Self-Supervised Learning for Robust Physiological Signal Representation in Wearable Devices

Yangjun Lei
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
yangjunlei879@buffalo.edu

Claude Warner
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
claudewarner1998@missouri.edu

Leo Lewis
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
lewis1986@unr.edu

Advait A. Venkataraman
Department of Computer Science, George Mason University, Fairfax, VA, USA.
advaitv502@gmu.edu

Abstract

Wearable physiological monitoring has become a central element of continuous health assessment, yet the translation of raw biosignals into clinically meaningful representations remains constrained by noise, inter-subject variability, motion artifacts, and the high cost of labeled data. Self-supervised learning offers a structurally significant alternative by exploiting large volumes of unlabeled physiological recordings to learn representations that can be transferred across downstream tasks. This paper presents a system-level examination of self-supervised representation learning for wearable physiological signals, with emphasis on architectural design, deployment constraints, robustness, data governance, fairness, and long-term sustainability. The study connects contrastive, predictive, and generative self-supervised paradigms to the specific characteristics of electrocardiography, photoplethysmography, and related modalities. It further analyzes how these representations can be embedded within resource-constrained wearable ecosystems while preserving utility across heterogeneous populations and real-world operating conditions. Drawing on cross-domain evidence from computer vision, natural language processing, speech, and clinical time series analysis, the paper identifies structural trade-offs between representation quality, computational cost, adaptability, and regulatory accountability. The discussion includes model governance considerations, bias amplification risks, data provenance, and the environmental implications of large-scale pretraining. Rather than proposing a single algorithmic solution, the paper develops an interdisciplinary perspective on how self-supervised representation learning can support robust, equitable, and operationally viable physiological monitoring infrastructures. The conclusion outlines forward-looking research directions that integrate technical advancement with institutional and policy-level safeguards.

Keywords

self-supervised learning; wearable devices; physiological signal processing; representation learning; robustness; health data governance.

1. Introduction

The proliferation of wearable devices has fundamentally altered the scale and temporality of physiological data collection. Modern wrist-worn, chest-mounted, and patch-based systems continuously record electrocardiographic, photoplethysmographic, electrodermal, temperature, and inertial signals across weeks or months, generating data streams that far exceed the volume of traditional clinical monitoring. Deep learning has substantially improved the ability to model these complex, high-dimensional signals [1]. Wearable biosensors now support applications ranging from arrhythmia screening to stress inference and sleep staging, and their clinical relevance continues to expand [2]. However, the performance of supervised deep learning depends on large annotated datasets, which are difficult to obtain for physiological signals because expert labeling is expensive, inter-rater variability is considerable, and many clinically significant events are rare.

Self-supervised learning has emerged as a powerful approach to reduce dependence on manual annotation. In computer vision, frameworks such as contrastive learning of visual representations have shown that features learned without labels can rival or exceed those obtained through fully supervised training [3]. Momentum-based contrastive methods have further stabilized large-batch representation learning by maintaining dynamic dictionaries of encoded samples [4]. In natural language processing, masked language modeling has established a strong precedent for learning contextual representations that transfer effectively to downstream tasks [5]. Contrastive predictive coding has likewise demonstrated that representations can be learned by predicting future observations in latent space, a principle that is especially relevant for sequential physiological data [6]. These developments suggest that wearable physiological monitoring could benefit from pretraining on large unlabeled corpora, followed by lightweight adaptation to clinical or wellness-oriented targets.

The central challenge is not simply the adoption of self-supervised objectives but the construction of representation learning systems that are robust across sensor types, body locations, activity states, and patient populations. Physiological signals are nonstationary, subject to strong noise processes, and deeply entangled with behavioral context. A representation that performs well in a controlled laboratory protocol may degrade substantially under free-living conditions. Consequently, the design of self-supervised physiological representation systems requires a systemic perspective that integrates algorithmic choices with hardware constraints, data governance structures, fairness considerations, and deployment pathways. This paper examines these interconnected dimensions. It does not focus narrowly on a single architecture or dataset. Instead, it develops a conceptual and technical analysis of how self-supervised learning can support robust physiological signal representation in wearable devices while remaining attentive to the social and institutional contexts in which these systems operate.

2. Background and Related Work

The analysis of long-duration physiological signals has traditionally relied on handcrafted features derived from time-domain morphology, frequency-domain power distributions, and nonlinear dynamics. Deep learning for time series classification has expanded this paradigm by enabling end-to-end feature extraction from raw or minimally preprocessed signals [11]. However, the transferability of deep representations remains a central concern. Transfer

learning provides a formal basis for adapting models trained in one domain to related domains with limited labels, and its principles are directly relevant to wearable physiological monitoring where target conditions are often sparsely annotated [12]. Strong baseline models for time series also demonstrate that carefully designed deep networks can learn useful features without extensive domain-specific preprocessing, but they still require substantial labeled data for each task [13].

The roots of self-supervised representation learning lie partly in early work on invariant mapping and metric learning, where the goal was to learn embeddings that bring semantically similar instances closer while separating dissimilar instances [14]. This conceptual foundation has been extended through contrastive and non-contrastive methods across multiple data modalities. In physiological signal analysis, self-supervised methods have been applied to electroencephalographic data to uncover structure that supports downstream clinical tasks [7]. Cardiac signal representations have been learned by contrasting recordings across spatial leads, temporal segments, and patient identities [8]. Self-supervised electrocardiographic representations have also been developed for emotion recognition, suggesting that representations learned without diagnostic labels can retain affective information relevant to mental health applications [9]. Public physiological repositories such as PhysioNet have been essential for these advances by providing large, heterogeneous collections of recordings and annotations [10].

Within wearable monitoring specifically, rapid extraction of vital signs from optical and electrical signals under varying physical activity intensity remains an active engineering problem [15]. Motion artifacts, sensor displacement, and autonomic fluctuations complicate the extraction of stable physiological parameters. This line of work illustrates the broader need for representations that are not only accurate but also resilient to real-world variability. The use of self-supervised pretraining in this context can be interpreted as a mechanism for learning invariances to common perturbations before task-specific adaptation.

Beyond algorithmic performance, the deployment of physiological machine learning systems raises important governance questions. Algorithmic fairness research has shown that clinical risk scores and healthcare allocation models can encode and amplify racial and socioeconomic biases [16]. High-performance medical artificial intelligence is often described as a convergence of human expertise and machine intelligence, but this convergence depends on careful integration into clinical workflows [17]. Guidelines for deep learning in healthcare emphasize the importance of external validation, interpretability, and attention to distributional shift [18]. These considerations are especially important for wearable systems that operate outside the controlled boundaries of hospital environments.

3. Foundations of Self-Supervised Physiological Representation Learning

The design space for self-supervised physiological representation learning can be organized around predictive, contrastive, and reconstructive objectives. Predictive methods derive supervision from the temporal structure of the signal itself, typically by requiring a model to forecast future signal segments or hidden states. Contrastive methods learn representations by distinguishing augmented variants of the same physiological segment from variants of other segments. Reconstructive methods require the model to recover corrupted or masked portions of the input. Each family offers distinct advantages for wearable physiological data. Predictive objectives align naturally with the sequential nature of electrocardiography and photoplethysmography, but they can become unstable when the signal contains high-frequency noise or abrupt movement artifacts. Contrastive objectives encourage invariance to

transformations such as amplitude scaling, time warping, and additive noise, but their effectiveness depends on the selection of physiologically meaningful augmentations. Reconstructive objectives can learn rich local structure, but they may emphasize low-level signal detail at the expense of clinically relevant semantic features.

Cross-domain experience provides useful guidance. Large-scale contrastive pretraining with natural language supervision has shown that aligned multi-modal representations can support robust zero-shot transfer [19]. Speech representation learning has similarly benefited from self-supervised objectives that combine contrastive losses with masked prediction over raw audio waveforms [20]. Denoising autoencoders introduced the idea that robust features can emerge from learning to reconstruct clean inputs from corrupted versions, a principle that is particularly relevant to movement-corrupted wearable signals [21]. Variational autoencoders provide a probabilistic framework for learning smooth latent spaces that can represent uncertainty and support generative data augmentation [22]. These cross-domain models demonstrate that self-supervised learning is not a single method but a design philosophy in which the supervisory signal is derived from structure internal to the data distribution.

For physiological signals, the most successful self-supervised systems are likely to combine multiple objectives. A wearable representation model might simultaneously predict the next latent state of a cardiac cycle, contrast different leads or sensor channels, and reconstruct masked segments of a photoplethysmographic waveform. This multi-objective design encourages the representation to capture slowly varying physiological state, rapid waveform morphology, and sensor-specific noise characteristics. However, multi-objective training introduces structural trade-offs. The computational cost of maintaining large-scale contrastive dictionaries or transformer-based encoders may exceed the energy and memory budgets of wearable devices. The representation dimensionality must balance expressive capacity with the need for efficient on-device inference. The choice of augmentation strategy must reflect physiological plausibility rather than naive data manipulation. These trade-offs are not resolved by a single optimization metric; they require system-level reasoning about the intended deployment context.

4. Architectural Design for Wearable Deployment

The architectural design of self-supervised physiological representation systems must reconcile two competing demands: the desire for large, expressive models that benefit from large-scale pretraining and the need for compact, energy-efficient models that can operate within wearable hardware. Cloud-based pretraining followed by on-device fine-tuning or distillation provides a practical pathway. In this model, a large encoder is trained on aggregated and de-identified physiological data in a secure computational environment. The resulting representations are then transferred to a smaller student model deployed on the wearable device. This separation allows the pretraining phase to use substantial computational resources without violating the power and thermal constraints of the wearable form factor.

The choice of encoder architecture has important implications for representation quality. Temporal convolutional networks can efficiently capture local waveform morphology and are well suited to low-power inference. Recurrent networks can model long-range dependencies but may be more difficult to parallelize and prune for embedded execution. Transformer-based architectures can capture complex temporal relationships across long sequences, yet their memory and computational requirements can be prohibitive for continuous on-device monitoring. Hybrid architectures that combine convolutional feature extraction with lightweight attention mechanisms have shown promise in clinical electrocardiographic

analysis [23]. Self-supervised representation learning from twelve-lead electrocardiographic data has also demonstrated that clinically useful features can be obtained from unlabeled recordings [24]. Deep networks combining convolutional and recurrent elements have been effective for arrhythmia classification from wearable electrocardiographic signals [25]. These results suggest that no single architecture dominates; the appropriate choice depends on the sensor modality, the downstream clinical task, and the operational constraints of the deployment environment.

The system architecture must also address data transmission and privacy. Continuous streaming of raw physiological data to cloud servers is often impractical because of bandwidth limitations, battery consumption, and privacy risks. A more sustainable approach is to perform feature extraction on the device and transmit only compact representations or task-specific summary statistics. Self-supervised representation learning supports this paradigm by producing embeddings that preserve clinically relevant information while reducing dimensionality. However, the embedding must be designed to prevent unintended re-identification. High-capacity representations may encode demographic, behavioral, or environmental attributes that could enable linkage attacks. Architectural choices therefore intersect with privacy engineering. Techniques such as adversarial representation learning, information bottleneck constraints, and differential privacy can be integrated into the training process to limit the leakage of sensitive attributes while retaining clinical utility.

5. Robustness and Generalization Across Contexts

Robustness in wearable physiological monitoring is not a single property but a multidimensional requirement. A representation must be robust to sensor noise, motion artifacts, changes in skin temperature, sensor placement, and hardware variability across device generations. It must also be robust to distributional shifts across populations, including differences in age, sex, body habitus, skin tone, and comorbid conditions. Self-supervised learning can improve robustness by exposing the model to a broad range of natural variability during pretraining. When the pretraining corpus includes recordings from diverse individuals, devices, and activity states, the resulting representation is more likely to encode physiological state rather than spurious environmental correlates.

Contrastive objectives are especially useful for learning invariance to known nuisance factors. By constructing positive pairs that differ only in a specific transformation, such as sensor displacement or motion artifact, the model can learn to disregard that transformation in the representation space. However, this approach requires careful specification of transformations that are physiologically plausible. Overly aggressive augmentation can destroy diagnostic information, while overly conservative augmentation may fail to induce sufficient invariance. The balance is particularly delicate for electrocardiography, where subtle ST-segment changes can carry critical ischemic information. A contrastive augmentation that smooths or warps the waveform too strongly may eliminate the very features that matter for downstream detection.

Transfer learning across modalities and tasks introduces additional generalization challenges. A representation trained primarily on electrocardiography may not transfer directly to photoplethysmography, even though both signals share cardiovascular origins. Multi-modal contrastive learning offers one path toward sensor-agnostic representations. By aligning representations of simultaneous electrocardiographic and photoplethysmographic recordings, the model can learn shared hemodynamic state variables while remaining robust to modality-specific noise. This approach is structurally similar to vision-language pretraining, where

aligned modalities enable zero-shot and few-shot transfer across tasks [19]. In wearable systems, multi-modal alignment could support a common physiological representation that remains useful when the available sensor suite changes between device generations or use contexts.

Robustness must also be evaluated under realistic temporal and behavioral conditions. Laboratory protocols often collect clean recordings during rest, whereas free-living monitoring includes transitions between sitting, walking, running, and sleeping. A rapid vital sign extraction approach designed for real-time monitoring under varying physical activity intensity levels highlights the difficulty of maintaining stable estimates under motion [15]. Self-supervised representations should be evaluated under similar ecologically valid conditions. Evaluation metrics should include not only average accuracy but also worst-case performance across activity states, sensor quality levels, and demographic subgroups. This perspective shifts attention from benchmark-centric model selection to system-level validation that reflects the operational reality of wearable monitoring.

6. Governance, Fairness, and Data Infrastructure

The deployment of self-supervised physiological representations in wearable devices is inseparable from questions of data governance and fairness. The pretraining data used to learn these representations may be drawn from heterogeneous sources, including research cohorts, clinical trials, consumer device users, and public repositories. Each source carries distinct consent terms, data quality characteristics, and population biases. Without careful governance, self-supervised pretraining can encode structural inequities that are difficult to detect because the training objective does not directly optimize for any clinical label. Bias may be embedded in the representation space and later amplified when the model is fine-tuned for a specific diagnostic or wellness task. Research on algorithmic bias in healthcare has demonstrated that risk scores can systematically disadvantage certain populations even when demographic variables are not explicitly included [16].

Fairness in representation learning requires more than balanced class labels. It requires attention to how physiological signals differ across groups due to normal biological variation, measurement technology, and social determinants of health. Photoplethysmographic signal quality, for example, can vary with skin tone, perfusion, and sensor contact pressure. If a self-supervised model is trained predominantly on recordings from one population, it may learn representations that are less informative for other populations. This failure mode may not appear in aggregate performance metrics, particularly when downstream evaluation sets are similarly biased. Governance frameworks should therefore require subgroup analysis, external validation, and transparent documentation of training data composition. The broader literature on deep learning in healthcare reinforces the importance of external validation and the systematic evaluation of distributional shift [18].

Data infrastructure is equally important. Self-supervised learning benefits from large-scale corpora, but the aggregation of physiological data raises privacy and security risks. Federated learning offers one architectural response by allowing models to be trained across distributed data sources without centralizing raw signals. This approach aligns with the operational structure of wearable ecosystems, where data are generated on millions of distributed devices. Federated self-supervised pretraining can learn population-level representations while preserving local data control. However, federated systems introduce new challenges, including communication efficiency, heterogeneity across devices, and the risk of adversarial participation. These challenges are not purely technical; they require institutional

arrangements that specify data use rights, audit mechanisms, and accountability for model behavior.

The convergence of human and artificial intelligence in medicine has been framed as a transformation of clinical practice [17]. For wearable physiological monitoring, this transformation extends beyond clinical settings into everyday life. Governance models must therefore consider the differing expectations of clinical validation, consumer wellness, and public health surveillance. A representation that is acceptable for general wellness may not satisfy regulatory requirements for diagnostic use. Conversely, a representation optimized for clinical specificity may be unnecessarily burdensome for wellness applications. A tiered governance framework that distinguishes among use cases while maintaining common infrastructure for data quality, privacy, and fairness can provide a practical balance.

7. Sustainability and Operational Implications

Sustainability in self-supervised physiological representation systems has several dimensions. Environmental sustainability concerns the energy and carbon costs of large-scale pretraining. Training large transformer-based models on extensive physiological corpora requires substantial computational resources. These costs may be justified when the resulting representations are shared across many downstream tasks, but they must be weighed against the environmental footprint of repeated experimentation and model retraining. Operational sustainability concerns the ability to maintain representation quality over time as devices, sensor technologies, and population characteristics evolve. A system that requires frequent cloud-based retraining may be less sustainable than one that supports incremental on-device adaptation.

Energy efficiency on the wearable device is a critical operational constraint. Continuous physiological monitoring already competes with other device functions for battery capacity. The added computational burden of neural feature extraction must be carefully managed through quantization, pruning, and architecture search. Self-supervised representations can support efficiency by enabling smaller downstream models, but the pretraining process itself may involve large models that are not deployable on-device. The system-level solution often involves a cascade of models: a lightweight on-device encoder extracts compact embeddings, and more complex inference occurs only when those embeddings indicate a need for detailed analysis. This cascade model reduces average power consumption while preserving the ability to detect rare or high-risk events.

Maintenance and adaptability also raise structural questions. Physiological signal distributions may change with firmware updates, sensor revisions, and shifts in user behavior. A self-supervised representation that performs well at launch may degrade if the user population begins to use the device differently. Continuous monitoring of representation drift is therefore necessary. Self-supervised objectives can themselves be used to detect distributional shift by measuring whether the pretext task performance declines on recent data. If the representation can no longer predict future signal segments or reconstruct masked portions as effectively, that degradation may signal a change in the underlying data distribution. This signal can trigger model retraining or recalibration, forming a closed-loop system that links technical performance to operational maintenance.

The economic sustainability of self-supervised systems depends on the reusability of learned representations. If a single pretrained encoder can support arrhythmia detection, sleep staging, stress assessment, and cardiorespiratory fitness estimation, the marginal cost of adding new

downstream capabilities decreases. This modular structure contrasts with task-specific supervised models that require separate data collection and labeling efforts for each application. The representation also functions as a shared substrate that can be validated once and then reused across multiple algorithmic products. However, this shared substrate can become a single point of failure or bias amplification. If the pretrained representation contains systematic errors, those errors may propagate across all downstream tasks. Governance and validation must therefore treat the shared representation as a critical infrastructure component rather than as an isolated model artifact.

8. Future Directions

Future research should advance self-supervised physiological representation learning along several interconnected fronts. Multi-modal representation learning that aligns electrocardiography, photoplethysmography, accelerometry, and contextual metadata can improve robustness to sensor-specific noise and support richer health state inference. Longitudinal self-supervision that exploits the temporal structure of wearable data across days and weeks may enable the representation of slowly evolving physiological processes, including circadian rhythms, illness onset, and recovery trajectories. Causal representation learning could help distinguish true physiological drivers from spurious correlations caused by behavioral and environmental confounders.

The integration of self-supervised representations with clinical workflows requires further study. Clinicians are unlikely to adopt black-box representations without evidence that the embeddings correspond to physiologically interpretable constructs. Representation probing and interpretability methods should be developed specifically for hemodynamic, autonomic, and respiratory applications. Regulatory frameworks also need to evolve to account for pretrained models that are fine-tuned for different purposes. A representation trained without labels in one institution may be adapted for a diagnostic task in another, raising questions about validation responsibility and model provenance.

Open benchmarking is necessary to compare self-supervised methods across diverse populations and conditions. Current benchmarks often emphasize clean clinical datasets and may not reflect the noise and variability of consumer wearable environments. A community-driven benchmark that includes motion artifacts, device heterogeneity, and demographic diversity would accelerate progress. Such benchmarks should report not only aggregate metrics but also subgroup performance and robustness to distributional shift. This orientation would align technical evaluation with fairness and equity goals.

The convergence of self-supervised learning, federated infrastructure, and privacy-preserving computation offers a path toward large-scale physiological representation learning without centralizing sensitive data. Future architectures may combine on-device pretraining with periodic encrypted aggregation of model updates. These systems could learn from millions of individuals while maintaining local control over raw signals. Realizing this vision will require advances in communication-efficient learning, secure aggregation, and verifiable privacy guarantees. It will also require sustained interdisciplinary collaboration among machine learning researchers, clinicians, human-computer interaction specialists, and health policy scholars.

9. Conclusion

Self-supervised learning holds substantial promise for robust physiological signal representation in wearable devices. By leveraging the vast unlabeled data generated by

continuous monitoring, it can reduce dependence on costly expert annotations and learn representations that transfer across tasks, sensors, and populations. However, the successful deployment of these systems requires more than algorithmic innovation. It demands careful attention to architectural trade-offs, on-device operational constraints, data governance, bias amplification, and long-term sustainability. Contrastive, predictive, and reconstructive objectives each offer distinct advantages, and their combination must be guided by the physiological characteristics of wearable signals rather than by conventions imported uncritically from other domains.

The system-level perspective developed in this paper emphasizes that representation learning is embedded within broader socio-technical infrastructures. Data collection, model pretraining, on-device inference, clinical integration, and regulatory oversight form a continuous pipeline in which technical decisions have institutional and ethical consequences. Robustness cannot be reduced to aggregate accuracy; it must include performance under motion, across demographic groups, and over time. Fairness cannot be reduced to demographic balancing; it must address the structural sources of bias in data provenance and downstream use. Sustainability cannot be reduced to model compression; it must encompass energy consumption, adaptability, and the economic viability of maintaining shared representation infrastructure.

Future progress will depend on the development of evaluation protocols that reflect real-world variability and on governance frameworks that support accountability across distributed deployment chains. Self-supervised representation learning is not a singular technology but a design paradigm that can be instantiated in many ways. Its ultimate value for wearable physiological monitoring will be determined not only by the quality of learned embeddings but also by the robustness, equity, and institutional legitimacy of the systems in which those embeddings are embedded.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
2. Kim, J., Campbell, A. S., de Ávila, B. E.-F., & Wang, J. (2019). Wearable biosensors for healthcare monitoring. Nature Biotechnology, 37(4), 389–406.
3. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597–1607). PMLR.
4. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729–9738). IEEE.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). ACL.
6. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

7. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D. A., & Gramfort, A. (2021). Uncovering the structure of clinical EEG signals with self-supervised learning. Journal of Neural Engineering, 18(4), 046020.

8. Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In Proceedings of the 38th International Conference on Machine Learning (pp. 5606–5615). PMLR.

9. Sarkar, P., & Etemad, A. (2020). Self-supervised ECG representation learning for emotion recognition. IEEE Transactions on Affective Computing, 13(3), 1541–1554.

10. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220.

11. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33(4), 917–963.

12. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.

13. Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In Proceedings of the International Joint Conference on Neural Networks (pp. 1578–1585). IEEE.

14. Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 1735–1742). IEEE.

15. Zheng, X., Dwyer, V. M., Barrett, L. A., Derakhshani, M., & Hu, S. (2023). Rapid vital sign extraction for real-time opto-physiological monitoring at varying physical activity intensity levels. IEEE Journal of Biomedical and Health Informatics, 27(7), 3107-3118.

16. 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.

17. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

18. 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.

19. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.

20. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (pp. 12449–12460). NeurIPS.

21. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (pp. 1096–1103). ACM.

22. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations. ICLR.

23. Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira, W., Jr., Schön, T. B., & Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications, 11, 1760.

24. Mehari, T., & Strodthoff, N. (2022). Self-supervised representation learning from 12-lead ECG data. Computers in Biology and Medicine, 141, 105114.

25. Oh, S. L., Ng, E. Y. K., Tan, R. S., & Acharya, U. R. (2018). Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with ECG signals. Computers in Biology and Medicine, 102, 278–287.