

Foundation Models for Generalizable Physiological Signal Representation in Wearable Health Applications

Blake D. Ortiz

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
ortiz2000@colostate.edu

Rasse Phelips

Department of Computer Science, University of Houston, Houston, TX, USA.
jessephillips752@uh.edu

Abstract

Wearable health systems generate continuous photoplethysmographic, electrocardiographic, accelerometric, and related physiological signals across highly diverse populations, device configurations, and environmental conditions. Conventional supervised models trained on isolated datasets often fail to generalize because physiological patterns vary with sensor placement, skin tone, motion, pathology, and recording hardware. Foundation models, which learn transferable representations from large-scale multimodal data, offer a promising system-level alternative. This paper examines the architectural, governance, deployment, robustness, fairness, and sustainability dimensions of foundation models for physiological signal representation. It discusses self-supervised pretraining, temporal and multimodal fusion, domain adaptation, and the structural trade-offs between centralized training and decentralized edge deployment. The analysis emphasizes that clinical utility depends not only on accuracy but also on data governance, regulatory alignment, infrastructure resilience, and equity. Wearable photoplethysmography and motion artifact mitigation serve as a representative challenge, while the discussion remains situated within broader wearable health infrastructure. The paper argues that sustainable foundation model ecosystems require careful coordination among data curation, federated learning, model auditing, and post-deployment monitoring. It concludes with future directions for cross-institutional collaboration, interoperable evaluation, and context-aware personalization.

Keywords

foundation models; wearable health; physiological signals; self-supervised learning; federated learning; fairness; deployment.

1. Introduction

The rapid expansion of wearable sensing platforms has transformed physiological monitoring from episodic clinical measurements into continuous, longitudinal data streams. Devices now capture photoplethysmographic, electrocardiographic, electrodermal, accelerometric, and temperature signals across daily life. However, the heterogeneity of skin properties, sensor placement, motion artifacts, and device specifications creates a foundational generalization challenge. Conventional supervised models trained on a single cohort or sensor configuration frequently exhibit brittle performance when moved across populations or hardware generations [1]. This limitation motivates the development of foundation models capable of

learning reusable representations from large corpora of unlabeled or weakly labeled physiological data. Transformers and other attention-based architectures have shown that scale and self-supervision can yield transferable representations in natural language and vision [2, 3]. The broader concept of foundation models formalizes this observation as a systems agenda for creating models that adapt to many downstream tasks through minimal fine-tuning or prompting [4]. In wearable health, such models promise to reduce dependence on scarce clinical annotations while improving robustness across heterogeneous deployment contexts.

Despite this promise, the adoption of foundation models for physiological signals involves substantial system-level complexity beyond model architecture. Data collection, annotation, privacy protection, edge computation, regulatory approval, fairness, and post-deployment drift all shape whether a representational model can be safely and equitably used. The present paper therefore adopts an interdisciplinary systems perspective. It examines the structural trade-offs in building and operating foundation models for wearable health, with particular attention to generalizable physiological signal representation. Rather than focusing narrowly on algorithmic accuracy, the analysis considers how representation learning, data governance, federated infrastructure, evaluation, and maintenance must be coordinated to achieve robust outcomes. This framing is necessary because physiological signals are not static perceptual inputs; they are dynamic, instrument-dependent, and entangled with individual anatomy, behavior, and health state. A model that performs well on clean laboratory data may fail when deployed in ambulatory environments where motion, low perfusion, and sensor displacement are common. The following sections develop this argument across multiple system layers.

2. Background and Problem Framing

Wearable health applications generate data that are simultaneously high-volume and high-context. A wrist-worn device may record heart rate, photoplethysmographic waveforms, and tri-axial acceleration at tens to hundreds of hertz. These data are correlated with sleep, stress, exercise, medication, hydration, and cardiovascular status. Yet labels for supervised learning are expensive and often unreliable because they require clinical adjudication or intrusive ground truth. In contrast, self-supervised learning creates training signals from the structure of the data itself, such as predicting masked segments of a temporal sequence or contrasting augmented views of the same signal [5, 6]. This paradigm is attractive for physiological time series, where unlabeled recordings are abundant but clinically labeled examples are limited.

Early deep learning successes in medical imaging and physical activity recognition relied on convolutional architectures that extract hierarchical features [7]. However, physiological signals possess long-range dependencies, irregular sampling, and multimodal coupling. A single heartbeat can be understood locally, but conditions such as arrhythmia, sleep apnea, or autonomic dysfunction may require context spanning minutes or hours. Attention mechanisms offer a flexible way to model long-range temporal dependencies, while self-supervised objectives can pretrain on large unlabeled collections [2]. The resulting representations may encode clinically meaningful structure without requiring task-specific labels. Previous work on deep learning in healthcare has highlighted the need for robust validation, interpretability, and integration with clinical workflows [8]. Foundation models for physiological signals extend this line of work by emphasizing transfer across tasks, populations, and devices rather than isolated task-specific models.

However, the translation from concept to operational system is not straightforward. High-performance medical artificial intelligence often depends on curated datasets, careful

preprocessing, and controlled deployment conditions [9]. In wearable contexts, these conditions are rarely stable. Sensors shift, batteries degrade, firmware updates alter filtering, and users vary in skin tone, tissue composition, and behavior. Moreover, machine learning in medicine has repeatedly shown that performance in one dataset does not guarantee performance in another [10]. A systems view therefore requires explicit attention to the ways in which data distributions, sensor characteristics, and user contexts interact. Without such attention, a foundation model may optimize for statistical shortcuts that disappear when devices or populations change. The remainder of this paper develops a systems-level account of foundation models for wearable physiological signals across architectural, infrastructural, deployment, equity, regulatory, and sustainability dimensions.

3. Architectural Foundations and Signal Processing Trade-offs

Physiological signals require representation learners that combine local morphological understanding with global temporal context. Convolutional encoders provide efficient local feature extraction but can be limited in capturing long-range dependencies across minutes or hours [7]. Attention-based transformers address this limitation by allowing each segment to interact with many other segments, enabling the model to integrate context such as circadian variation, activity patterns, and slowly evolving physiological states [2, 3]. In wearable health, a hybrid design often emerges: shallow convolutional layers extract local waveform features, while transformer layers aggregate these features across long sequences. This structure is attractive because it balances computational efficiency with representational power, but it introduces trade-offs. Deeper transformers increase memory and energy consumption, which is problematic for battery-constrained wearables. At the same time, aggressive downsampling or fixed-length segmentation may discard transient events such as ectopic beats or motion artifacts that are clinically informative.

Self-supervised objectives provide a natural training signal for unlabeled physiological data. Masked signal modeling encourages the model to reconstruct missing segments and thus to learn latent temporal dynamics, while contrastive approaches train the model to identify different augmented views of the same physiological event [5, 6]. Each objective imposes different biases. Masked modeling tends to preserve fine-grained temporal structure, which is valuable for detecting subtle waveform changes. Contrastive learning emphasizes invariance to transformations such as amplitude scaling or small time shifts, which can improve robustness to sensor drift. However, if augmentation is too aggressive, contrastive objectives may erase clinically meaningful variation, such as changes in waveform morphology that distinguish benign from pathological conditions. The choice of pretraining objective is therefore not merely an accuracy decision; it is a systems decision that determines what information is retained, what is suppressed, and how the model will behave under deployment shift.

Another architectural consideration concerns the integration of preprocessing and representation. Physiological signals from wearable devices are often corrupted by motion artifacts, baseline wander, and intermittent sensor contact. Many traditional pipelines apply adaptive filtering or artifact removal before model input. Foundation models may either rely on such preprocessing modules, learn to handle artifacts internally, or combine both approaches. If artifact removal is fully internalized, the model may become more robust to unseen noise but may also require greater capacity and larger datasets. If preprocessing remains external, the representation model can be lighter, but its performance becomes

coupled to the correctness of the filtering algorithm. A system-level design may benefit from configurable preprocessing where the filtering intensity is adjusted based on accelerometric context and signal quality estimates, allowing the foundation model to receive cleaner input without fully obscuring informative variability.

Multimodal fusion is another key architectural dimension. Wearable devices simultaneously capture multiple signals, including photoplethysmography, acceleration, skin conductance, and temperature. These modalities are complementary but not synchronized in their noise properties. Early fusion architectures concatenate raw features, which is simple but can be dominated by high-amplitude accelerometric signals. Mid-level fusion aligns latent representations from separate encoders, allowing the model to learn cross-modal interactions such as the relationship between motion and heart rate. Late fusion combines task-level predictions, which preserves modularity but may miss subtle interactions. Foundation models designed for generalizable physiological representation increasingly adopt cross-modal attention, where one modality can query information from another. This approach is promising but introduces substantial computational costs and complicates privacy-preserving distributed training. The architectural choice therefore shapes not only representational quality but also the feasibility of edge deployment and federated optimization.

4. Data Infrastructure, Governance, and Federated Coordination

Foundation models depend on large, diverse, and well-governed data collections. Wearable physiological data are generated in the private sphere, which creates tensions between scale and privacy. Centralized data repositories are efficient for model training and auditing but concentrate risk. Federated learning offers an alternative in which models are trained across distributed devices or institutions without centralizing raw data [11]. In this paradigm, local updates are aggregated through secure protocols, allowing the model to learn from population diversity while reducing direct data exposure. Federated health learning has been advanced as a way to enable multi-institutional collaboration without sharing sensitive patient records [12, 13]. However, federated training introduces structural complexity: devices have heterogeneous compute capabilities, network connectivity is intermittent, and local data distributions are non-identically distributed. The global model must reconcile updates from participants whose data may reflect different age groups, skin tones, activity levels, or sensor hardware. Without careful weighting and personalization, the aggregated representation may overlook minority subpopulations or converge toward the most common data regime.

Data governance also involves consent, provenance, and lifecycle management. Wearable users often provide broad consent for research or commercial use without a clear understanding of how their data may be reused across multiple downstream models. Foundation models amplify this concern because pretrained representations may be fine-tuned for unforeseen clinical tasks. Organizations must implement mechanisms for data provenance, usage restrictions, and model lineage tracking. A governance framework should specify how data are curated, how quality is assessed, how sensor metadata are preserved, and how users can withdraw consent. These requirements are not merely administrative; they shape the structure of the training system. For example, if consent withdrawal requires removing the influence of a user's data on a trained model, the system may need to adopt unlearning techniques or maintain strict data partitioning and checkpointing.

Interoperability and metadata standards further influence downstream generalization. Physiological recordings without device metadata, sampling rates, sensor locations, and activity contexts are difficult to combine safely. A foundation model trained on heterogeneous

datasets must learn to condition on such metadata or to infer it from signal statistics. This can be achieved through metadata-aware encoders, domain adversarial training, or prompting mechanisms that supply context to the transformer. Yet these approaches require data curators to collect and standardize metadata, which is often missing in legacy datasets. The absence of standardized formatting creates a hidden tax on foundation model development and limits cross-study evaluation. A systems-level program should therefore invest in open data standards, secure data enclaves, and computational provenance frameworks that enable diverse datasets to be used without sacrificing governance.

5. Deployment, Robustness, and Edge System Trade-offs

Deployment of foundation models for wearable health spans a continuum from cloud-based inference to fully on-device representation. Cloud deployment simplifies model updates and enables large models but introduces latency, connectivity dependence, and privacy concerns. Edge deployment reduces data transmission and enables real-time feedback, but it constrains model size, memory, and energy. Most practical deployments adopt a tiered architecture: a lightweight encoder runs on the wearable to extract compact representations, while more expensive downstream reasoning occurs on a paired smartphone or a cloud service. This tiering aligns with the foundation model paradigm if the wearable encoder produces generalizable representations that can support multiple health tasks without retraining the entire stack. However, edge compute imposes severe constraints on attention-based models. Model compression through distillation, quantization, and pruning can reduce resource usage but may degrade representation quality or introduce quantization errors that disproportionately affect low-amplitude physiological events.

Continuous monitoring of representation quality is necessary to detect model degradation in real-world settings. Sensor drift, skin changes, and device replacement can slowly shift input distributions. Representational drift may be measured through downstream task performance, signal reconstruction error, or stability of latent embeddings. When degradation is detected, the system must trigger recalibration or retraining [14]. This monitoring is especially important for foundation models that are intended to remain in service across many months or years. Without such monitoring, a foundation model may silently degrade even if its offline benchmarks remain stable.

Robustness to motion artifact and sensor variation is essential. Wearable photoplethysmography is particularly sensitive to motion, and adaptive filtering strategies can recover signal quality in challenging physical activity conditions [15]. A foundation model deployed downstream of such filtering may benefit from cleaner input, but the filtering itself must adapt to individual movement patterns. If the model is expected to operate across different filtering configurations or raw signals, it must learn invariances that do not depend on a single preprocessing pipeline. This can be achieved through data augmentation during pretraining, including simulated motion artifact, sensor dropouts, and baseline shifts. Robustness evaluation should therefore treat preprocessing as a source of distribution shift rather than an assumed fixed stage.

On-device personalization adds another layer of complexity. A global foundation model can be adapted to an individual using small amounts of local data through lightweight fine-tuning or adapter modules. This improves personalization while preserving shared knowledge. However, personalization must be carefully constrained to avoid overfitting to transient states, such as acute illness or temporary medication effects. The deployment system must decide when personalized parameters should be updated, how often they should be synchronized

with a global model, and how to maintain interpretability for clinicians who may need to understand whether a prediction is based on global population patterns or on an individual's recent history.

6. Fairness, Bias, and Equity

Representation learning does not automatically yield equitable performance. If training data overrepresent certain demographics, device types, or activity conditions, the learned representations may encode harmful biases that propagate across downstream tasks. In healthcare algorithms, bias can arise from unequal access, historical disparities, and underrepresentation in training cohorts [17]. For wearable physiological signals, skin tone is a known factor in photoplethysmographic signal quality because optical absorption varies with melanin content. A model trained predominantly on lighter skin tones may learn features that depend on signal amplitude rather than underlying cardiovascular dynamics, leading to degraded performance for darker skin tones. Fairness therefore requires deliberate data collection across diverse populations, transparent reporting of demographic composition, and subgroup-level evaluation.

Bias can also arise from sensor placement and device cost. Wrist-worn devices may be less accessible to certain populations, and their signal quality may differ across body mass index, age, and peripheral perfusion. A foundation model that learns from a narrow device base may fail to generalize to lower-cost or alternative form-factor sensors. Fairness considerations should therefore extend beyond demographic categories to include device equity and condition-specific representation. Auditing foundation models requires more than aggregate accuracy; it requires stratified evaluation across sensor metadata, clinical subgroups, and environmental contexts. When disparities are detected, mitigation may involve targeted data collection, reweighting strategies, or hierarchical modeling that separates population-level patterns from individual-level factors. These approaches must be integrated into the training and governance systems, not applied as post hoc corrections.

7. Evaluation, Monitoring, and Regulatory Implications

Evaluation of foundation models for physiological signals requires careful benchmark construction. Current datasets often emphasize healthy volunteers in laboratory settings, which systematically differs from clinical populations in ambulatory environments. Benchmark tasks should include representation quality measures, downstream clinical tasks, and robustness under distribution shift. Because a foundation model is intended to transfer across tasks, evaluation must also assess adaptation efficiency: how many labeled examples are required to reach clinically acceptable performance on a new task or population? This question is important for real-world feasibility, since clinical annotation is expensive. Examples of strong deep learning performance in cardiology demonstrate the potential of large labeled datasets, but such datasets are rarely available for every wearable context [14]. Foundation models reduce this burden, but their adaptation efficiency must be characterized rather than assumed.

Clinical validation must also be connected to real clinical workflows. Previous work has emphasized the importance of integrating deep learning systems with electronic health record data and clinician-facing interfaces to ensure that model outputs are actionable and interpretable [18]. A representation model that is accurate but detached from clinical context may generate alerts that are difficult for clinicians to interpret or reconcile with other patient

information. Therefore evaluation should include clinician interaction, decision support quality, and workflow integration, not solely standalone predictive metrics.

Continuous post-deployment monitoring is a system-level requirement. Unlike static software, learned models can drift as populations, devices, and behaviors change. Monitoring should include representation drift, calibration, fairness metrics, and task-specific performance where ground truth is available. In many wearable settings, ground truth is sparse, so proxies such as signal quality indices and physiological plausibility may be used [16]. When performance degradation is detected, the system must support safe rollback, model update, and incident reporting. Regulatory frameworks increasingly emphasize lifecycle management for adaptive algorithms. The U.S. Food and Drug Administration has outlined expectations for AI-based software as a medical device, including plans for iterative improvement and real-world performance monitoring [19]. Similar considerations are emerging in international regulatory guidance. A foundation model used in wearable health must therefore be embedded within a quality management system that links model versioning, data provenance, deployment configuration, and clinical oversight.

8. Sustainability and Long-Term Ecosystem Design

The sustainability of foundation models depends on computational, environmental, and organizational factors. Pretraining large models on continuous physiological data can require significant energy and hardware resources [20]. Although wearable models are typically smaller than language models, large-scale multimodal pretraining still consumes considerable compute. Sustainable design requires careful attention to model size, data efficiency, and hardware utilization. Distillation from larger models to compact wearable encoders can reduce deployment energy, but the pretraining cost remains. Organizations should report energy consumption and consider trade-offs between model scale and marginal performance gains.

Long-term ecosystem design also involves maintaining shared infrastructure across institutions. Open model weights, standardized evaluation suites, and community benchmarks can reduce redundant training and enable scientific reproducibility. However, open releases raise governance questions about misuse, especially if representations could be used for unregulated health inferences. A sustainable foundation model ecosystem therefore balances openness with responsible access controls, perhaps through staged releases, model cards, and structured evaluation agreements. Additionally, versioning across models, datasets, and filtering pipelines must be coordinated so that deployed systems can be reproducibly audited. Without such coordination, incremental updates may create hidden inconsistencies across devices and institutions.

9. Conclusion

Foundation models for physiological signal representation offer a systemic response to the generalization challenges of wearable health. By learning reusable representations from large-scale, multimodal, and self-supervised data, these models can reduce dependence on scarce clinical labels and improve adaptability across tasks, devices, and populations. Yet their value depends on more than representation quality. The architecture must balance long-range temporal modeling with edge efficiency. Data governance must reconcile scale with privacy, consent, and inclusion. Deployment must integrate preprocessing, robustness monitoring, and personalization without sacrificing reliability. Fairness must be treated as a first-class system property, not an afterthought. Regulatory and sustainability considerations must guide the

entire lifecycle from data collection through post-market surveillance. Adaptive signal processing methods, including motion artifact mitigation, remain important components of this pipeline and illustrate the need for close coupling between signal conditioning and representation learning [15]. As wearable health systems continue to expand, interdisciplinary coordination across machine learning, clinical medicine, data governance, and systems engineering will determine whether foundation models fulfill their promise or reproduce the brittleness of previous generations of health algorithms.

References

1. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
<https://doi.org/10.1109/TKDE.2009.191>
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.
4. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
5. Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
6. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597–1607.
7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
8. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.
9. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
10. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(6), e1002683.
11. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, 1273–1282.

12. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.
13. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D. S., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598.
14. Hannun, A. Y., Rajpurkar, P., Haghighpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65–69.
15. Zheng, X., Dwyer, V. M., Barrett, L. A., Derakhshani, M., & Hu, S. (2022). Adaptive notch-filtration to effectively recover photoplethysmographic signals during physical activity. *Biomedical Signal Processing and Control*, 72, 103303.
16. Li, Q., & Clifford, G. D. (2012). Signal quality and data fusion for false alarm reduction in the intensive care unit. *Journal of Electrocardiology*, 45(6), 596–603.
17. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
18. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18.
19. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
20. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.