overlapping windows. For each window $\mathbf{x}^{(i)}, i = 1, \ldots, N$, we compute its normalized shape $\hat{\mathbf{x}}^{(i)} = \frac{\mathbf{x}^{(i)} - \mu_i}{\sigma_i}$. The normalized shape vector $\hat{\mathbf{x}}^{(i)}$, the scalar mean μ_i, and the scalar standard deviation σ_i are embedded independently with a linear layer followed by layer normalization:

$$\mathbf{e}_i^{\text{shape}} = \text{LayerNorm}(\mathbf{W}^{\text{shape}} \hat{\mathbf{x}}^{(i)} + \mathbf{b}^{\text{shape}}),$$
$$\mathbf{e}_i^{\mu} = \text{LayerNorm}(\mathbf{W}^{\mu} \mu_i + \mathbf{b}^{\mu}), \quad \mathbf{e}_i^{\sigma} = \text{LayerNorm}(\mathbf{W}^{\sigma} \sigma_i + \mathbf{b}^{\sigma}).$$

These embeddings are concatenated to form the initial token $t_i^{(0)} = [\mathbf{e}_i^{\text{shape}}; \mathbf{e}_i^{\mu}; \mathbf{e}_i^{\sigma}]$. An additional linear layer with layer normalization is then applied: $t_i = \text{LayerNorm}(\mathbf{W}^{\text{final}} t_i^{(0)} + \mathbf{b}^{\text{final}})$. Finally, positional encoding $\mathbf{p}_i$ is added to each token to incorporate temporal order, yielding the final token sequence $\mathbf{T} = [t_1 + \mathbf{p}_1; t_2 + \mathbf{p}_2; \ldots; t_N + \mathbf{p}_N]$.

2.2 Masked Waveform Pre-Training with Transformers

We feed the token sequence into a Flan-T5 transformer encoder, which has been successfully applied to time-series data in models such as MOMENT [7]. The Flan-T5 encoder, composed of stacked self-attention layers and feed-forward networks, captures both local and global temporal dependencies by learning contextual representations over the entire sequence.

The core of the transformer encoder is the multi-head self-attention mechanism. For an input token sequence $\mathbf{T} \in \mathbb{R}^{N \times d}$, the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are computed as linear projections: $\mathbf{Q} = \mathbf{T}\mathbf{W}^Q, \mathbf{K} = \mathbf{T}\mathbf{W}^K, \mathbf{V} = \mathbf{T}\mathbf{W}^V$, where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_k}$ are learned parameter matrices. The scaled dot-product attention is then computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

Multi-head attention concatenates the outputs of h separate attention heads: $\text{MultiHead}(\mathbf{T}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$ where each head is $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$, and $\mathbf{W}^O \in \mathbb{R}^{hd_k \times d}$ is a learned projection matrix. Following the multi-head attention, position-wise multilayer

perceptron (MLP) are applied independently to each token:

$$\mathrm{MLP}(x) = \mathrm{GELU}(x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{mlp}}, \mathbf{W}_2 \in \mathbb{R}^{d_{mlp} \times d}$. Layer normalization and residual connections are employed around both attention and feed-forward sublayers to stabilize training $\mathbf{Z}^{(l)} = \mathrm{LayerNorm}\big(\mathbf{Z}^{(l-1)} + \mathrm{MultiHead}(\mathbf{Z}^{(l-1)})\big)$, and $\mathbf{Z}^{(l)} = \mathrm{LayerNorm}\big(\mathbf{Z}^{(l)} + \mathrm{MLP}(\mathbf{Z}^{(l)})\big)$ where $\mathbf{Z}^{(0)} = \mathbf{T}$.

To train the model in a self-supervised manner, a fraction of input tokens are masked, and the model predicts the masked values from the surrounding context. After passing through L such Transformer layers, the contextual embeddings $\mathbf{Z}^{(L)}$ corresponding to masked positions are passed through a position-wise linear decoder. Let $\mathbf{T}_{\text{mask}}$ denote the masked tokens, $\hat{\mathbf{T}}_{\text{mask}}$ the corresponding model predictions, and $\mathcal{M}$ the set of masked indices. The masked modeling loss is defined as:

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \mathbf{T}_{\text{mask}}^{(i)} - \hat{\mathbf{T}}_{\text{mask}}^{(i)} \right\|^2$$

2.3 Downstream Model Training with Stochatic Weight Averaging

During downstream model training, the pretrained embeddings $\mathbf{Z}$ are fed into a MLP. We do not use a linear model because the pretrained features are strong but inherently nonlinear [11]. However, the induced loss landscape over these fixed embeddings can be relatively sharp and multimodal: foundation-model features may cluster in high-curvature regions that create multiple narrow local optima, especially when labeled data are limited. To address this, we apply Stochastic Weight Averaging (SWA) [12], which averages model weights sampled from different points along the optimization trajectory. This procedure can smooth over the rugged loss surface defined by the frozen features, pulling the solution toward flatter, wider basins that generalize better and avoid overfitting to spurious local minima. As a result, SWA provides a more stable and robust downstream classifier, enabling a single lightweight architecture to transfer reliably across diverse clinical tasks. The running average weight $\bar{\mathbf{w}}$ at step m is updated as:

$$\bar{\mathbf{w}}_m = \frac{m-1}{m}\bar{\mathbf{w}}_{m-1} + \frac{1}{m}\mathbf{w}_m$$

Weight averaging is performed in the later phase of training by traveling in small steps along connected paths of low loss between different models, enabling effective ensembling of diverse solutions. Using $\bar{\mathbf{w}}$ as final weights leads to flatter minima with better generalization.

We use focal loss to handle class imbalance and focus learning on hard examples. For binary classification, the focal loss is:

$$\mathcal{L}_{\text{focal}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where p_t is the predicted probability for the true class, α_t balances class weights, and γ controls down-weighting of easy samples. Focal loss not only improves accuracy but also helps reduce model overconfidence, resulting in better calibrated predictions [13].

3 Experimental Results

3.1 Datasets

Pre-Training Dataset: We curated a large-scale dataset of arterial blood pressure waveforms collected over the course of more than a decade using Edwards Lifesciences ClearSight (non-invasive finger-cuff) and FloTrac (invasive arterial line) monitoring systems. The dataset comprises 12,267,124 waveform segments, each 10 seconds long, from 11,967 unique subjects, spanning a total of 34,075 recording hours. Each waveform was sampled at 100 Hz and and underwent mean removal. To reduce high-frequency noise while preserving physiologically relevant dynamics, signals were low-pass filtered at 10 Hz prior to segmentation. The dataset's diversity across patient demographics, clinical conditions, and hemodynamic states provides a strong basis for developing foundation models. Figure 2 presents the demographic distributions for the pretraining cohort.

Task-specific Dataset: We used two downstream datasets focused on detecting moderate to severe aortic stenosis (AS) and reduced left ventricular ejection fraction (LowEF, defined as LVEF < 40%).

Labels were derived from transthoracic echocardiography (TTE), the clinical gold standard. This retrospective study, conducted at the Cleveland Clinic from December 2020 to March 2022, included patients referred nationwide, ensuring broad geographic diversity. Data from December 2020 to October 2021 were used for model training, and data from October 2021 to March 2022 formed an independent test set. The AS dataset included 1,444 subjects (183 positive), and the LowEF dataset included 3,956 subjects (336 positive).

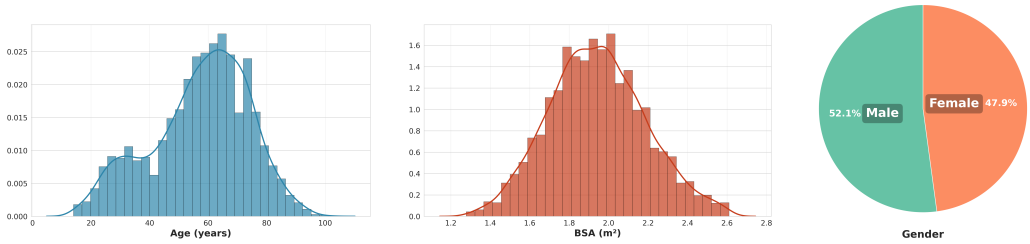

Figure 2: Demographic distributions of the pre-training dataset.

Table 1: Performance on age and biological sex prediction on the Cleveland Clinic dataset.

Model	**Age (MAE)**	**Biological Sex (AUROC)**
Task-specific CNN	8.049	0.866
Ours	**7.257**	**0.873**
MOMENT [7]	8.244	0.829
bioFAME [14]	8.060	0.855

3.2 Implementation Details

The Flan-T5-Large encoder serves as our foundation backbone, consisting of 24 transformer blocks with hidden size 1024, 16 attention heads, and feed-forward size 2816, using absolute sinusoidal positional embeddings. We pretrain with 50% token masking (patch size 8), Adam optimizer, batch size 250, and learning rate 1×10^{-4} for 10 epochs. For downstream tasks, frozen embeddings are pooled to 1024 features and passed to a two-layer MLP with hidden size 512, with the learning rate increasing from 3×10^{-5} to 1×10^{-3} in the first 10% of epochs and decaying over the final 30%.

3.3 Results and Discussion

We compare our model to task-specific CNN baselines—supervised models separately optimized for each clinical task—and state-of-the-art time-series foundation models, including MOMENT (pretrained on large-scale public data) and bioFAME (pretrained in the frequency domain on our dataset). To assess the general utility of the learned embeddings beyond disease classification, we first evaluate representation quality on two auxiliary tasks that reflect fundamental physiological attributes: age, biological sex, and cardiac output (CO) estimation. These tasks probe whether the foundation model encodes meaningful cardiovascular structure independent of specific diseases.

Tables 1 and 2 show that our model learns broadly meaningful cardiovascular representations across demographic and functional tasks. For age and sex prediction, the model achieves an age MAE of 7.257 and a biological sex AUROC of 0.873, indicating that the embeddings capture physiological aging signals and sex-associated hemodynamic patterns—attributes that strongly influence waveform morphology and are essential for unbiased downstream modeling. For continuous cardiac output estimation, the model achieves a correlation of 0.801 and a percent precision error of 26.6%, outperforming the supervised baseline and prior foundation models and approaching agreement levels typical of arterial-line estimators versus intermittent cardiac output using thermodilution. Importantly, this performance is obtained in a cohort of cardiac surgery and liver transplant patients, groups with highly abnormal and variable hemodynamics, yet the error profile remains comparable to reports on healthy or mildly ill populations, highlighting the robustness of the learned representations under challenging clinical conditions. This reflects the advantages of large-scale pretraining in capturing subtle cycle-level biomarkers tied to stroke volume, contractility, and arterial compliance. Together, these results confirm that the proposed foundation model internalizes general hemodynamic princi-

Table 2: Performance on cardiac output (CO) estimation from arterial-line waveforms in cardiac surgery and liver-transplant patients. CO values had a mean of 6.1 L/min (SD 3.4), median 4.9 L/min (IQR 4.0–6.8), and ranged from 2.0 to 19.5 L/min.

Model	**Bias**	**% Bias**	**Precision**	**% Precision**	**Correlation**
Task-specific CNN	-0.031	-5.593	2.197	28.992	0.769
Ours	**-0.274**	**-8.202**	**2.047**	**26.607**	**0.801**
MOMENT [7]	1.600	17.920	3.296	40.657	0.288
bioFAME [14]	-0.120	-6.334	2.140	28.265	0.780

Table 3: Performance on aortic stenosis (AS) and low ejection fraction (LowEF) detection from finger-cuff waveforms.

AS detection

Model	AUROC (wr.r.t training subject percentage)					Sensitivity	Specificity
	5%	10%	30%	50%	100%		
Task-specific	0.724 ± 0.065	0.788 ± 0.041	0.832 ± 0.029	0.857 ± 0.025	0.881	0.802	0.77
Ours	**0.760 ± 0.060**	**0.821 ± 0.032**	**0.878 ± 0.018**	**0.896 ± 0.008**	**0.918**	**0.802**	**0.836**
MOMENT [7]	0.740 ± 0.049	0.780 ± 0.017	0.780 ± 0.024	0.786 ± 0.016	0.802	0.802	0.654
bioFAME [14]	0.742 ± 0.043	0.769 ± 0.024	0.803 ± 0.022	0.814 ± 0.016	0.828	0.802	0.685

LowEF detection

Model	AUROC (w.r.t training subject percentage)					Sensitivity	Specificity
	5%	10%	30%	50%	100%		
Task-specific	0.747 ± 0.004	0.810 ± 0.010	0.875 ± 0.004	0.884 ± 0.006	0.895	0.806	0.818
Ours	**0.812 ± 0.018**	**0.860 ± 0.012**	**0.900 ± 0.006**	**0.903 ± 0.003**	**0.908**	**0.806**	**0.851**
MOMENT [7]	0.705 ± 0.033	0.774 ± 0.009	0.840 ± 0.008	0.851 ± 0.009	0.865	0.806	0.726
bioFAME [14]	0.761 ± 0.067	0.825 ± 0.011	0.866 ± 0.011	0.872 ± 0.003	0.880	0.806	0.801

ples rather than relying on task-specific cues, providing the basis for its strong downstream disease detection performance.

Table 3 shows that our model consistently outperforms both the task-specific CNN baselines and previous foundation models across all data regimes for both cardiovascular conditions using only noninvasive ClearSight finger-cuff waveforms. The benefit is most pronounced in low-data settings: even with only 5% or 10% of the training data, our model sustains higher AUROC scores, highlighting its strong data efficiency and generalizability under limited supervision. Notably, with just 30% of the training data, our model matches the performance of the task-specific CNN trained on the full 100%, underscoring significant improvements in training efficiency. At full data availability, our model achieves AUROC scores of 0.918 for AS and 0.908 for LowEF detection—exceeding the CNN baseline (0.881 and 0.895), MOMENT (0.802 and 0.865), and bioFAME (0.828 and 0.880). For fair comparison, all models are evaluated at the same high sensitivity level, prioritizing the detection of positive cases; under this setting, our approach achieves substantially higher specificity, reflecting improved discrimination of true negatives.

The strong performance of our model is driven by three key components. First, pressure-aware tokenization preserves physiologically meaningful features such as pulse pressure dynamics—particularly critical for AS detection—that are often lost in other foundation models relying on instance normalization [15]. Second, large-scale masked waveform pretraining equips the model to learn rich cardiovascular representations, enabling strong generalization even in low-data regimes. Finally, weight averaging during downstream training enhances robustness and supports hyperparameter transferability across diverse clinical tasks, resulting in consistent performance gains.

4 Conclusion

This work introduces a foundation model for hemodynamic waveforms, demonstrating that large-scale self-supervised learning on cardiovascular signals enables accurate and data-efficient screening for critical conditions. By moving beyond task-specific modeling, we show how waveform foundation models can unlock scalable, noninvasive, and cost-effective cardiovascular assessment at the point of care. Looking forward, this paradigm opens new opportunities for real-time monitoring, longitudinal risk stratification, and integration into broader multimodal healthcare systems, paving the way toward more accessible and proactive cardiovascular medicine.

References

[1] David N Ku. Blood flow in arteries. *Annual review of fluid mechanics*, 29(1):399–434, 1997.

[2] Timothy W Secomb. Hemodynamics. *Comprehensive physiology*, 6(2):975–1003, 2016.

[3] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

[4] Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

[5] Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

[6] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

[7] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[8] Chenguo Lin, Xumeng Wen, Wei Cao, Congrui Huang, Jiang Bian, Stephen Lin, and Zhirong Wu. Nutime: Numerically multi-scaled embedding for large-scale time-series pretraining. *arXiv preprint arXiv:2310.07402*, 2023.

[9] Blase A Carabello and Walter J Paulus. Aortic stenosis. *The lancet*, 373(9667):956–966, 2009.

[10] Ateet Kosaraju, Amandeep Goyal, Yulia Grigorova, and Amgad N Makaryus. Left ventricular ejection fraction. 2017.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[12] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[13] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

[14] Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

[15] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.