

Contrastive Learning for Motion-Robust Representation of Wearable Photoplethysmographic Signals

Martins Lyons

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

Stanley Larsen

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
stanleyl@buffalo.edu

Abstract

Wearable photoplethysmographic sensors have become central to continuous health monitoring, but their real-world reliability is undermined by motion artifacts, sensor displacement, skin variability, and environmental interference. Conventional signal processing pipelines rely on heuristic filtering and accelerometer-assisted noise rejection that often fail under nonstationary daily activity. This paper examines contrastive learning as a systems-oriented strategy for learning motion-robust representations from wearable photoplethysmographic signals. By constructing latent spaces in which augmented views of the same physiological context remain close while dissimilar contexts are separated, contrastive models can suppress nuisance distortions while preserving diagnostically relevant signal structure. The analysis extends beyond algorithmic accuracy to address architectural trade-offs, edge-cloud partitioning, training infrastructure, governance, fairness, sustainability, and regulatory implications. It argues that motion-robust representation is not merely a model-level property but a structural outcome of careful augmentation design, encoder selection, deployment architecture, and continuous monitoring. The paper further considers how federated learning and on-device inference can support privacy-preserving representation learning, while stratified evaluation and governance mechanisms are needed to avoid widening health disparities. By integrating technical and policy perspectives, the discussion provides a research agenda for building robust, equitable, and maintainable wearable sensing infrastructures.

Keywords

contrastive learning; photoplethysmography; wearable health monitoring; motion artifact; fairness; governance; infrastructure.

1. Introduction

Wearable photoplethysmographic sensing has expanded rapidly across consumer devices, clinical remote monitoring, and occupational health applications. Photoplethysmographic sensors measure blood volume changes in microvascular tissue through optical absorption and reflection, enabling noninvasive estimation of heart rate, oxygen saturation, respiratory patterns, and vascular dynamics [1]. The widespread deployment of wrist-worn and earbud-based devices has created a continuous stream of physiological data that can support early detection, chronic disease management, and longitudinal health assessment. However, the

transition from controlled laboratory settings to unstructured daily life has intensified a fundamental challenge: the raw optical signal is highly vulnerable to motion, sensor movement, ambient illumination, tissue compression, and differences in skin properties. These disturbances degrade waveform morphology and introduce uncertainty into downstream physiological estimates, limiting the clinical and operational value of wearable sensing systems.

The limitations of conventional processing pipelines are now widely recognized. Traditional approaches combine filtering, peak detection, spectral tracking, and heuristic plausibility checks, often using accelerometer signals to identify and suppress motion-related frequencies [2]. Although such methods can perform well in constrained scenarios, they tend to rely on assumptions about stable sensor contact, stationary noise, and predictable motion patterns. Daily activities violate these assumptions through complex, nonstationary movements and heterogeneous sensor coupling. As a result, the accuracy of motion-contaminated photoplethysmographic interpretation can vary significantly across users, devices, and contexts. This variability raises concerns not only about performance but also about fairness, clinical trust, and the long-term viability of wearable monitoring infrastructures.

Recent advances in contrastive representation learning offer a compelling path toward more robust photoplethysmographic analysis. Contrastive methods learn encoders by maximizing agreement between semantically related views of data while minimizing agreement between unrelated views. This paradigm has achieved notable success in computer vision, speech, and natural language processing, and its principles can be adapted to time-series bio signals. For wearable photoplethysmography, contrastive learning provides a mechanism to build representations that are explicitly invariant to motion-related distortions while retaining physiological content. The present paper addresses this emerging area from a systems perspective, emphasizing architectural design, computational trade-offs, governance, deployment sustainability, and policy implications rather than isolated algorithmic novelty.

2. Motion Degradation in Wearable Photoplethysmographic Systems

Motion artifact in wearable photoplethysmography arises from the complex interplay of optical, mechanical, and physiological factors. The optical path depends on stable contact between the light emitter and the photodetector against the skin. Relative displacement alters the scattering geometry and introduces low-frequency drift and high-frequency spikes unrelated to cardiac pulsation. Mechanical vibration and arm swing produce periodic acceleration whose harmonics can overlap with heart rate frequencies, confounding spectral methods. Physiological phenomena such as vasoconstriction, localized sweating, and venous pulsation further modulate the signal in ways that cannot be predicted from accelerometer amplitude alone. Consequently, motion robustness must be understood as an inferential property of the system rather than a simple artifact removal stage. Systematic reviews of wrist-worn photoplethysmographic sensing demonstrate high inter-subject variability in motion tolerance [3]. Clinical validation studies further show that passive monitoring during everyday activity introduces errors that challenge algorithmic thresholds established in laboratory settings [4].

Traditional multi-sensor fusion techniques use accelerometer data as a reference channel to detect and suppress motion-corrupted segments. While this approach can reduce gross artifacts, it does not capture subtle morphological changes caused by tissue deformation or light path variation. The nonlinear and nonstationary nature of these disturbances implies that no fixed transformation of the raw signal can fully recover the underlying cardiac waveform.

A representation learning approach therefore becomes appealing because it can learn data-driven invariances that adapt to the statistical structure of real-world motion without requiring explicit noise models.

3. Contrastive Representation Learning for Physiological Time Series

Contrastive learning is a self-supervised paradigm that trains encoders by pulling augmented versions of the same example together in an embedding space while pushing different examples apart [5]. The resulting representations are useful for downstream tasks because they encode high-level semantic content rather than low-level nuisance variation. Contrastive predictive coding extended this principle to sequential data by predicting future representations from past observations [6]. For physiological monitoring, time series contrastive methods have been applied to electrocardiography and photoplethysmography to learn features that are invariant across sensor placements, patients, and recording conditions [7]. Self-supervised contrastive models have also been shown to improve emotion recognition from cardiac signals, indicating that the learned representations retain diagnostically and behaviorally relevant information [8].

Adapting contrastive learning to wearable photoplethysmography requires careful consideration of what constitutes a positive pair. In computer vision, geometric and color augmentations define identity-preserving transformations. In photoplethysmography, plausible transformations include time warping, amplitude scaling, noise injection, and segmentation of adjacent windows. Each augmentation must preserve the physiological content of interest while simulating realistic motion corruptions. Overly aggressive augmentations can destroy diagnostic features such as pulse wave morphology, whereas weak augmentations produce representations that remain sensitive to motion.

4. Architectural Design and Encoding Strategies

The choice of encoder architecture shapes the capacity to extract co-occurring temporal and spectral features. Convolutional neural networks over raw temporal windows are common, but recurrent and attention-based architectures provide stronger sequence modeling for variable-length segments. Multi-scale temporal convolutions can capture local pulse morphology and longer respiratory modulations simultaneously. A systems perspective emphasizes that the encoding strategy must be co-designed with deployment constraints, including memory, power, and latency budgets on wearable devices. While larger transformer-based encoders may improve representation quality, they can exceed the computational envelope of low-power microcontrollers. The contrastive training process therefore includes a structural trade-off between representational expressiveness and on-device feasibility.

Temporal contrastive objectives can be computed over augmented windows drawn from the same subject and recording context, while global contrastive objectives may use different channels such as multiple wavelengths. The use of cross-channel information aligns with opto-physiological monitoring systems that combine green, red, and infrared photoplethysmographic signals to estimate oxygen saturation and heart rate. Structural trade-offs also appear in the granularity of the contrastive unit. Window-level contrast discriminates short perturbed segments but may overlook session-level drift, whereas segment-level contrast can capture slower changes associated with vasomotor tone and autonomic state. The tension between local and global contrastive objectives reflects a broader systems design question: whether robustness should be optimized for short-term artifact rejection or for long-

term physiological stability across changing contexts. Hybrid objectives that combine window-level and session-level contrast may balance these goals, but they increase computational cost and complicate hyperparameter tuning. Furthermore, the learned representations must remain interpretable enough to support clinical validation and regulatory review, placing constraints on the degree of abstraction that can be introduced without losing traceability to physiological features.

Multi-channel optical signals provide complementary information, but their integration raises architectural choices. Green light is often preferred for heart rate monitoring because of its relative motion tolerance, whereas red and infrared channels contain richer information for oxygen saturation. Contrastive learning can be structured so that different wavelength views are treated as positive pairs, encouraging representations that are stable across optical configurations. This strategy may improve generalization across device manufacturers and sensor placements, but it also requires large and diverse training corpora that include multiple device types and skin tones. Without such diversity, the contrastive objective may learn invariances that are specific to a narrow population or hardware platform.

5. Training Infrastructure, Privacy, and Data Governance

The development of motion-robust contrastive models depends on access to large and heterogeneous photoplethysmographic datasets that reflect realistic daily motion, sensor variability, and demographic diversity. Such datasets are difficult to centralize because physiological records are sensitive personal data and are increasingly subject to privacy regulations. Federated learning provides a structural alternative in which model updates rather than raw signals are transmitted to a central coordinator [9]. This approach aligns with the need for large-scale representation learning while reducing the risk of centralized data exposure. In a federated contrastive framework, local encoders can be trained on device-specific data and shared through encrypted aggregation, preserving user privacy while capturing population-level variability. However, federated systems introduce additional governance challenges, including synchronization overhead, heterogeneous hardware, and the risk of model inversion attacks that may partially reconstruct training inputs from gradient information. The system design must therefore include differential privacy mechanisms, secure aggregation protocols, and audit trails that document data provenance and model updates.

Data governance for wearable photoplethysmographic learning must also address consent, ownership, and secondary use. Continuous physiological monitoring generates longitudinal records that are far more revealing than isolated clinical measurements. The contrastive training pipeline should be embedded within a governance framework that distinguishes between model training, internal validation, and post-deployment monitoring. Clear policies are needed for how long edge devices retain data, which transformations are applied before transmission, and how users can withdraw consent. These governance choices are not peripheral to the technical design; they shape the data distributions available for training and therefore influence the robustness and fairness of the learned representations. A system that relies on implicit consent and ambiguous data retention may accumulate certain demographic groups while underrepresenting others, producing encoders that perform unevenly across populations.

6. Fairness, Bias, and Evaluation of Motion-Robust Representations

Fairness evaluation for wearable photoplethysmographic systems requires going beyond aggregate accuracy metrics. Motion robustness can vary with skin tone, body mass, sensor tightness, ambient temperature, and activity patterns. If the training corpus contains predominantly young, healthy, light-skinned subjects with high device compliance, the learned representation may fail for older adults, individuals with darker skin, or users with peripheral vascular conditions. The fairness literature in machine learning emphasizes that demographic and contextual disparities must be measured explicitly rather than inferred from overall performance [10]. For contrastive learning, this implies stratified evaluation across subpopulations and motion conditions, with separate analyses of false negatives and estimation bias. Relying on a single global error metric can mask systematic underperformance for minority groups.

Evaluation protocols should also account for distribution shift between training and deployment. Contrastive representations are often validated using linear probes or fine-tuning on downstream tasks such as heart rate estimation, respiratory rate inference, or arrhythmia detection. However, the choice of validation task influences the apparent robustness of the learned features. A model may achieve low mean absolute error on average while exhibiting high variance among users with high motion levels. Cross-dataset and cross-device validation is therefore essential. Furthermore, the benchmark datasets themselves may encode historical inequities in sensor design, because optical photoplethysmography was originally validated on limited populations. Systematic audits of training data composition and evaluation protocols are necessary to prevent contrastive learning from amplifying these inherited biases.

7. Edge-Cloud Deployment and Sustainability

Deploying contrastive photoplethysmographic encoders in wearable devices introduces competing constraints related to energy, memory, and latency. High-capacity models may achieve superior motion robustness but cannot be executed continuously on battery-constrained hardware. Model compression techniques such as pruning, quantization, and knowledge distillation can reduce footprint, but aggressive compression may erode the very invariances learned during contrastive training [11]. The systems design must therefore balance the accuracy gains of larger encoders against the operational cost of frequent recharging and device heating. One viable strategy is hierarchical inference, in which a lightweight on-device encoder performs initial representation extraction and a more powerful cloud model refines ambiguous cases. However, this architecture increases communication overhead and may introduce latency unsuitable for real-time safety-critical applications.

Sustainability is another structural consideration. Continuous physiological monitoring generates large volumes of data, and contrastive training over long time series is computationally intensive. The carbon footprint of training large models must be weighed against the clinical benefit of improved motion robustness. Edge-cloud partitioning can reduce the need for centralized retraining, but frequent model updates and communication round trips consume energy. Sustainable deployment requires not only algorithmic efficiency but also lifecycle assessment of the hardware, including device replacement rates, battery degradation, and data center energy use. A motion-robust representation that requires frequent cloud synchronization may be less sustainable than a moderately less accurate but fully on-device model. These trade-offs should be made explicit in system design documentation and procurement decisions.

8. Attention Mechanisms and Multi-Sensor Integration

Recent work on opto-physiological monitoring has explored joint attention mechanisms to improve robustness during physical activity [12]. Attention layers can dynamically weight temporal segments, spectral bands, or sensor channels according to their reliability, effectively learning to ignore motion-corrupted portions while preserving clean signal segments. When integrated with contrastive objectives, attention can be trained to align with the same invariances that the encoder is learning. A joint attention architecture may operate across multiple wavelengths and accelerometer channels, producing a fused representation that is more stable under motion than any single channel alone. This approach aligns with the broader movement toward multimodal wearable sensing, where photoplethysmographic signals are combined with electrodermal activity, skin temperature, and accelerometry. The structural challenge is that attention mechanisms increase model complexity and require careful regularization to prevent overfitting to specific motion patterns. However, their capacity for interpretability, through attention weights, can support clinical review and regulatory acceptance if the weights correspond to physiologically meaningful events.

The integration of attention with contrastive learning also raises questions about the ordering of training stages. Attention modules may be trained jointly with the contrastive encoder, or they may be added after a pretrained encoder has been frozen. Joint training can improve representational quality but may destabilize convergence, especially with nonstationary photoplethysmographic data. Staged training separates the representation learning from the attention fusion, allowing more controlled debugging but potentially missing interactions between channel reliability and motion. A systems perspective suggests that the deployment context should determine this choice: safety-critical applications may favor the interpretability and stability of staged training, while exploratory research may tolerate the greater flexibility of joint optimization.

9. Regulatory and Policy Considerations

The translation of contrastive photoplethysmographic representations into clinical and consumer health products involves multiple regulatory pathways. In many jurisdictions, algorithms that estimate heart rate or oxygen saturation may be considered medical devices if they are intended for diagnosis, treatment, or disease management. The self-supervised nature of contrastive learning complicates validation because the training objective does not directly optimize a clinical endpoint. Regulatory reviewers increasingly require evidence that learned representations are stable across intended use populations and environments. Benchmarks for TinyML systems provide one reference for documenting hardware-level performance, but clinical validation demands additional evidence of physiological accuracy and safety [13]. Systematic post-market surveillance is necessary to detect performance drift as user behavior and device firmware evolve.

Policy mechanisms should also address the data lifecycle of wearable photoplethysmographic systems. Wearable technology applications in healthcare have expanded beyond wellness into disease management, creating responsibilities for manufacturers and healthcare providers [14]. Transparent reporting of training data composition, augmentation strategies, and failure modes can support accountability. Regulatory frameworks should require that motion robustness be evaluated not only under standardized laboratory protocols but also under free-living conditions representative of diverse users. The governance of these systems must also clarify liability when motion artifacts contribute to erroneous health estimates that lead to harm. A clear chain of responsibility, from sensor manufacturer to algorithm developer to clinical integrator, is essential for maintaining trust and enabling safe innovation.

10. Clinical Translation and Data Integration

The clinical value of motion-robust photoplethysmographic representations depends on their integration with broader health information systems. Electronic health records contain longitudinal patient data, medications, laboratory results, and diagnoses that can contextualize wearable signals. Mining electronic health records has matured as a field, offering methods to link physiological time series with clinical outcomes [15]. High-performance medicine increasingly relies on the convergence of high-resolution sensor data and artificial intelligence to inform decision-making [16]. Machine learning in medicine more broadly has demonstrated the potential to extract predictive patterns from heterogeneous data, but also highlighted the need for rigorous validation and prospective trials [17]. For contrastive photoplethysmographic models, this implies that motion-robust representations should be evaluated not only on signal reconstruction or heart rate accuracy but also on clinically meaningful endpoints such as exacerbation detection, hospitalization risk, or therapeutic response.

Ethical challenges arise when machine learning systems are embedded in clinical workflows without clear accountability structures [18]. The same contrastive learning pipeline that improves motion robustness may also obscure the physiological basis of its predictions, complicating clinician trust and patient communication. Big data infrastructures for health monitoring have expanded rapidly, creating opportunities for population-scale analytics but also risks of data fragmentation and inconsistent curation [19]. Communication-efficient decentralized learning methods, including federated averaging, can mitigate some of these risks by enabling collaborative training across institutions without centralizing raw physiological data [20]. However, the governance of such distributed systems must be designed alongside the technical architecture to ensure that data sharing, model aggregation, and clinical deployment respect patient consent and institutional policies.

11. Conclusion

Contrastive learning offers a promising foundation for building motion-robust representations of wearable photoplethysmographic signals, but its success depends on more than algorithmic advances. The design of augmentations, encoder architectures, training infrastructures, and deployment strategies must be treated as interconnected elements of a larger socio-technical system. Motion robustness is best understood as an emergent property of this system, influenced by data diversity, privacy safeguards, fairness evaluation, energy constraints, and regulatory oversight. Attention-based and multi-channel approaches can further enhance the learned invariances, but they introduce structural trade-offs that must be managed across the entire lifecycle. A research agenda that integrates technical innovation with governance, sustainability, and equity will be necessary to move wearable photoplethysmographic monitoring from controlled demonstrations to dependable, population-scale health infrastructures.

References

1. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), R1–R39. <https://doi.org/10.1088/0967-3334/28/3/R01>
2. Zhang, Z., Pi, Z., & Liu, B. (2015). TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical

- exercise. *IEEE Transactions on Biomedical Engineering*, 62(2), 522–531.
<https://doi.org/10.1109/TBME.2014.2359372>
3. Biswas, D., Simões-Capela, N., Van Hoof, C., & Van Helleputte, N. (2019). Heart rate estimation from wrist-worn photoplethysmography: A review. *IEEE Sensors Journal*, 19(16), 6560–6570. <https://doi.org/10.1109/JSEN.2019.2914166>
 4. Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine*, 3, 18.
<https://doi.org/10.1038/s41746-020-0226-6>
 5. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning*, 1597–1607.
 6. Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
 7. Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. *Proceedings of the International Conference on Machine Learning*, 5606–5615.
 8. Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3), 1541–1554.
 9. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.
<https://doi.org/10.1038/s41746-020-00323-1>
 10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
<https://doi.org/10.1145/3457607>
 11. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations*.
 12. Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. *Biomedical Signal Processing and Control*, 113, 108949.
 13. Banbury, C., Reddi, V. J., Lam, M., Fu, W., Fazel, A., Holleman, J., ... & Whatmough, P. (2020). Benchmarking TinyML systems: Challenges and direction. *arXiv preprint arXiv:2003.04821*.
 14. Wu, M., & Luo, J. (2019). Wearable technology applications in healthcare: A literature review. *Online Journal of Nursing Informatics*, 23(3).
 15. Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6), 1–40.
 16. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
 17. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.

18. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), e1002689.
19. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171–209.
20. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.