On-device Foundation Models for Wearable Signals

Simon A. Lee[*,1,2], Cyrus Tanade[1], Hao Zhou[1], Juhyeon Lee[1], Megha Thukral[1], Minji Han[1], Rachel Choi[1], Md Sazzad Hissain Khan[1], Baiying Lu[1] and Sharanya Arcot Desai [1]

[*]Work done during AI Residency, [1]Digital Health Team, Samsung Research America, [2]Department of Computational Medicine, University of California Los, Angeles

We propose a lightweight foundation model for wearable signals that leverages convolutional inductive biases within a masked autoencoder and U-Net CNN backbone. By explicitly encoding temporal locality and multi-scale structure, our approach aligns more naturally with the nonstationary dynamics of physiological waveforms than attention-based transformers. Pretrained on 80k hours of photoplethysmogram (PPG), the model matches or surpasses larger state-of-the-art baselines across ten clinical classification tasks. At the same time, it achieves two to three orders of magnitude reductions in parameters (0.31M vs. 110M), memory footprint (3.6MB vs. 441.3MB), and compute, while delivering substantial speedups ($\sim$4$\times$ CPU, $\sim$20$\times$ GPU) with resolution flexibility. Together, these results establish compact convolutional self-supervised models as both scientifically aligned and practically scalable for real-time on-device healthcare monitoring.

Keywords: On-device, SSL, Inductive Bias, Efficiency

Corresponding author(s): simonlee711@g.ucla.edu/s.lee5@partner.samsung.com

1. Introduction

Continuous physiological monitoring through wearable sensors has the potential to transform healthcare by enabling scalable, real-time assessment of cardiovascular, metabolic, and systemic states. Recent advances in wearable technology (from smartwatches to medical-grade devices) now generate vast streams of multimodal physiological data that are typically unlabeled, noisy, and high-frequency (Lee and Akamatsu, 2025). Extracting structure from these signals in a computationally feasible way remains a core challenge.

Foundation models (Khan et al., 2025; Zhou et al., 2024) provide a promising paradigm, learning general-purpose representations from large volumes of unlabeled data via self-supervised pretraining (Abbaspourazad et al., 2023; Lee and Akamatsu, 2025; Narayanswamy et al.). However, when applied to wearables, their deployment is hindered by enormous parameter counts and high memory requirements that make inference impractical on resource-constrained edge devices.

Transformer-based architectures (Narayanswamy et al., 2024; Vaswani et al., 2017) dominate current foundation model design, yet their use for physiological signals reveals limitations. Photoplethysmography (PPG), for instance, is highly nonstationary, with a morphology that can change by participant. Its waveforms combine quasi-periodic rhythms with subtle aperiodic variations from arrhythmias, vascular tone, and motion artifacts (Almarshad et al., 2022; Nitzan and Ovadia-Blechman, 2022). Capturing such dynamics requires sensitivity to both local temporal structure and longer-range dependencies. Transformers, lacking explicit inductive biases for locality, often force a trade-off: smaller models tend to underfit, while larger ones rely on brute-force capacity.

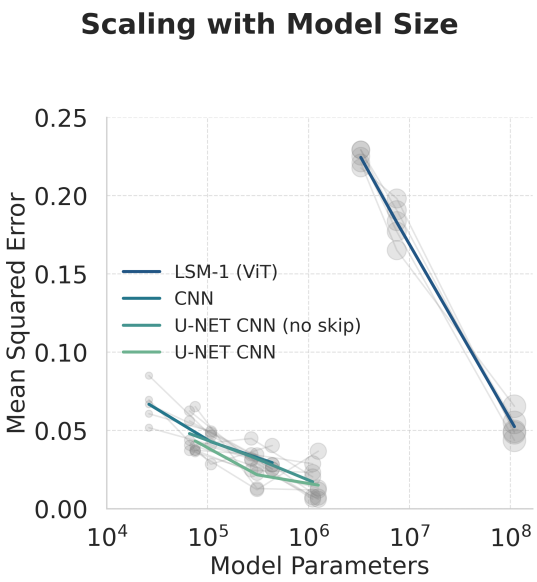

Figure 1 | Model Scaling: The U-NET CNN exhibits a graceful scaling behavior despite smaller capacity, outperforming the Transformer and standard CNN models, likely due to its use of hierarchical feature learning.

Convolutional networks (LeCun et al., 1989; O'Shea and Nash, 2015) offer a more natural alternative. They are parameter-efficient, inherently local, and scale linearly with sequence length. Moreover, U-Net–style hierarchies (Ronneberger et al., 2015) enable multi-resolution feature extraction that aligns with physiological waveforms and scale gracefully (Figure 1).

In this work, we introduce a lightweight masked autoencoding framework (He et al., 2022) built on a U-Net CNN backbone (Ronneberger et al., 2015). When pretrained on 80,000 hours of PPG, the model matches or surpasses state-of-the-art transformer and contrastive methods across ten clinically motivated tasks. At the same time, its convolutional hierarchy yields reductions of two to three orders of magnitude in parameters, FLOPs, and memory relative to representative transformer baselines, alongside substantial inference speedups. Taken together, these results suggest that inductive bias is as critical as scale: convolutional masked autoencoders provide efficient, resolution-flexible, and scientifically aligned foundations for wearable sensing, establishing a practical and principled path toward on-device foundation models.

2. Related Work

2.1. Self-Supervised Pretraining Objectives for Wearable Signals

Wearable Devices equipped with photoplethysmography (PPG), electrocardiography (ECG), and accelerometry generate long, multi-channel time series that encode diverse physiological and behavioral phenomena, including cardiovascular dynamics (Castaneda et al., 2018), activity patterns (Xu et al., 2025; Yuan et al., 2024), sleep cycles (Li et al., 2021; Logacjov et al., 2025; Thapa et al., 2024), and other latent processes. These data streams are abundant, passively collected, and predominantly unlabeled, making them ideal candidates for large-scale self-supervised learning (SSL) (Bommasani et al., 2021; Liang et al., 2024; Zhou et al., 2024; ?).

Self-supervised learning (SSL) has become the dominant paradigm for wearable time-series representation learning, reflecting both the scarcity of labeled data and the abundance of unlabeled sequences collected in free-living settings. Among SSL strategies, masked autoencoding has risen to prominence, inspired by its success in vision (He et al., 2022; Vaid et al., 2023) and language modeling (Devlin et al., 2019). In this framework, random patches of the signal are occluded, and the model is tasked with reconstructing the missing regions. This simple but powerful objective forces the network to capture latent physiological structure and temporal regularities (Kong et al., 2023; Zhang et al., 2022). Recent large-scale efforts, including Google's LSM series (Narayanswamy et al., 2024; Xu et al., 2025), rely heavily on masked autoencoding, establishing it as a foundation for pretraining on multi-modal wearable datasets. Despite its effectiveness for local pattern recovery, masked autoencoding by itself often struggles to capture long-range temporal dependencies unless explicitly paired with architectures designed for hierarchical structure.

A complementary line of work is contrastive learning, which encourages invariance by drawing semantically similar samples closer in latent space while pushing dissimilar samples apart (Jaiswal et al., 2020; Schmitt and Kuljanin, 2008). For wearable signals, the main challenge lies in defining positive and negative pairs without explicit labels. A proposed solution is participant-level contrastive training, where segments from the same individual are positives and those from different individuals are negatives—an approach adopted in Apple's ECG and PPG foundation models (Abbaspourazad et al., 2023), echoing the SimCLR framework (Chen et al., 2020b). More domain-specific innovations attempt to design physiologically meaningful augmentations: PaPaGei leverages PPG morphology to construct contrastive pairs (Pillai et al., 2024), while SleepFM extends the paradigm across multiple modalities (EEG, ECG, EMG) to enforce cross-modal consistency (Thapa et al., 2024). Additional embedding-level regularizers, such as differential entropy constraints (Abbaspourazad et al., 2023; Jing et al., 2021), further enrich representation quality. Nevertheless, contrastive methods are often sensitive to augmentation heuristics, computationally intensive, and limited in interpretability, providing little insight into which temporal structures are preserved.

3. Methods

3.1. Pre-training Data

Our pre-training corpus comprises a collection of approximately $80,000$ hours of wearable photoplethysmography (PPG) signals. This data was aggregated from seven independent studies collected internally across Samsung, encompassing $47,644$ participants and collected across seven distinct Samsung Galaxy Watch devices. Following preprocessing (Signal Quality Index (SQI) check, bandpass filters, and z-score normalization rescaling), the continuous signals are segmented into fixed-length intervals for masked autoencoding. We use 10-second windows

sampled at 100 Hz, resulting in sequences of $1,000$ timesteps each. These time intervals were selected in accordance to clinical practice where clinically collected waveforms are collected in 10 second intervals (e.g., Electrocardiograms).

3.2. Masked Autoencoding with a U-Net CNN

We extend the masked autoencoding paradigm (He et al., 2022) to single-channel 1-D physiological time series by pairing a patch-masking self-supervised objective with a U-Net–style convolutional encoder–decoder (Ronneberger et al., 2015). The model takes an input sequence $x \in \mathbb{R}^L$, partitions it into $N = L/P$ non-overlapping patches of length P, and applies a binary mask $m \in \{0,1\}^N$ sampled i.i.d. from a Bernoulli distribution with parameter r, the masking ratio. We adopt random masking, identical to the strategy used in the Large Signal Model (LSM) (Narayanswamy et al., 2024), where masked patch indices are drawn uniformly without replacement from the N available positions. The mask is expanded to match the temporal resolution, yielding $m' \in \{0,1\}^L$. The observed signal is $\tilde{x} = x \odot (1 - m')$, where $\odot$ denotes element-wise multiplication. This process removes large contiguous regions of the input, forcing the model to infer missing dynamics from incomplete temporal context.

The encoder f_θ is a hierarchical convolutional network composed of residual blocks and strided convolutions, halving the temporal resolution at each stage. This progressive downsampling expands the receptive field exponentially (Table 1); in our configuration, the bottleneck spans approximately ($\approx 1000/2^5$) input timesteps, while earlier layers capture fine-scale fluctuations. Skip connections link encoder and decoder stages, preserving high-frequency detail alongside coarse abstractions.

Table 1 | Temporal resolution and cumulative receptive field through the encoder. T denotes the input length in samples. R_ℓ is the receptive field after layer ℓ and J_ℓ the effective input stride ("jump").

Layer	Kernel k	Stride s	Output length	R_ℓ / J_ℓ
Enc1-conv1	5	2	$T/2$	5 / 2
Enc1-conv2	5	1	$T/2$	13 / 2
Enc2-conv1	5	2	$T/4$	21 / 4
Enc2-conv2	5	1	$T/4$	37 / 4
Enc3-conv1	5	2	$T/8$	53 / 8
Enc3-conv2	5	1	$T/8$	85 / 8
Enc4-conv1	5	2	$T/16$	117 / 16
Enc4-conv2	5	1	$T/16$	181 / 16
Enc5-conv1	5	2	$T/32$	245 / 32
Enc5-conv2	5	1	$T/32$	373 / 32

The decoder g_ϕ mirrors the encoder, employing transposed convolutions to upsample back to the original resolution. When skip connections are enabled, upsampled features are concatenated with encoder activations and refined via convolution, injecting localized detail without discarding global temporal context. A final transposed convolution with bounded nonlinearity outputs $\hat{x} \in \mathbb{R}^L$, matching the amplitude range of the input signal.

Training minimizes a masked reconstruction loss $\mathcal{L}_{\text{MSE}}(\theta, \phi) = \frac{\|(\hat{x}-x)\odot m'\|_2^2}{\sum_{t=1}^{L} m'_t}$ which computes mean squared error only over masked positions. This formulation forces the model to estimate $p(x_{\mathcal{M}} \mid x_{\mathcal{O}})$, where $\mathcal{M}$ and $\mathcal{O}$ denote masked and observed indices, discouraging trivial copying of visible segments and encouraging temporally coherent, multi-scale representations.

We optimize with AdamW (?) and a warmup–cosine learning rate schedule. Masking patterns are resampled independently for every sequence and iteration, improving diversity and reducing overfitting to specific occlusions. All benchmarks use subject-disjoint train/validation/test splits to prevent identity leakage, and no subjects overlap between pretraining and evaluation, ensuring that benchmark performance reflects generalization. All hyperparameters are reported in our Appendix Section C.2.

Model	Params (↓)	FLOPs (↓)	Memory (↓)
U-NET CNN	1.2M	0.0647 gFLOPS	4.8 MB
Efficient-Net	7.8M	0.70 gFLOPS	31.1 MB
Swin-Transformer	110.6M	11.89 gFLOPS	423.8 MB
LSM-Base	110.6M	15.94 gFLOPS	441.3 MB

Model	GPU Lat. (↓)	GPU Thr. (↑)	CPU Lat. (↓)	CPU Thr. (↑)
U-NET CNN	0.039 ms	25.8k/s	0.99 ms	1.2k/s
Efficient-Net	0.082 ms	12.2k/s	1.42 ms	0.704k/s
Swin-Transformer	0.704 ms	1.42k/s	2.95 ms	0.456k/s
LSM-Base	0.80 ms	1.24k/s	3.36 ms	0.298k/s

Table 2 | Model efficiency and on-device inference: Sample on-device detections on Samsung Watch 8 device. Size, compute cost, memory footprint, and CPU latency (ms per sample, batch size 2048) measured over a 10s sequence at 100Hz. Latency (ms per sample, batch size 2048) and throughput (samples/sec) measured over 10 s windows.

3.3. Experimental Design

Inference Efficiency Protocol

To assess the feasibility of real-time digital health monitoring on resource-constrained devices, we evaluated the U-Net CNN's inference performance against the transformer baseline (LSM-Base & Swin Transformer) as well as Efficient-Net. Efficiency was measured across three dimensions: (i) model footprint and gFLOPs per 10-second window at 100 Hz (Table 2); (ii) latency, defined as mean per-sample forward-pass time (ms) at batch size 1; and (iii) throughput, the maximum number of samples processed per second (Table 2).

All experiments were run on a Samsung Watch Series 8. Benchmarks were run on-device, using Exynos W1000 CPUs. We also tested on a T4 GPU for potential mobile device deployment; although the T4 is a datacenter GPU, modern mobile processors like the Qualcomm Adreno 750 found on commercial phones are optimized for high-performance ML and can deliver comparable efficiency (Buber and Banu, 2018; Wesolowski et al., 2021), underscoring the practicality of on-device deployment.

Classification Performance

We additionally evaluate binary classification across ten clinically motivated tasks: premature ventricular contractions (PVC) detection, hypertension status, and eight laboratory abnormality screens (Potassium, Sodium, Platelets, A1C, Creatinine, Hemoglobin, LDL, CO_2). These tasks span acute events (PVC), chronic conditions (hypertension), and systemic dysregulation (biochemical markers), thereby probing both transient and long-term predictive capacity from PPG signals.

PVC labels were derived from expert annotations on our internal Samsung datasets. Hypertension was defined under specialist guidance as systolic blood pressure ≥ 130,mmHg, with labels aggregated from two independent cohorts in a free world and lab setting. For the remaining eight tasks, we used Tulane University institution data linking continuous wearable signals with temporally aligned laboratory results. Each laboratory task was cast as an out-of-range screen: positives correspond to values exceeding institution-defined clinical thresholds, negatives otherwise.

We compare our proposed U-Net CNN model against three strong baselines: (i) a SimCLR variant, (ii) PaPaGei (Pillai et al., 2024), an open-source PPG foundation model, and (iii) the Large Sensor Model (LSM) (Narayanswamy et al., 2024), a transformer-based masked autoencoder.

Model	PVC	Hypertension	Potassium	Sodium	Platelets	A1C	Creatinine	Hemoglobin	LDL	CO_2
UNET-CNN	0.802	0.663	0.837	0.614	0.685	0.695	0.552	0.547	0.579	0.457
LSM	0.722	0.646	0.698	0.485	0.541	0.793	0.447	0.586	0.572	0.489
SimCLR	0.715	0.609	0.664	0.492	0.587	0.623	0.531	0.583	0.497	0.401
PaPaGei	0.802	0.646	0.811	0.607	0.770	0.754	0.530	0.585	0.648	0.461

Table 3 | Downstream Classification Performance: AUROC on ten classification tasks. Best performing column are underlined.

Together, these cover contrastive, transformer and PPG-specific FM paradigms for representation learning. Performance is all measured via linear probing reported using AUROC as the primary metric, which is robust to class imbalance across tasks (McDermott et al., 2024).

4. Results

4.1. Classification Results

Table 3 reports AUROC across ten binary tasks. Our U-NET consistently secures the majority of wins, frequently outperforming or matching models in this benchmark. While PaPaGei and LSM achieve isolated wins, both rely on substantially larger or more specialized architectures. In contrast, our model achieves comparable or superior performance with a model footprint two to three orders of magnitude smaller. Taken together, these results demonstrate that convolutional inductive biases, when paired with masked autoencoding, can rival or surpass transformer-based approaches in both efficiency and predictive accuracy.

4.2. Few Shot Learning

A central challenge in the wearable domain is that labels are scarce across tasks. Models that can adapt quickly from generic pretraining to specific detection tasks with limited supervision are therefore essential. Figure 2 illustrates this setting: U-NET CNN provides strong representations that can be adapted to diverse tasks such as PVC detection or hypertension monitoring with only a handful of labeled examples as reflected by the shape of the learning curves on the few-shot learning experiments. By reducing the supervision required to reach high performance, U-NET CNN enables new tasks to be supported on-device without the prohibitive cost of large curated datasets which help bolster its practical utility.

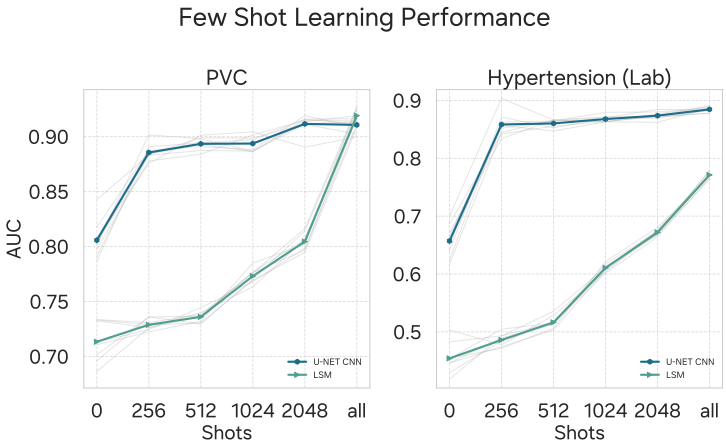

Figure 2 | Few-shot adaptation. U-NET CNN adapts efficiently to new wearable tasks under sparse labels indicated by curve shape over transformer baselines.

4.3. On-Device Benchmarking

A central novelty of U-NET CNN is that it is, to our knowledge, the first SSL method compact enough to run entirely on-watch, rather than on phone-class hardware. We evaluate on-device PVC detection on smartwatch-class CPUs sampled at 100 Hz (Figure 3). U-NET CNN is exceptionally lightweight (1.2M parameters, 0.0647 gFLOPs, 4.8 MB) and achieves 0.99 ms latency per sample, equivalent to processing ≈1,010 samples/s or ≈2.8 hours of signal per minute of wall time. By contrast it shows massive performance gains against transformer baselines, Swin-Transformer (110M parameters, 11.9 gFLOPs, 423 MB) and LSM-Base (110M, 15.9 gFLOPs, 441 MB). U-NET CNN also outperforms optimized models like Efficient-Net B1 (Tan and Le, 2020) providing context to the latency and compactness of our model. U-NET CNN is thus ∼3–4× more efficient compared to transformers while fitting fully on-watch (without quantization (Jacob et al., 2017)), enabling continuous, private inference at the point of signal collection. This prototype is strictly for research and is not deployed commercially.

Figure 3 | Model efficiency and on-device inference: Sample on-device detections on Samsung Watch 8 device. Size, compute cost, memory footprint, and CPU latency (ms per sample, batch size 2048) measured over a 10s sequence at 100Hz.

Model footprint. Our U-Net CNN is two to three orders of magnitude lighter than the transformer baseline while maintaining accuracy (Table 2). Parameters drop from 110M to 0.31M ($\sim 355\times$ fewer), FLOPs per 10s window at 100Hz fall from 15.94G to 0.0647G ($\sim 246\times$ fewer), and memory shrinks from 441.3MB to 3.6MB ($\sim 123\times$ smaller). This compactness enables storage within mobile or embedded caches and allows multiple task heads to co-reside on a single device, while also supporting dense multiplexing in server deployments.

Latency. On GPU, mean per-sample latency improves from 0.80,ms to 0.039,ms ($\sim 20\times$ faster), and on CPU from 3.93,ms to 0.99,ms ($\sim 4\times$ faster) (Table ??). At 100 Hz, these latencies leave ample headroom for real-time streaming. Although our benchmarks used NVIDIA Tesla T4 GPUs, modern smartphone chipsets—such as those equipped with the Qualcomm Adreno 750 GPU—are optimized for high-performance ML, and thus are likely to deliver comparable efficiency in practice.

Throughput. The model also sustains markedly higher throughput. On GPU, throughput rises from 1.24k to 25.8k samples/s ($\sim 21\times$ higher), equivalent to compressing ~ 71 hours of continuous 100 Hz data per real-time second versus ~ 3.4 hours/s for LSM. On CPU, throughput improves from 0.255k to 1.2k samples/s ($\sim 5\times$ higher), corresponding to ~ 3.3 hours of signal per second. Put differently, a single GPU could process an entire week of PPG data in under two minutes, while CPU inference remains feasible for continuous, always-on monitoring.

Taken together, these results highlight that our model delivers substantial speedups over transformer baselines at a fraction of the cost, combining a 3.6 MB footprint with sub-millisecond latency and orders-of-magnitude higher throughput. This efficiency enables future research on-deployment for mobile and wearable devices while reducing energy consumption and thermal load during inference.

5. Discussion

Summary. We introduce a single–channel U-Net CNN masked autoencoder for wearable PPG that, when pretrained on 80k hours of data, achieves competitive or superior AUROC across ten clinically relevant tasks. Crucially, our model is two to three orders of magnitude smaller and faster than transformer baselines, translating into sub-millisecond latency and high throughput on both GPU and CPU. This efficiency makes real-time, on-device inference practical without sacrificing accuracy.

Acknowledgments

We extend our gratitude to the leadership of Samsung Research America in Digital Health: Subramaniam Venkatraman, Matthew Wiggins, and Praveen Raja, for their invaluable feedback and guidance throughout this project. Additionally, we express our appreciation to the numerous researchers, designers, and developers with whom we engaged in discussions regarding this work, which contributed significantly to its successful completion.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: a system for large-scale machine learning. In Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16, page 265–283, USA, 2016. USENIX Association. ISBN 9781931971331.

S. Abbaspourazad, O. Elachqar, A. C. Miller, S. Emrani, U. Nallasamy, and I. Shapiro. Large-scale training of foundation models for wearable biosignals. arXiv preprint arXiv:2312.05409, 2023.

M. A. Almarshad, M. S. Islam, S. Al-Ahmadi, and A. S. BaHammam. Diagnostic features and potential applications of ppg signal in healthcare: A systematic review. In Healthcare, volume 10, page 547. MDPI, 2022.

U. An, M. Jeong, S. A. Lee, A. Gorla, Y. Yang, and S. Sankararaman. Raptor: Scalable train-free embeddings for 3d medical volumes leveraging pretrained 2d foundation models. arXiv preprint arXiv:2507.08254, 2025.

B. Arnrich, E. Choi, J. A. Fries, M. B. McDermott, J. Oh, T. Pollard, N. Shah, E. Steinberg, M. Wornow, and R. van de Water. Medical event data standard (meds): Facilitating machine learning for health. In ICLR 2024 Workshop on Learning from Time Series For Health, pages 03--08, 2024.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

E. Buber and D. Banu. Performance analysis and cpu vs gpu comparison for deep learning. In 2018 6th International Conference on Control Engineering & Information Technology (CEIT), pages 1--6. IEEE, 2018.

D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. International journal of biosensors & bioelectronics, 4(4):195, 2018.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597--1607. PmLR, 2020a.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020b. URL https://arxiv.org/abs/2002.05709.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171--4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

A. Hannun, J. Digani, A. Katharopoulos, and R. Collobert. MLX: Efficient and flexible machine learning on apple silicon, 2023. URL https://github.com/ml-explore.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000--16009, 2022.

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL https://arxiv.org/abs/1712.05877.

A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. Technologies, 9(1):2, 2020.

L. Jing, P. Vincent, Y. LeCun, and Y. Tian. Understanding dimensional collapse in contrastive self-supervised learning. arXiv, 2021. doi: 10.48550/arxiv.2110.09348. URL https://arxiv.org/abs/2110.09348.

W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang. A comprehensive survey of foundation models in medicine. IEEE Reviews in Biomedical Engineering, 2025.

L. Kong, M. Q. Ma, G. Chen, E. P. Xing, Y. Chi, L.-P. Morency, and K. Zhang. Understanding masked autoencoders via hierarchical latent variable models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7918--7928, 2023.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541--551, 1989.

S. A. Lee and K. Akamatsu. Foundation models for physiological signals: Opportunities and challenges. 2025.

S. A. Lee and T. Lindsey. Can large language models abstract medical coded language? arXiv preprint arXiv:2403.10822, 2024.

S. A. Lee, C. Tanade, H. Zhou, J. Lee, M. Thukral, B. Lu, and S. A. Desai. Towards on-device foundation models for raw wearable signals. In NeurIPS 2025 Workshop on Learning from Time Series for Health.

S. A. Lee, J. Lee, and J. N. Chiang. Feet: A framework for evaluating embedding techniques. arXiv preprint arXiv:2411.01322, 2024.

S. A. Lee, S. Jain, A. Chen, K. Ono, A. Biswas, Á. Rudas, J. Fang, and J. N. Chiang. Clinical decision support using pseudo-notes from multiple streams of ehr data. npj Digital Medicine, 8(1):394, 2025.

Q. Li, Q. Li, A. S. Cakmak, G. Da Poian, D. L. Bliwise, V. Vaccarino, A. J. Shah, and G. D. Clifford. Transfer learning from ecg to ppg for improved sleep staging from wrist-worn wearables. Physiological measurement, 42(4):044004, 2021.

Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, S. Pan, and Q. Wen. Foundation models for time series analysis: A tutorial and survey. In Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, pages 6555--6565, 2024.

Y. Lin, Z. B. Yu, and S. Lee. A case study exploring the current landscape of synthetic medical record generation with commercial llms. arXiv preprint arXiv:2504.14657, 2025.

A. Logacjov, K. Bach, and P. J. Mork. Long-term self-supervised learning for accelerometer-based sleep--wake recognition. Engineering Applications of Artificial Intelligence, 141:109758, 2025.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

M. McDermott, H. Zhang, L. Hansen, G. Angelotti, and J. Gallifant. A closer look at auroc and auprc under class imbalance. Advances in Neural Information Processing Systems, 37:44102--44163, 2024.

M. B. McDermott, J. Xu, T. S. Bergamaschi, H. Jeong, S. A. Lee, N. Oufattole, P. Rockenschaub, K. Stankevičiūtė, E. Steinberg, J. Sun, et al. Meds: Building models and tools in a reproducible health ai ecosystem. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pages 6243--6244, 2025.

G. Narayanswamy, X. Liu, K. Ayush, Y. Yang, X. Xu, J. Garrison, S. A. Tailor, J. Sunshine, Y. Liu, T. Althoff, et al. Scaling wearable foundation models. In The Thirteenth International Conference on Learning Representations.

G. Narayanswamy, X. Liu, K. Ayush, Y. Yang, X. Xu, S. Liao, J. Garrison, S. Tailor, J. Sunshine, Y. Liu, et al. Scaling wearable foundation models. arXiv preprint arXiv:2410.13638, 2024.

M. Nitzan and Z. Ovadia-Blechman. Physical and physiological interpretations of the ppg signal. In Photoplethysmography, pages 319--340. Elsevier, 2022.

K. Ono and S. A. Lee. Text serialization and their relationship with the conventional paradigms of tabular machine learning. arXiv preprint arXiv:2406.13846, 2024.

K. O'Shea and R. Nash. An introduction to convolutional neural networks, 2015. URL https://arxiv.org/abs/1511.08458.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

A. Pillai, D. Spathis, F. Kawsar, and M. Malekzadeh. Papagei: Open foundation models for optical physiological signals. arXiv preprint arXiv:2410.20542, 2024.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234--241. Springer, 2015.

N. Schmitt and G. Kuljanin. Measurement invariance: Review of practice and implications. Human resource management review, 18(4):210--222, 2008.

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

R. Thapa, B. He, M. R. Kjaer, H. M. Iv, G. Ganjoo, E. Mignot, and J. Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. In International Conference on Machine Learning, pages 48019--48037. PMLR, 2024.

M. Thukral, C. Tanade, S. A. Lee, J. Lee, and S. A. Desai. Wavelet-based masked multiscale reconstruction for ppg foundation models. In NeurIPS 2025 Workshop on Learning from Time Series for Health.

A. Vaid, J. Jiang, A. Sawant, S. Lerakis, E. Argulian, Y. Ahuja, J. Lampert, A. Charney, H. Greenspan, J. Narula, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. NPJ Digital Medicine, 6(1):108, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

L. Wesolowski, B. Acun, V. Andrei, A. Aziz, G. Dankel, C. Gregg, X. Meng, C. Meurillon, D. Sheahan, L. Tian, et al. Datacenter-scale analysis and optimization of gpu machine learning workloads. IEEE Micro, 41(5):101--112, 2021.

M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah. The shaky foundations of large language models and foundation models for electronic health records. npj digital medicine, 6(1):135, 2023.

M. A. Xu, G. Narayanswamy, K. Ayush, D. Spathis, S. Liao, S. A. Tailor, A. Metwally, A. A. Heydari, Y. Zhang, J. Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. arXiv preprint arXiv:2506.05321, 2025.

H. Yuan, S. Chan, A. P. Creagh, C. Tong, A. Acquah, D. A. Clifton, and A. Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. NPJ digital medicine, 7(1):91, 2024.

Q. Zhang, Y. Wang, and Y. Wang. How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems, 35:27127--27139, 2022.

C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. International Journal of Machine Learning and Cybernetics, pages 1--65, 2024.

A. Author Contribution

We attribute proper credit to the following authors for their contributions in this project.

Table 4 | Overview of author contributions.

Author	Concept	Experiment Design	Coding	Analysis	Writing	Visualization	Project Mgmt.	Discussion	Resources
Simon A. Lee	✓	✓	✓	✓	✓	✓	✓	✓	
Cyrus Tanade			✓	✓	✓		✓	✓	✓
Hao Zhou			✓	✓		✓		✓	
Juhyeon Lee				✓	✓			✓	✓
Megha Thurkal				✓				✓	✓
Minji Han						✓		✓	✓
Rachel Choi						✓		✓	✓
Md Sazzad Hissain Khan			✓	✓				✓	✓
Baiying Liu				✓				✓	
Sharanya Desai		✓			✓		✓	✓	✓

B. Reproducibility Statement

Table 5 | U-NET CNN architecture components.

Encoder--Decoder

Layer	Output Shape
Input	[B, 1, T]
EncoderConvBlock(1→16)	[B, 16, T/2]
EncoderConvBlock(16→32)	[B, 32, T/4]
EncoderConvBlock(32→64)	[B, 64, T/8]
EncoderConvBlock(64→128)	[B, 128, T/16]
EncoderConvBlock(128→256)	[B, 256, T/32]
DecoderSkipBlock(256→128)	[B, 128, T/16]
DecoderSkipBlock(128→64)	[B, 64, T/8]
DecoderSkipBlock(64→32)	[B, 32, T/4]
DecoderSkipBlock(32→16)	[B, 16, T/2]
Final Deconv (16→1)	[B, 1, T]
Tanh	[B, 1, T]

EncoderConvBlock

Layer
Conv1d ($k = 5$, s=2, p=2)
BatchNorm
GELU
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
Conv1d ($k = 1$, s=2) + BN
GELU

DecoderSkipBlock

Layer
ConvTranspose1d ($k = 5$, s=2, p=2, op=1)
Concat skip connection
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
GELU
Conv1d ($k = 5$, s=1, p=2)
BatchNorm
GELU

Due to restrictions around data licensing and industry policies, we are unable to release the full source code associated with U-NET CNN. To mitigate this limitation, we provide complete details of the model architecture, layer configurations, and hyperparameters in Table 5. This includes all encoder, decoder, and skip connection blocks, along with kernel sizes, strides, padding, activation functions, and normalization layers. Together, these descriptions are sufficient to re-implement the model faithfully in any modern deep learning framework (Abadi et al., 2016; Bradbury et al., 2018; Hannun et al., 2023; Paszke et al., 2019). In addition, we report all training settings (e.g., optimizer, learning rate schedule, and batch size) in the Appendix Section C to further support reproducibility. Our goal is to ensure that, while the exact implementation cannot be shared, independent researchers can replicate the methodology and validate the findings presented in this work given the state of AI in Health Research (Arnrich et al., 2024; McDermott et al., 2025).

C. Baselines and Model Configuration

Self Supervised methods have become a dominant paradigm for health to study a variety of applications (An et al., 2025; Lee and Lindsey, 2024; Lee et al., 2024, 2025; Lin et al., 2025; Ono and Lee, 2024; Thukral et al.; Wornow et al., 2023). Foundation models for one-dimensional signals are predominantly repurposed from architectures designed for vision, with adaptations that reinterpret temporal structure as a flattened analogue of spatial correlation. In this section we describe our baseline models and configurations

C.1. Baselines

LSM (Narayanswamy et al., 2024) introduces a large-scale foundation model trained on multimodal wearable sensor data. The approach adopts a vision transformer architecture trained via masked autoencoding with random masking. The model is designed as a general-purpose foundation, transferring effectively across a range of downstream tasks in physiological sensing and human activity recognition. In our work, we do not replicate the full multimodal design; instead, we adapt and constrain the model to a unimodal setting.

SimCLR (Chen et al., 2020b) establishes contrastive learning as a competitive self-supervised paradigm. The core idea is to maximize agreement between augmented views of the same signal in a latent space while pushing apart representations of different images. This is implemented using a ResNET encoder (He et al., 2015), a projection head, and a contrastive loss (NT-Xent (Chen et al., 2020a)).

PaPaGei (Pillai et al., 2024) is a domain-specific foundation model designed for optical physiological sensing, particularly photoplethysmography (PPG). It adapts ResNET-style CNN architectures to learn robust, generalizable representations from large-scale optical physiological datasets. PaPaGei releases both model weights and datasets to support reproducibility and broader adoption in physiological signal analysis. In our work, we used their source code to benchmark their method by pre-training on our volume of data to ensure fair comparison.

C.2. Model Hyperparameters

Table 6 | Hyperparameter Configurations for Different Models

Configuration	U-Net CNN	LSM	SimCLR	PaPaGei
Training Steps	50000			15000
Warmup Steps	2500			---
Optimizer	AdamW (Loshchilov and Hutter (2017))			
Opt. momentum $[\beta_1, \beta_2]$	[0.9, 0.95]	[0.9, 0.95]	[0.9, 0.99]	---
Base learning rate	1e-3	5e-3	1e-3	1e-4
Batch size	2048			256
Weight decay	1e-4			---
Gradient clipping	1.0	1.0	3.0	---
Dropout	0.0			---
Learning rate schedule	Linear Warmup & Cosine Decay			---
Loss Function	Mean Squared Error		Contrastive Loss	
Data resolution	1 (signal) - 100 Hz (Sampling rate) $\times$ 10 (seconds)			
Augmentation	Flip, Time Warping, Noise			

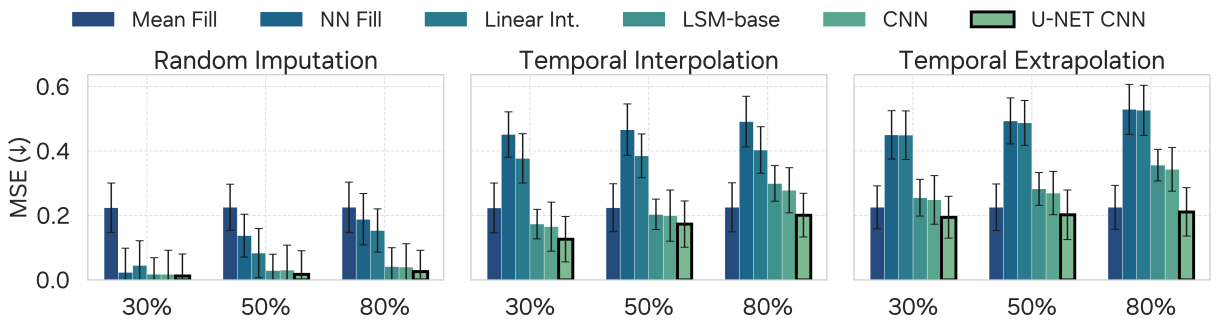

Figure 4 | Performance on generative benchmarks. Mean squared error for random imputation, temporal interpolation, and temporal extrapolation at varying missingness levels. Bold outline indicates best performing model.

D. Additional Results

D.1. Generative Performance

We conduct a generative benchmarks, where our U-NET CNN consistently outperforms all baselines across random imputation, temporal interpolation, and temporal extrapolation tasks (Figure 4). In terms of mean squared error, U-NET CNN achieves the lowest reconstruction error in every setting, including cases with heavy missingness. By achieving the lowest reconstruction error even in challenging extrapolation scenarios, U-NET CNN demonstrates reconstruction ability beyond naive heuristics (e.g., mean fill, nearest neighbor, or linear interpolation). Together, these results establish U-NET CNN as a strong generative model for missing data problems, with advantages that persist across scaling regimes and input corruption patterns.

D.2. Model Configurations Ablations

We conducted a comprehensive ablation study of U-NET CNN on a 100 Hz dataset comprising ten million segments (roughly 30k hours). The experiments systematically varied architecture and hyperparameters to understand their effect on reconstruction quality (Extrapolation task from our generative benchmark in tables where it is not explicitly stated as previously done in (Narayanswamy et al., 2024)), with multiple independent training runs averaged to reduce variance from stochastic initialization and data sampling. Unless otherwise noted, all training employed AdamW with a learning rate of 3×10^{-4}, cosine decay scheduling, and a batch size of 512.

Architecture.

We evaluated U-NET CNN alongside CNN baselines across increasing network depths, defined by the sequence of hidden channel dimensions $[16, 32, 64]$, $[16, 32, 64, 128]$, and $[16, 32, 64, 128, 256]$. Table 7 lists the parameter counts, showing a modest growth for U-NET CNN compared to CNN baselines, with the skip-connected U-NET CNN exhibiting slightly higher capacity than its no-skip variant.

Table 7 | Model Parameters (in K or M)

Model Depth	U-NET CNN-tiny [16,32,64]	U-NET CNN-small [16,32,64,128]	U-NET CNN-Base [16,32,64,128,256]
CNN	26.2 K	108 K	437 K
U-NET CNN-no skip	66.1 K	271 K	1.10 M
U-NET CNN	75.3 K	309 K	1.25 M

The impact of network depth on mean absolute error (MAE) and mean squared error (MSE) is summarized in Table 8. Increasing depth consistently reduced both MAE and MSE for U-NET CNN, with the deepest configuration yielding the lowest reconstruction error. Skip connections were critical, as U-NET CNN consistently outperformed its no-skip variant across all depths.

Table 8 | MAE and MSE for Different Network Depths

Model Depth	U-NET CNN-tiny [16,32,64]		U-NET CNN-small [16,32,64,128]		U-NET CNN-Base [16,32,64,128,256]	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	0.4052	0.2345	0.4177	0.2491	0.4008	0.2315
U-NET CNN-noskip	0.4031	0.2365	0.4006	0.2465	0.3975	0.2339
U-NET CNN	0.4008	0.2309	0.3892	0.2232	0.3827	0.2210

Patch Size.

We varied the spatial-temporal patch sizes over 1, 5, 10, and 20. The results in Table 9 indicate that 5 provided the best trade-off between local resolution and generative performance. Smaller patches increased flexibility but slightly degraded performance due to reduced receptive field per token, while overly large patches caused loss of fine-grained structure.

Convolution Kernel Size.

Kernel size was varied over $\{1, 5, 10, 20\}$. Table 10 shows that 5 yielded the lowest errors across all models, suggesting moderate receptive fields match the temporal and spatial scales of our

Table 9 | Model Performance for Different Patch Sizes

Model	1		5		10		20	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	0.4140	0.2391	0.4008	0.2315	0.4122	0.2449	0.4274	0.2613
U-NET CNN-noskip	0.4069	0.2398	0.3976	0.2339	0.4037	0.2462	0.4195	0.2629
U-NET CNN	0.3899	0.2268	0.3827	0.2210	0.3861	0.2312	0.4039	0.2479

data. Very small kernels restricted context aggregation, while very large kernels oversmoothed latent features.

Table 10 | Model Performance Across Convolution Kernel Sizes

Model	1		5		10		20	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	0.4162	0.2413	0.4010	0.2309	0.4103	0.2418	0.4241	0.2576
U-NET CNN-noskip	0.4090	0.2427	0.3959	0.2331	0.4032	0.2440	0.4208	0.2591
U-NET CNN	0.3921	0.2283	0.3821	0.2206	0.3885	0.2316	0.4047	0.2485

Stride.

We evaluated stride values of 2, 4, and 8 (Table 11). Smaller strides yielded the best performance, particularly for U-NET CNN, by preserving high temporal resolution in early feature maps. Performance degraded monotonically with stride increases.

Table 11 | Model Performance Across Stride Values

Model	2		4		8	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
CNN	0.4016	0.2312	0.4139	0.2445	0.4318	0.2678
U-NET CNN-noskip	0.3976	0.2334	0.4098	0.2471	0.4272	0.2702
U-NET CNN	0.3829	0.2209	0.3928	0.2325	0.4103	0.2504

Masking Ratio.

Finally, we explored the effect of varying the latent masking ratio in the masked autoencoding objective for generative tasks, with ratios from 0.5 to 0.9. As shown in Table 12, interpolation and extrapolation both improved when increasing the ratio up to 0.8, after which performance degraded for interpolation and collapsed for extrapolation.

Final Selection.

These controlled experiments informed the final U-NET CNN configuration: the deepest architecture $[16, 32, 64, 128, 256]$ with skip connections, patch size 5, kernel size 5, stride 2, and a masking ratio of 0.8, which jointly achieved the best trade-off between reconstruction fidelity and parameter efficiency.

Table 12 | MAE and MSE for U-NET CNN Across Different Masking Ratios Evaluated on Generative Tasks

U-NET CNN Masking Ratio	Temporal Interpolation		Temporal Extrapolation	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓
0.5	0.3972	0.2292	0.4077	0.2519
0.6	0.3889	0.2223	0.3975	0.2294
0.7	0.3848	0.2207	0.3963	0.2278
0.8	0.3796	0.2183	0.3879	0.2217
0.9	0.3818	0.2219	0.2881	0.2216